Bounding the Impact of a Gifted Program On Student Retention using a Modified Regression Discontinuity Design

Billie Davis, John Engberg, Dennis Epple, Holger Sieg and Ron Zimmer*

January 24, 2013

*Billie Davis, Carnegie Mellon University. John Engberg, RAND Corporation. Dennis Epple, Carnegie Mellon University and NBER. Holger Sieg, University of Pennsylvania and NBER. Ron Zimmer, Vanderbilt University. We would like to thank an anonymous referee, William Dickens, Robert Gary-Bobo, Glenn Ellison, Guido Imbens, Hendrik Juerges, Brian Junker, Edwin Leuven, John Pepper, Allen Ruby, Robert Siegler, and seminar participants at numerous conferences and universities for comments and suggestions. We would also like to thank the "mid-sized urban school district" for sharing their data. Andrea Phillips provided excellent research assistance. Financial support for this research is provided by the Institute of Education Sciences (IES R305A070117 and R305D090016).

Abstract

This paper examines whether gifted programs can help urban districts retain students. Gifted programs often employ IQ thresholds for admission, which creates strong incentives to manipulate the IQ score of students to increase access to the program. In the presence of local manipulation, the standard regression discontinuity estimator does not identify the local average treatment effect. However, we can construct a lower bound for the effectiveness of the program by exploiting a weak monotonicity assumption regarding the mean outcome in the untreated state. Our point estimates suggest that the gifted program has a strong positive effect on retention for higher income students.

Keywords: Regression Discontinuity Design, Manipulation of Forcing Variable, Monotone Treatment Effects, Gifted Program, Urban School District.

JEL classification: C21, I21, H75

1 Introduction

Student retention has increasingly become an important issue for urban districts, as nearly half of large urban districts in cities classified by the National Center for Education of Statistics lost students between school years 1999-2000 and 2009-2010.[1] Cities in the Midwest and on the East Coast have been hit especially hard. Urban districts in Buffalo, Cincinnati, Cleveland, Detroit, Kansas City, Milwaukee, Pittsburgh, and Philadelphia have lost thousands of students over the last several decades. State funding, which is allocated on a per-pupil basis, has shrunk dramatically, and the declining urban population has led to a lower local tax base. However, these districts have buildings, staffing, and pension systems designed for a much larger enrollment base. Downsizing an urban school district is not only costly, but can also impose serious disruptions for students whose schools are closed and who need to transfer (Engberg, Gill, Zamarro & Zimmer, 2011). Moreover, a decline in enrollment is often accompanied by a decline in peer quality as high achieving students are more likely to opt out of public schools. As a consequence, declining enrollments often lead to declining achievement for the students that remain in the urban district, exacerbating already existing problems in many urban schools.

These pressures have caused urban districts to search for ways to help maintain enrollment numbers. Districts often look to specialized programs – such as magnet and gifted programs – to attract and retain students and households, especially middle class households that have many options in the educational market place.

[1] Information gathered from National Center for Educational Statistics Common Core data available at http://nces.ed.gov/ccd/bat/index.asp. Using these data, we examined school districts that are either in large or mid-size cities by the common core and have at least 10 schools in the 2009-10 school year. In total there were 260 such school districts in 2009-10 school year.

In searching for a school, families may consider not only the quality of facilities, curriculum, and instruction, but also the quality of educational opportunities and peers. Gifted programs create opportunities for students to be stimulated and challenged and have positive peer influences. Often, smaller suburban districts may not have the scale to offer such programs, and, as a result, gifted programs may be a mechanism for retaining strong students within an urban district. The purpose of this paper is to estimate the treatment effect of admittance into a gifted program on student retention using new and unique data from an anonymous urban district.

Evaluating the impact of gifted programs is challenging. Gifted programs are not randomly assigned to districts and may develop in districts as a function of unobservable district characteristics. Moreover, students are typically not randomly assigned to gifted programs. Often, students are admitted based upon IQ scores. The admission by IQ scores raises the possibility of using a regression discontinuity (RD) design, which was recently used by Bui, Craig & Imberman (2011) to estimate the impact of an urban district's gifted program on test score outcomes.[2] In theory, using an RD design to estimate the retention effect of the gifted program should be straightforward. Since students are admitted into gifted programs based on an IQ threshold, we could just use the IQ score as the forcing variable for an RD design. However, IQ examinations are oral tests, which provides psychologists some discretion in scoring. A psychologist may "give the benefit of the doubt" in assessing performance of students who are near the threshold for admission to a gifted program. In our data, many test scores appear to be

[2]The regression discontinuity design was first used by Thistlethwaite & D.Cambell (1960) . Some well-known applications of the RD design in education include Angrist & Lavy (1999) and van der Klaauw (2002), DiNardo & Lee (2004), and Jacob & Lefgren (2004). For a guide for implementing RD designs, see Imbens & Lemieux (2008).

manipulated. Similar issues arise in other settings as concerns for transparency can lead to promulgation of the criteria for admission into programs. Knowing the criteria for admission, participants may undertake activities that alter the reported outcome on the variable that determines admission. For example, students who fall below the threshold on a test determining whether they will be subject to remediation may retake the test to attempt to obtain a score above the threshold (Calcagno & Long, 2008).

In our case, theory suggests that manipulation of IQ scores should be local in nature. The costs of manipulation are too large for students that have IQ scores below a threshold since they would not benefit from the advanced curriculum and might create negative peer effects for more advanced students in the gifted program. Moreover, there is obviously no need to manipulate the test scores of students who have scores above the admission threshold. Local manipulation of IQ scores then gives rise to a non-standard error in variables problem.

We provide two tests for manipulation of IQ scores. These tests help us determine the range over which IQ scores are likely to be manipulated. Both tests exploit the local nature of the manipulation. The first test is based on a monotonicity property of the density of the IQ distribution near the admission threshold. Because IQ scores are standardized to a normal distribution, when the admission threshold is above the mean, the density of the IQ score is monotonically decreasing around the threshold. If there is no manipulation, we should see a higher fraction of students below versus above the threshold. With local manipulation, however, some students with true IQ scores below the threshold have reported IQ scores above the threshold. If the fraction of students whose scores are manipulated is sufficiently large, we will observe the opposite of our expectation, i.e.,

there will be more students with scores above versus below the threshold.[3]

The second test for manipulation is based on scores on achievement tests that are often given along with the IQ test. Achievement test performance is not directly referenced in the admission guidelines of the gifted program, and, more importantly, the scores do not appear to be manipulated. We can, therefore, use the observed IQ score to predict the achievement score. Local manipulation creates an error-in-variables problem over a bounded interval of values. Using a regression model to predict the achievement score, we can test for parameter stability over the manipulation range. We implement this stability test using a standard Chow test.

One consequence of manipulation is that IQ scores at the admission threshold may not be valid instruments for program participation, which invalidates the key identifying assumption of the standard regression discontinuity design. However, we can exploit the local nature of the manipulation process and construct an estimator for a lower bound of the relevant treatment effect. The key additional assumption that we need to invoke is that the mean outcome in the untreated state is monotonically declining in the IQ score. In our application, this assumption is natural and implies that parents with students that have higher IQ's are less likely to stay in the district in the absence of a gifted program. More able students are likely to have more alternatives, including being sought after by private schools. Under this assumption, we can construct a lower bound of the treatment effect of the gifted program. The basic idea is to compare the mean outcome of students

[3] McCrary (2007) provides an alternative framework for testing for manipulation. He tests the null hypothesis of continuity of the underlying density function at the program cut-off points. In our application, the IQ test is measured in discrete increments which leads to our alternative tests.

with the highest non-manipulated IQ score below the threshold with the outcome of students with the lowest non-manipulated IQ score above the threshold. Our approach draws on the pioneering work of Manski (1997) who first suggested exploiting monotone treatment responses to construct bounds for treatment effects. We show that a similar idea can be used within a regression discontinuity design with a locally manipulated forcing variable.[4] Moreover, we show that we can implement this estimator using a modified RD estimator.

Our application focuses on a gifted program operated by a mid-sized urban school district that prefers not to be identified. We implement our estimation strategy for a sample of students tested for the gifted program while attending a district school in school years 2004-05 to 2007-08. We find large, significant point estimates which provide strong evidence for a favorable retention for higher income students.

With the development of a bound analysis within an RD framework, this paper not only makes a contribution to the important policy question of whether gifted programs are a mechanism for retaining students, but also provides a new method that can be used in other educational contexts in which test scores, or other measures, are manipulated to gain entrance into a program. The rest of the paper is organized as follows. Section 2 provides information about our data set and describes the testing and admission procedures used in the gifted program studied in our application. Section 3 discusses estimation and inference in an RD design with local manipulation. Section 4 presents the empirical findings of our paper. Section 5 offers some conclusions.

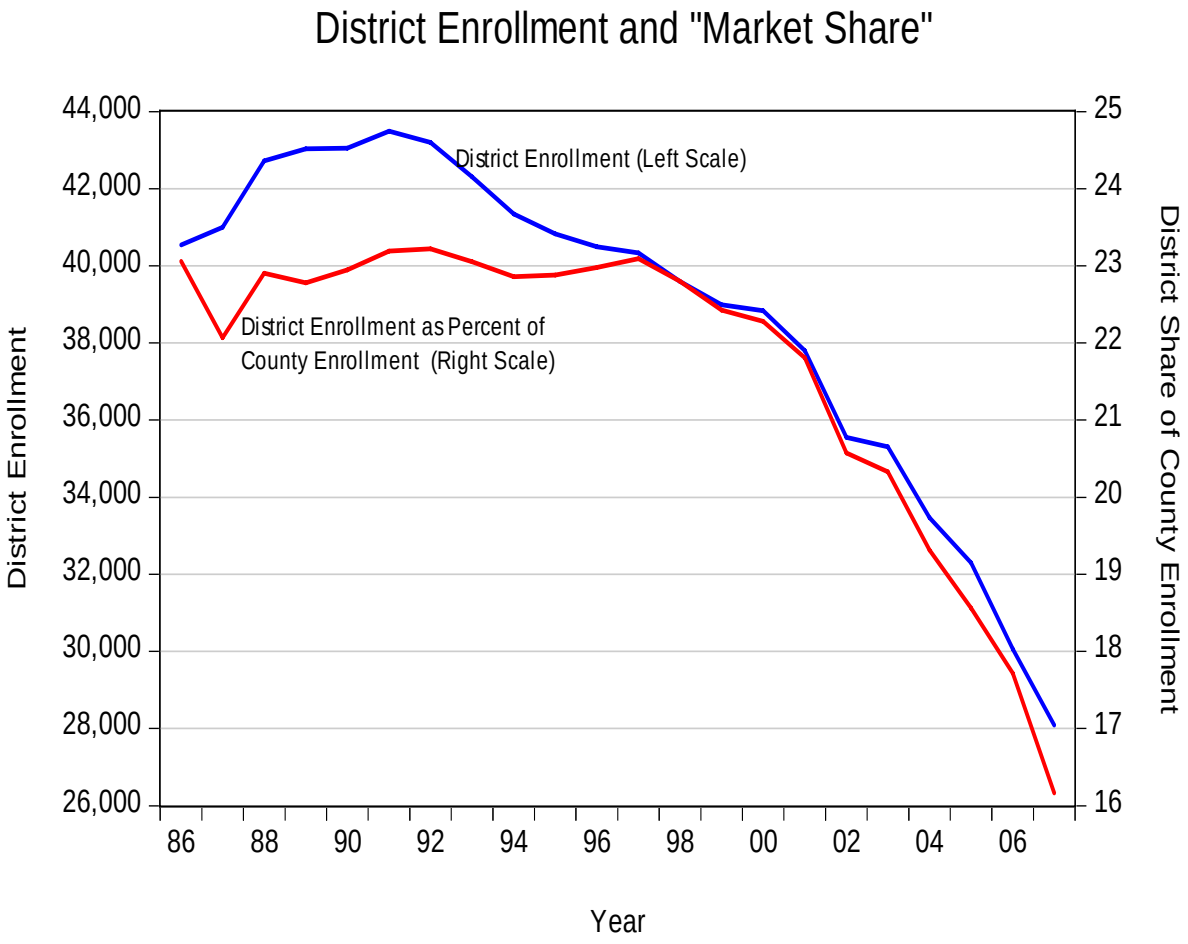
[4]See also Manski & Pepper (2000) who extend this framework using monotone instrumental variables.

2 Institutional Background and Data

Student retention has become one of the key challenges faced by urban school districts. Figure 1 plots the market share of the urban district considered in this paper relative to the broader educational market (measured by all districts in the county). The district was maintaining its student share during the 1990's when enrollment was rising in the market. When countywide enrollment began to decline in 1998, the district not only shared in the countywide decline in the student population, but experienced a further decline as more affluent households exited the city and moved up the school district income hierarchy ("voting with their feet"). The combination of these two effects resulted in the district bearing 75 percent of the countywide decline in public school enrollment.

One promising tool to make an urban school district attractive to students and parents is to offer special education programs that cannot be provided by smaller districts. Gifted programs are one prominent example of such programs. Gifted and talented programs have a long history in the U.S., dating back to the late 19th century. However, gifted programs did not receive federal support until 1958 when the federal government established the National Defense Education Act. This act initiated federal support for specialized programs for math, science, and foreign languages (Bhatt, 2009). More recently, the federal government expanded its support to gifted programs through the Jacob Javits Gifted and Talented Educational Act in 1988 and the No Child Left Behind Act in 2002. Through these initiatives, gifted programs have gained popularity, especially in urban districts. For urban districts, these programs have the dual objective of engaging and challenging gifted students to reach advanced levels of achievement as well as attracting and retaining students who might otherwise leave for sub-

Figure 1: Market Share of the Urban District



urban or private schools. Despite receiving federal support, gifted programs are not mandated by the federal government. Individual states or districts decide if and how to use gifted programs, including how students are identified (Shaunessy, 2003).

Despite the fact that there are currently over 3 million students in gifted programs, these programs have generally been ignored by researchers (Bhatt, 2011). In reviewing the small literature up to that point, Vaughn, Feldhusen and Asher (1991) conducted a meta-analysis of nine papers and found that participation in pull-out gifted programs led to improved achievement, critical thinking, and creativity, but student's self-concepts were not affected. However, these studies often had difficulties dealing with endogeneity issues associated with students self-selecting into the programs. More recently, Bui, Craig, and Imberman (2011) using an RD design to address the self-selection problem, examined the impact of a gifted program on test scores in an anonymous urban district and found that these programs did have a positive impact on language and science test scores after two years of participation, but did not have an effect in other subjects.[5]

The school district that we study in this paper operates a gifted program that is quite large in scope. Approximately 10 percent of the students in the district participate in some type of gifted education. Gifted students in grades 1 to 8 participate in a one-day-per-week pull out program at a designated location away from the student's home school. Students enroll in programs designed to enhance creative problem solving and leadership skills and are offered specially designed

[5] A closely related literature considers school tracking, which sorts students into different tracks based upon ability (Zimmer, 2003; Figlio and Page, 2002; Betts and Shkolnik, 2000; Argys, Rees, and Brewer, 1996; Hoffer, 1992; Kerckhoff, 1986). However, this research has not examined the impact of gifted or tracking programs on retaining students.

instruction in math, science, literature, and a variety of other fields. For high school students, gifted education is available within the school and involves the annual design of an individualized education program, full-time curricula, and a number of other enhancements.

The district adheres to state regulations concerning gifted students and services. The state regulations outline a multifaceted approach used to identify whether a student is gifted and whether gifted education is needed. A mentally gifted student is defined as someone with an IQ of at least 130 points or someone who shows outstanding intellectual and creative ability using other educational criteria. Further, to qualify for gifted services, the district must show that the student requires services or programs not available in regular education.

The state guidelines stress that IQ cannot be the only factor used in determining gifted ability. Specifically, low scores in memory or processing speed tests cannot be used alone to disqualify a student. Also, even if a student has an IQ below 130, she may be deemed gifted based on above grade level achievement on standardized tests, a superior rate of acquisition or retention of new academic content or skills, excellence in specific academic areas, or other factors that indicate superior functioning. Additionally, the guidelines specifically note that the gifted decision must account for any potential masking of gifted abilities because of disability, socio/cultural deprivation, gender or race bias, or English as a second language. Further, it is emphasized that the gifted decision may not be based on a single test or type of test. For limited English proficiency or students of racial-, linguistic-, or ethnic-minority background, it is specifically noted that an IQ score may not be used as the only measure to show low aptitude.

The evaluation process begins when a parent, teacher, administrator, or student requests a gifted evaluation. Once the student's parents are notified and give

consent for the evaluation, a team consisting of parents, a certified school psychologist, teachers, and others familiar with the student's educational experience and performance or cultural background conducts the evaluation. The evaluation must include information on academic functioning, learning strengths, and educational needs. The information and findings from the evaluation of the student's educational needs and strengths is combined by the team into a written report. This report includes the team's recommendation as to whether the student is gifted and in need of specially designed instruction. Finally, the report is evaluated in a team meeting where the decision is made regarding the student's eligibility for gifted education.

One way to support a claim of giftedness is to show superior performance (above 130 points) on an intelligence test. In our district, every student considered for the gifted program is given some type of intelligence test. During the timeframe of our analysis (school years 2004-2005 to 2007-2008) district psychologists mainly used the Wechsler Intelligence Scale for Children, 4th edition (WISC4) test instrument. The WISC4 gives four index scores measuring verbal comprehension (VCI), perceptual reasoning (PRI), working memory (WMI), and processing speed (PSI). It also gives a "Full Scale IQ" (FSIQ) which combines the results from the four indexes and a "Generalized Ability Index" (GAI) which combines the results from the VCI and PRI. The FSIQ, indexes, and GAI are normed, by age, to be representative of the current population of children in the United State and have a mean of 100 and a standard deviation of 15. Thus, a score of 130 is two standard deviations above the mean.

In the district, each student being evaluated for gifted admission takes an intelligence test and is then categorized as meeting the IQ criteria if the FSIQ or

GAI is 130 or above.[6] Students with a FSIQ or GAI of 125 to 129 or a VCI or PRI of 130 or above do not meet the IQ criteria but do qualify for special further consideration through a portfolio evaluation. Students who score below these cutoffs may still be considered gifted based on a further review of other factors. In practice, the probability of a student being admitted into the gifted program increases most at the portfolio cutoff of 125 points. (See Figure 2 in Section 4.)

To investigate the impact of the gifted program on student retention, we start with a set of 1726 students first tested for the gifted program in school years 2004-05 to 2007-08 while attending a district public school.[7] Of these students, we keep 806 who do not receive subsidized lunch.[8] Students eligible for subsidized lunch face different IQ thresholds to be considered for admission into the gifted program. Moreover, retention effects are likely to be more important for higher income households since they have access to more educational options for their children. We exclude 141 students who have more than one set of test results for gifted consideration reported at any point up until summer 2009. Dropping observations with multiple tests removes potential manipulation of the type uncovered by Calcagno and Long (2008). Next, we drop 21 students not tested by a district psychologist, since a private psychologist hired by a student's parent may have incentives to inflate scores. Scores would likely only be reported to

[6]Note that the GAI excludes processing speed and working memory sub-tests. Hence, the GAI offers a way to address the state requirement pertaining to not excluding students based solely on low scores in memory and processing speed.

[7]Note that this excludes students attending a private school when tested. This is in order to isolate retention effects of the program. While attraction effects are also of interest, they are beyond the scope of this paper.

[8]Throughout our analysis, a designation of subsidized lunch means that the student was tagged as receiving subsidized lunch at some point in the district's data from school years 1999 to 2009. Thus, it is a constant variable by student (as are race and sex).

the district if the score is above the admission cutoffs. Additionally, parents who hire a private psychologist may differ on unobservables. We also omit 26 students with nonverbal test scores. Recall that nonverbal tests are given to students who are culturally or linguistically diverse. Thus, these students likely differ on unobservables. Finally, for data comparability, we exclude 9 students with invalid or missing data and 69 students not administered the WISC4 test instrument. This leaves us with our final sample of 540 students.

In addition to IQ and achievement test results for students being evaluated for gifted admission, we have longitudinal student level data for all district students in school years 2004-05 to 2007-08. The data includes information about race, gender, standardized test scores, subsidized lunch status, school attended, home census tract, etc. Table 1 reports descriptive statistics for the state (data column 1), the district (data column 2), and students in our sample (data columns 3 and 4).[9]

Comparing the district to the state, we find that the district has a similar proportion of males and a larger proportion of African American students than the state. Also, district students, on average, live in neighborhoods with lower median income and fewer college educated adults and perform below the state average on a 5th grade state-wide standardized test in math and reading.

Next, comparing students in our sample to all district students, we see that tested students have a similar proportion of males but a smaller proportion of African American students than the district. Also, tested students come from

[9]State averages for race, gender, and test scores are for school year 2005-06. State averages for income and college are from Census 2000. District averages are for students in kindergarten through 12th grade enrolled in the district at some time between 2004 and 2007. Standard deviations are in parentheses.

relatively richer and more educated neighborhoods and score higher on the standardized tests. Admitted students come from even richer and more educated neighborhoods and score even higher on both math and reading. These demographic differences are partly due to the fact that our sample excludes students who receive subsidized lunch.

Considering just the students in our sample, we see that tested and admitted students have a higher average IQ than students who were tested but not admitted. Finally, Table 1 shows that the means of one and two year retention for admitted students is higher than the means for students tested but not admitted. This suggests that admittance into the gifted program may have some positive impact on retention.

3 Estimation and Inference in an RD Design with Local Manipulation

3.1 The Fuzzy Regression Discontinuity Design

We adopt standard notation in the program evaluation literature and consider a model with two potential outcomes.[10] Let D denote an indicator variable that is equal to one if a person receives treatment and zero otherwise. In our application, treatment is admittance into the gifted program. Let Y_1 denote the outcome with treatment and Y_0 the outcome without treatment, where the outcome of interest is retention in the district. The researcher observes:

$$Y = D\,Y_1 + (1 - D)\,Y_0 \tag{1}$$

[10] See Quandt (1972), Rubin (1974, 1978), and Heckman (1978, 1979) .

Table 1: Descriptive Statistics

	State	District	Tested & Admitted	Tested & Not Admitted
Male	0.514	0.506	0.506	0.514
African American	0.161	0.559	0.079	0.176
Income	40,106	28,868	43,575	37,238
		(11,153)	(17,268)	(12,000)
College	0.224	0.185	0.443	0.283
		(0.157)	(0.234)	(0.209)
Math	1420	1308	1732	1512
		(221)	(178)	(155)
Reading	1310	1246	1585	1446
		(217)	(142)	(155)
IQ			131.635	112.072
			(7.630)	(8.238)
1 year retention			0.918	0.847
			(0.274)	(0.361)
2 year retention			0.840	0.743
			(0.368)	(0.438)
Count			318	222

Standard deviations in parentheses.

The difference from receiving the treatment is defined as

$$\Delta \;=\; Y_1 \;-\; Y_0 \tag{2}$$

and note that this gain is unobserved for every single person in the sample. In terms of the treatment effect, the model can be written as

$$Y \;=\; Y_0 \;+\; D\,\Delta \tag{3}$$

We focus on the "fuzzy" regression discontinuity design in which the probability of receiving the treatment changes discontinuously at certain points in the support of a forcing variable. In our application the forcing variable is ability measured by an IQ score. Let Z denote the observed IQ score. Under the fuzzy design D is a random variable given Z. The propensity score is defined as:

$$E[D|Z] \;=\; Pr\{D = 1|Z\} \tag{4}$$

The RD design is then based on three assumptions. The first assumption formalizes the notion that the probability of admission into the program is discontinuous at a known point z_0:

Assumption 1 *$Pr\{D = 1|Z\}$ is known to be discontinuous at z_0.*

Following Hahn, Todd, & van der Klaauw (2001), identification of the treatment effect can be established using the following argument. Let $e > 0$ denote an arbitrarily small positive number. Then

$$\begin{aligned} E[Y|Z = z_0 + e] \;-\; E[Y|Z = z_0 - e] \;&=\; E[\Delta\, D|Z = z_0 + e] \;-\; E[\Delta\, D|Z = z_0 - e] \\ &+\; E[Y_0|Z = z_0 + e] \;-\; E[Y_0|Z = z_0 - e] \end{aligned} \tag{5}$$

The second assumption of the RD design guarantees that the second term in equation (5) vanishes:

Assumption 2 *$E[Y_0|Z]$ is continuous at z_0.*

Hence

$$\lim_{e \to 0} E[Y_0|Z = z_0 + e] \; - \; E[Y_0|Z = z_0 - e] \; = \; 0 \tag{6}$$

The third assumption of the RD design is:

Assumption 3 *D is conditionally independent of Δ near z_0.*

This assumption implies that the first term in equation (5) can be written as:

$$\begin{aligned} E[\Delta \; D|Z = z_0 + e] - E[\Delta \; D|Z = z_0 - e] \; &= \; E[\Delta|Z = z_0 + e] \; E[D|Z = z_0 + e] \\ &- \; E[\Delta|Z = z_0 - e] \; E[D|Z = z_0 - e] \end{aligned} \tag{7}$$

Assumptions 1-3 together imply that the treatment effect is identified from the following equation:

$$E[\Delta|Z = z_0] \; = \; \lim_{e \to 0+} \frac{E[Y|Z = z_0 + e] \; - \; E[Y|Z = z_0 - e]}{E[D|Z = z_0 + e] \; - \; E[D|Z = z_0 - e]} \tag{8}$$

3.2 Local Manipulation of the IQ Score

If the observed IQ score Z is manipulated, the true IQ score must be treated as a latent variable. We observe instead a manipulated score denoted by $\tilde{Z}$. We make the following assumption regarding the extent of the manipulation:

Assumption 4

1. *The IQ is not manipulated for values that are sufficiently far away from the cut off point z_0, i.e., there exists a known value $\underline{z}$ such that if $Z_i < \underline{z}$, then $\tilde{Z}_i = Z_i$.*

2. *Scores above z_0 are not manipulated, i.e., $Z_i \geq z_0$ implies $\tilde{Z}_i = Z_i$.*

Assumption 4 implies that manipulation is local in nature. Only IQ scores between $\underline{z}$ and z_0 are potentially manipulated. This assumption is plausible since district psychologists have no incentive to manipulate scores above the admission threshold. Moreover, the costs of admitting unqualified students to the program is increasing in the distance from the threshold. The detrimental effects to the program through lower peer quality are one example. This implies that students with sufficiently low scores should not be admitted into the program. In that case, district psychologists have no incentive to manipulate the scores of students below a given lower bound denoted by $\underline{z}$.[11]

However, not all IQ scores in the relevant range are manipulated. Let M denote a random variable which is equal to one if the observation is subject to manipulation and zero otherwise. We assume that district psychologists engage in selective manipulation:

Assumption 5

1. *$M = 1$ with probability P^M. In that case*

$$\tilde{Z} = z_0 + E \tag{9}$$

where E is an individual specific manipulation error.

2. *Moreover there exists a known value $\overline{z}$ such that $z_0 \leq \tilde{Z} \leq \overline{z}$.*

Assumptions 4 and 5 imply that observations outside the range $[\underline{z}, \overline{z}]$ are not manipulated.

[11] It is fairly straightforward to derive this result in a simple theoretical model in which a decision maker optimally manipulates scores to increase access to the gifted program.

Note that it is straightforward to allow the probability of manipulation to be a function of observables and unobservables. Strategic manipulation arises if students that are more likely to be retained are receiving favorable treatment and vice versa.

3.3 Bounding the Treatment Effect

With local manipulation of the IQ score the treatment effect is no longer point identified under the assumptions made above. Here we construct a lower bound of the treatment effect based on a monotonicity assumption of the treatment effect, exploiting the fact that manipulation is only local in nature.

Assumption 6 *For values in the interval* $[\underline{z}, \overline{z}]$*, we assume that* $E[Y_0|Z]$ *is monotonically decreasing in* Z

This type of monotonicity assumption was first used by Manski (1997) who showed how to exploit monotone treatment responses to construct bounds for treatment effects. Here we show that a similar idea can be used within a regression discontinuity design with locally manipulated forcing variables to construct bounds for treatment effects.

A slightly weaker assumption is that $E[Y_0|Z = \underline{z}] > E[Y_0|Z = \overline{z}]$. In our application, this assumption means that parents with students who have higher IQ's are less likely to stay in the district in the absence of a gifted program. As we noted earlier, this is a natural assumption because more intelligent students will have more attractive outside alternatives.

Finally, we assume that the expected marginal treatment effect is constant over the interval $[\underline{z}, \overline{z}]$:

Assumption 7 $E[\Delta|\tilde{Z} = \underline{z}] = E[\Delta|\tilde{Z} = z_0] = E[\Delta|\tilde{Z} = \overline{z}]$

We then obtain the following result:

Proposition 1 *A lower bound for the treatment effect of the program is given by:*

$$LB \quad = \quad \frac{E[Y|\tilde{Z} = \overline{z}] - E[Y|\tilde{Z} = \underline{z}]}{Pr\{D = 1|\tilde{Z} = \overline{z}\} - Pr\{D = 1|\tilde{Z} = \underline{z}\}} \tag{10}$$

Proof:

Assumptions 4 and 5 imply that $Z(= \tilde{Z})$ is observed without error outside the manipulation range $[\underline{z}, \overline{z}]$. Thus all the conditional expectations in equation (10) are identified.

Assumption 1 implies that the denominator of the right hand side of equation (10) is greater than zero. Replacing $\tilde{Z}$ by Z, we have

$$\begin{aligned} E[Y|Z = \overline{z}] \ - \ E[Y|Z = \underline{z}] \quad = \quad & E[\Delta\ D|Z = \overline{z}] \ - \ E[\Delta\ D|Z = \underline{z}] \\ & + E[Y_0|Z = \overline{z}] \ - \ E[Y_0|Z = \underline{z}] \end{aligned} \tag{11}$$

Assumption 6 implies that

$$E[Y_0|Z = \overline{z}] \ - \ E[Y_0|Z = \underline{z}] < 0 \tag{12}$$

Assumptions 3 and 7 imply that

$$E[\Delta\ D|Z = \overline{z}] - E[\Delta\ D|Z = \underline{z}] = E[\Delta|Z = z_0]\ (Pr\{D = 1|Z = \overline{z}\} - Pr\{D = 1|Z = \underline{z}\}) \tag{13}$$

Note that we need the slightly stronger assumption of conditional independence for all values in $[\underline{z}, \overline{z}]$, not just those near z_0. Combining these results, we have:

$$\frac{E[Y|Z = \overline{z}] - E[Y|Z = \underline{z}]}{Pr\{D = 1|Z = \overline{z}\} - Pr\{D = 1|Z = \underline{z}\}} < E[\Delta|Z = z_0] \tag{14}$$

We thus have a lower bound for the treatment effect. Q.E.D.

3.4 Detecting Local Manipulation

Here we discuss two tests to detect local manipulation of the IQ score. The first test is based on a monotonicity assumption of the density of the IQ distribution.

Assumption 8 *We assume that the IQ threshold, z_0, that determines access to the programs is sufficiently high that the density of the true IQ score is monotonically decreasing around the cut-off point.*

The monotonicity assumption is realistic for the whole underlying population. However, the students that take the IQ test are not a random sample of the underlying population. If parents and teachers can accurately predict who will pass the admission test, one may expect that this assumption could be violated in the selected sample of test takers.

In practice, IQ scores are reported in discrete intervals. Therefore, to construct a test for local manipulation it is useful to treat Z as a discrete random variable. The proportion of students with scores that are within a bandwidth of h below the cut-off is denoted by $Pr\{z_0 - h \leq Z < z_0\}$. Similarly, the proportion of students in the same length interval above the threshold is $Pr\{z_0 \leq Z < z_0 + h\}$. In the absence of manipulation, $Pr\{z_0 - h \leq Z < z_0\} > Pr\{z_0 \leq Z < z_0 + h\}$. We test to see if this holds.

The second test exploits the availability of a non-manipulated achievement test score denoted by X. By definition, we have

$$X = E[X|Z] + V \tag{15}$$

Notice that $E[X|Z]$ is not known for values in the range $[\underline{z}, \overline{z}]$ since IQ scores are potentially manipulated in that range. We, therefore, need to impose a functional form assumption:

Assumption 9 *The functional form of $E[X|Z]$ is known up to a parameter value that can be consistently estimated based on the non-manipulated subsample.*

For example, if $E[X|Z] = \alpha_0 + \alpha_1 Z$, we can consistently estimate the parameters of this function based on the non-manipulated sample using OLS. In the absence of manipulation a regression of X on $\tilde{Z}$ should give a consistent estimator of the parameters of the regression model whether or not the observations in the range $[\underline{z}, \overline{z}]$ are used. If manipulation occurs, however, a regression of X on $\tilde{Z}$ using the full sample will not yield a consistent estimator due to the error-in-variables problem that is created by manipulation. Regressions using observations below $\underline{z}$ and/or above $\overline{z}$ will give consistent estimators in the presence of manipulation. This suggests a strategy for investigating manipulation. In particular, we conduct Chow tests to assess whether the parameters estimated using observations in the interval $[\underline{z}, \overline{z}]$ differ significantly from parameters estimated using observations above and below that interval. In addition, Assumption 9 can be tested by investigating stability between observations below and observations above interval $[\underline{z}, \overline{z}]$.[12]

An alternative and potentially more robust way to implement this test is to (i) estimate the parametric model for $E[X|Z]$ on the range $(-\infty, \underline{z})$; (ii) test whether this model holds on $(\overline{z}, \infty)$, and (iii) test whether it holds on $[\underline{z}, \overline{z}]$. If the test is not rejected on $(\overline{z}, \infty)$, this can be viewed as a good indication that the model is valid for the full support.

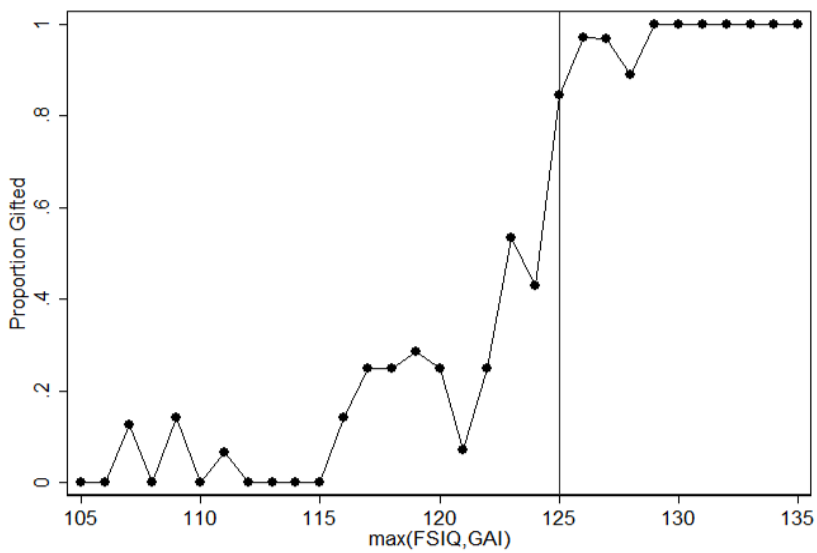
[12] We have assumed that $\underline{z}$ and $\overline{z}$ are known to the econometrician or at least can be bounded from above and below. In most applications that assumption will be valid. In any empirical application, it is useful to conduct some robustness analyses. We discuss these checks in the next section.

4 Empirical Results

4.1 Manipulation of IQ Scores

Since the district's regulations are in terms of both FSIQ and GAI, a natural starting point for the analysis is to consider the maximum of these two scores as the forcing variable that determines access to the gifted program. Figure 2 shows the proportion of students who are gifted as a function of the maximum of the FSIQ and GAI score.

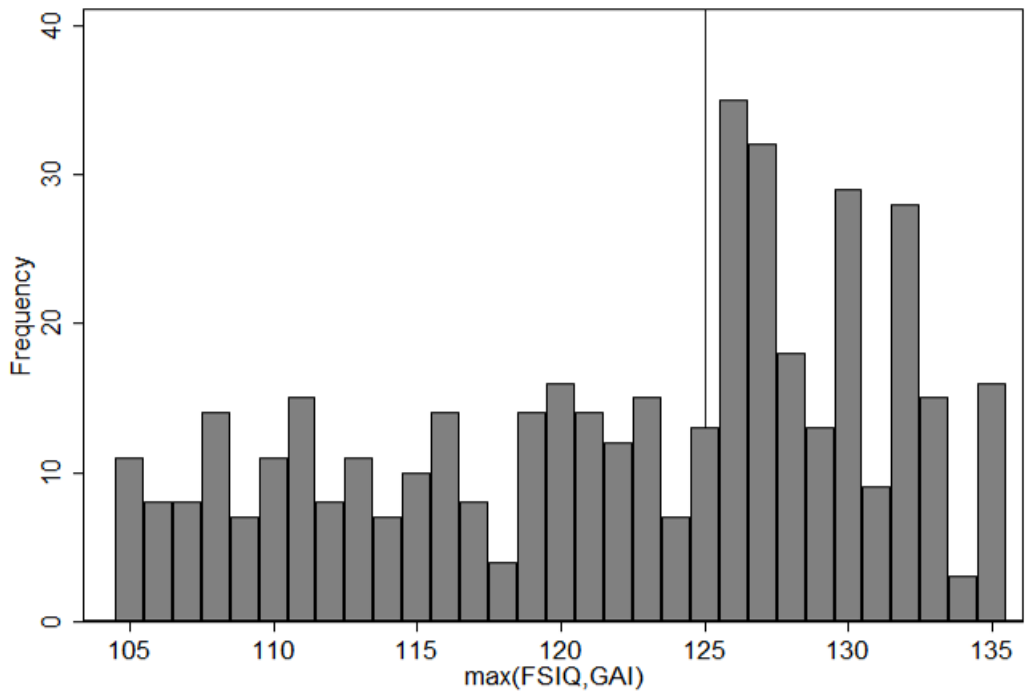
Figure 2: Proportion Gifted by Max(FSIQ,GAI)



We find that higher scores generally correspond to a higher proportion of students admitted into the gifted program. The largest jump in proportion gifted occurs at the value 125.[13]

[13]Note that there are some students who are admitted into the program without meeting the IQ requirements. This is consistent with the requirement that students not be rejected solely on failure to meet the IQ thresholds.

Figure 3: Score Distribution



For the fuzzy RD design to be valid, the distributions of any covariates, including the forcing variable, should not show a discontinuity at the cutoff. Here we encounter a puzzling feature of the distribution of the main forcing variable. The maximum of FSIQ and GAI does not exhibit a smooth frequency distribution. Figure 3 plots a histogram of the distribution. We find that the distribution is heavily skewed to the right around the cut-off point of 125. This finding is robust to a number of sensitivity checks. For example, the patterns are not driven by any one administering psychologist or by testing in one particular grade or school year.[14]

Next, we implement the two tests discussed above to provide more formal evidence of manipulation of the forcing variable.

The first test exploits a monotonicity assumption of the non-manipulated IQ distribution. Recall that this means that in the absence of manipulation we should see $Pr\{z_0 - h \leq Z < z_0\} > Pr\{z_0 \leq Z < z_0 + h\}$. We implement the first test for a variety of different bandwidth parameters. We also consider two samples: the full sample and a subsample of students for whom we also have additional achievement test scores, namely from the Wechsler Individual Achievement Test (WIAT). Note that we can only implement the second manipulation test for the WIAT subsample. The empirical results associated with the monotonicity test are summarized in Table 2.

Table 2 provides strong evidence to suggest that IQ scores are locally manipulated. The null hypothesis that the proportion of observations below the cut-off

[14]The protocol for converting raw scores into IQ creates spikes at certain values of the distribution, as evidenced by spikes at 132 and 135. This phenomenon may also be a contributor to spikes earlier in the distribution, but this phenomenon alone cannot account for the evidence of manipulation documented below.

Table 2: Monotonicity Test

	Full Sample		WIAT Subsample	
Bandwidth	Proportion Below Cutoff	Observations in Interval	Proportion Below Cutoff	Observations in Interval
2	0.31	70	0.27	44
3	0.30	114	0.28	76
4	0.33	146	0.32	104
5	0.37	175	0.36	127
6	0.36	218	0.36	163

is at least 50% is rejected at the 1% level at all bandwidths for the full sample and for the WIAT subsample. That is, we find that, contrary to the expected relation, the proportion of students in the interval at or above the IQ cut-off of 125 is significantly higher than the proportion below the cut-off. For instance, at a bandwidth of 6, 36% of all students in the interval are below the 125 cut-off while the remaining 64% are at or above the cut-off.

To implement the second test for manipulation, we need a non-manipulated achievement test score. Natural candidates are the WIAT scores for numerical operations and word reading. These scores are not used in the gifted admission decision and thus are unlikely to be manipulated. Using the WIAT subsample, we implement Chow tests for IQ manipulation. Table 3 summarizes the main findings of the Chow tests using the two WIAT scores as dependent variables.

Test I in Table 3 tests Assumption 9 by investigating the stability of parameters across observations below (Group A) and observations above (Group C) the manipulation range. Consistent with Assumption 9, we find no evidence to reject

Table 3: Chow Test

	Test I		Test II		Test III	
	Group A vs Group C		Groups A&C vs Group B		Group A vs Group B	
	Num Op	Word	Num Op	Word	Num Op	Word
IQ	0.518**	0.680***	0.684***	0.539***	0.518**	0.680***
2nd Group (dummy)	-7.297	30.18	-75.70**	-24.30	-93.98**	-8.66
2nd Group * IQ	0.0876	-0.248	0.589**	0.198	0.755**	0.0573
Constant	48.14**	33.75**	29.86***	49.39***	48.14**	33.75**
Observations	232	232	382	382	279	279
R-squared	0.375	0.39	0.275	0.313	0.162	0.271
Joint Test:						
F(2, n-4)	0.45	0.59	3.19	0.52	2.76	0.40
P-value	0.6385	0.5533	0.0422	0.5941	0.0652	0.6683

Robust standard errors, ***$p < 0.01$, **$p < 0.05$, * $p < 0.1$

Group A: Observations below manipulation range: [100,119]

Group B: Observations in the manipulation range: [120,130]

Group C: Observations above the manipulation range: [131,150]

Test I: Chow test of Group A vs Group C; 2nd Group is Group C

Test II: Chow test of combined Groups A&C vs Group B; 2nd Group is Group B

Test III: Chow test of Group A vs Group B; 2nd Group is Group B

the hypothesis of parameter stability across these two intervals outside the manipulation range. In particular, the tests of parameter stability for the two WIAT scores yield p-values of 0.64 and 0.55. In Test II, we compare parameter stability for observations outside the manipulation range (Groups A and C) to observations in the manipulation range (Group B). For Numeric Operations we find a p-value of .04, which provides relatively strong evidence of manipulation. We do not find evidence for manipulation when we use Word Reading as the dependent variable for Test II. Test III compares observations below the manipulation range (Group A) to observations in the manipulation range (Group B). Here the findings echo those from Test II. We find evidence of manipulation using Numerical Operations but not using Word Reading. Overall, the two testing strategies provide relatively strong evidence of manipulation.

4.2 Retention Effects

Given the evidence above of local manipulation of IQ scores, the standard RD estimator is likely to be biased.[15] Therefore, we implement our modified RD estimator which provides a lower bound of the treatment effect. We can estimate this lower bound using the Imbens' (2007) approach for standard RD estimators. In our application, a conservative estimate for the manipulation range is $\underline{z} = 120$ and $\overline{z} = 130$.[16] We can then write the outcome equation as:

$$Y_i = \beta_0 + \beta_1 1\{Z_i < 120\}(120 - Z_i) + \beta_2 1\{Z_i > 130\}(Z_i - 130) + \Delta D_i + U_i \quad (16)$$

where Z_i is the observed IQ score. For a bandwidth of five, for instance, we only use data in the intervals of $[115, 119]$ and $[131, 135]$ to estimate the model. The

[15] For the sake of completeness, we report the results from the standard RD approach in Appendix B. of the paper. Not surprisingly, we do not find any significant retention effects.

[16] Robustness analysis regarding these cut-off points is discussed below.

new instrument is $1\{Z_i > 130\}$.

Table 4: First Stage of the Modified RDD

Full Sample						
	Constant			Linear		
	BW4	BW5	BW6	BW4	BW5	BW6
Difference	0.775***	0.820***	0.842***	0.661***	0.624***	0.626***
St. Err	(0.067)	(0.055)	(0.049)	(0.024)	(0.017)	(0.014)
Obs.	95	121	140	95	121	140
WIAT subsample						
	Constant			Linear		
	BW4	BW5	BW6	BW4	BW5	BW6
Difference	0.742***	0.805***	0.830***	0.636***	0.588***	0.591***
St. Err	(0.080)	(0.063)	(0.055)	(0.030)	(0.021)	(0.017)
Obs.	73	95	111	73	95	111

We implement this estimator for a variety of different bandwidths and use both local linear and local constant estimators. Table 4 reports the first stage of the 2SLS estimator that implements our lower bound.

The first stage results show that the proportion of students admitted into the gifted program is larger above the IQ cutoff. These results are statistically significant at various bandwidths in both the constant and linear specifications. For instance, in the full sample, there are 40 students with IQ's 116, 117, 118, or 119 (bandwidth 4 below the manipulation range) and 55 students with IQ's 131, 132, 133, or 134 (bandwidth 4 above the manipulation range). Of those below the cutoff, 22.5% are admitted while 100% of those above are admitted. Thus, the

difference in the constant specification is 77.5%.

Next, we investigate whether admittance into the gifted program helps the district retain students. We use one and two year retention (whether a student is in a district school one year or two years after being tested) as the outcome variables.[17] The main results are summarized in Table 5. Again we conduct a variety of robustness checks and implement the estimator on two different samples.

Table 5 provides no clear evidence that the gifted program increases one year retention. The point estimates are positive in the majority of the reported cases, but standard errors are large and the overall effects are insignificant. This result for one year retention is not too surprising since it would likely take some time for families of non-admitted students to respond and find suitable alternatives. Considering two year retention, though, we find positive and statistically significant results. For the constant model, the treatment effect is approximately 25 percentage points. The point estimates are even larger for the local linear specification which is the preferred method for implementing RDD estimators.

Finally, we conducted sensitivity analysis with respect to the lower limit of the manipulation range. In the previous analysis, we used 120 as the lower limit. In Table 6 we show results using 119 as the lower limit for the range. The results between the two analyses are quite similar. If anything, the main findings are stronger using 119 as the lower bound since one year retention effects are then significant for the constant specification. We also implemented the estimator using 121 as the lower limit of the manipulation range and again we find the same results. We thus conclude that our main empirical findings are robust to the choice of the manipulation range.

[17] A student is considered in the district two years after testing if she graduated from a district school one year after testing.

Table 5: Second Stage of the Modified RDD

Full sample												
	One Year Retention						Two Year Retention					
	Constant			Linear			Constant			Linear		
	BW4	BW5	BW6	BW4	BW5	BW6	BW4	BW5	BW6	BW4	BW5	BW6
Gifted	0.114	0.112*	0.068	-0.175	-0.124	0.062	0.279**	0.228**	0.162*	0.447	0.448	0.575**
St. Error	(0.075)	(0.061)	(0.056)	(0.115)	(0.089)	(0.098)	(0.113)	(0.093)	(0.085)	(0.316)	(0.283)	(0.271)
p-value	0.129	0.068	0.226	0.128	0.166	0.53	0.013	0.014	0.058	0.158	0.113	0.034
Obs.	95	121	140	95	121	140	95	121	140	95	121	140
WIAT subsample												
	One Year Retention						Two Year Retention					
	Constant			Linear			Constant			Linear		
	BW4	BW5	BW6	BW4	BW5	BW6	BW4	BW5	BW6	BW4	BW5	BW6
Gifted	0.066	0.075	0.027	-0.108	-0.117	0.062	0.318**	0.233**	0.152	0.681*	0.658	0.762**
St. Error	(0.085)	(0.066)	(0.061)	(0.124)	(0.106)	(0.122)	(0.139)	(0.109)	(0.10)	(0.387)	(0.350)	(0.338)
p-value	0.435	0.257	0.656	0.382	0.271	0.609	0.022	0.033	0.127	0.078	0.06	0.024
Obs.	73	95	111	73	95	111	73	95	111	73	95	111

Table 6: Robustness Analysis: Second Stage of the Modified RDD

Full sample: Lower Limit 119												
	One Year Retention						Two Year Retention					
	Constant			Linear			Constant			Linear		
	BW4	BW5	BW6	BW4	BW5	BW6	BW4	BW5	BW6	BW4	BW5	BW6
Gifted	0.151*	0.126*	0.131**	0.139	0.248	0.161	0.228**	0.188**	0.147*	0.84	0.703*	0.578*
St. Error	(0.079)	(0.065)	(0.062)	(0.233)	(0.20)	(0.195)	(0.104)	(0.090)	(0.080)	(0.589)	(0.417)	(0.305)
p-value	0.055	0.051	0.035	0.551	0.215	0.407	0.028	0.036	0.067	0.154	0.092	0.058
Obs.	91	114	137	91	114	137	91	114	137	91	114	137

Summarizing the empirical findings, the relatively large positive point estimates for two year retention suggest that there is a favorable effect on retention for students in our sample, i.e., those not on subsidized lunch. These effects are statistically significant in most cases.

5 Conclusions

Student retention has increasingly become an important issue for urban districts that have experienced large declines in student enrollment during the past decade. This paper examines whether a gifted program can help an urban district retain talented students. We focus on students from higher income families (i.e., those not on subsidized lunch) since such students and families are the most mobile yet the most important for districts to keep in order to maintain the tax base.

Many urban districts operate gifted programs that are large in scope and can be expensive to maintain, making the effectiveness of retaining students an important policy question. Gifted programs often employ IQ thresholds for admission, with those above the threshold being admitted. These types of admission criteria are often mandated by state rules. Unfortunately, such rules create strong incentives to manipulate students' IQ scores in order to increase access to the program. We have proposed two tests that can be used to detect local manipulation of IQ scores. Our tests show that manipulation of IQ scores is present in our data. One consequence of manipulation is that the standard regression discontinuity estimator does not identify the local average treatment effect of the gifted program.

In this paper, we show how to modify the standard RD approach to construct a lower bound for the effectiveness of the program. This lower bound can be estimated using a modified RD estimator. This estimator may prove fruitful for

studying retention in other RD applications – especially in education settings since test scores are often used for admission into schools or specialized programs and in awarding scholarships. In these settings, as in ours, manipulation may be "local," i.e., the scope for successful manipulation will be greatest in the vicinity of the threshold. Such settings offer the potential for application of the bounding approach developed in this paper.

Our point estimates suggest that there is a favorable effect on retention for more affluent students, i.e., students who are not eligible for subsidized lunch. These results are encouraging and suggest that parents value gifted programs and are more likely to keep their children in urban schools if the children are recognized as gifted and obtain special instruction.

References

Angrist, J. & Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *Quarterly Journal of Economics*, *114*, 533–75.

Argyst, L., Rees, D. & D. Brewer (1996). Detracking Americas Schools: Equity at zero cost? *Journal of Policy Analysis and Management*, *15 (4)*, 623-45.

Betts, J. & Shkolnik, J. (2000). Key difficulties in identifying the effects of ability grouping on student achievement. *Economics of Education Review*, *19 (1)*, 21-6.

Bhatt, R. (2009). The Impacts of Gifted and Talented Education. Working Paper.

Bhatt, R. (2011). A Review of Gifted and Talented Education in the United States. Education, Finance, and Policy.

Bjorklund, A. & Moffitt, R. (1987). The Estimation of Wage and Welfare Gains in Self-Selection Models. *Review of Economics and Statistics*, *69*, 42–49.

Bui, S., Craig, S. & S. Imberman (2011). Is Gifted Education a Bright Idea? Assessing the Impact of Gifted and Talented Programs on Achievement. *NBER Working Paper 17089*.

Calcagno, J. & Long, B. (2008). The Impact of Postsecondary Remediation Using a Regression Discontinuity Approach: Addressing Endogenous Sorting and Noncompliance. CCRC Working Paper.

Chay, K., McEwan, P., & Urquiola, M. (2005). The Central Role of Noise in Evaluating Interventions that USE Test Scores to Rank Schools. *American Economic Review*, *95*, 1237–58.

DiNardo, J. & Lee, D. (2004). Economic impacts of new unionization on private sector employers: 19842001. *Quarterly Journal of Economics*, *119 (4)*, 1383-1441.

Engberg, J., Gill, B., Zamarro, G. & Zimmer, R. (2011). Closing schools in a shrinking district: Do student outcomes depend on which schools are closed? *Journal of Urban Econonomics*, *forthcoming*.

Engberg, J., Epple, D., Imbrogno, J., Sieg, H. & Zimmer, R. (2012). Evaluating Education Programs that have Lotteried Admission and Selective Attrition. *Journal of Labor Econonomics*, *forthcoming*.

Figlio, D. & Page, M. (2002). School choice and the distributional effects of ability tracking: does separation increase inequality? *Journal of Urban Economics*, *51*, 497-514.

Fisher, R. (1935). *Design of Experiments*. Hafner. New York.

Gunderson, C., Kreider, B. & Pepper, J. (2011). The Impact of the National Lunch Program on Child Health: A Nonparametric Analysis. *Journal of Econonometrics*, *forthcoming*.

Hahn, J., Todd, P., & van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects With Regression Discontinuity Design. *Econometrica*, *69 (1)*, 201–09.

Heckman, J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica*, *46*, 931–960.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, *47 (1)*, 153–161.

Heckman, J. & Robb, R. (1985). Alternative Methods for Evaluating the Impact of Interventions. In *Longitudinal Analysis of Labor Market Data*. Cambridge University Press.

Heckman, J. & Vytlacil, E. (2007). Econometric Evaluation of Social Programs. In: *Handbook of Econometrics 6b*. Elsevier North Holland.

Hoffer, T. (1992). Middle school ability grouping and student achievement in science and mathematics.. *Educational Evaluation and Policy Analysis*, *14 (3)*, 205-27.

Imbens, G. (2007). Regression Discontinuity Design. NBER Lecture Notes: What's New in Econometrics?

Imbens, G. & T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of Labor Economics*, *142*, 615-35.

Jacob, B. & Lefgren, L. (2004). The Impact of Teacher Training on Student Achievement: Quasi Experimental Evidence from School Reform Efforts in Chicago. *Journal of Human Resources*, *39*, 50–79.

Kerckhoff, A. (1986). Effects of Ability Grouping in British Secondary Schools. *American Sociological Review*, *51*, 842-58.

Lee, D. & Card, D. (2007). Regression Discontinuity Inference with Specification Error. NBER Technical Working Paper 322.

Lee, L. (1979). Identification and Estimation in Binary Choice Models with Limited (Censored) Dependent Variables. *Econometrica*, *47*, 977–996.

Manski, C. (1997). Monotone Treatment Response. *Econometrica*, *65*, 1311–1334.

Manski, C. & Pepper, J. (2000). Monotone Instrumental Variables: With an Application to the Returns to Schooling. *Econometrica*, *68*, 997-1010.

Martorell, P. & McFarlin, I. (2008). Help or Hindrance? The Effects of College Remediation on Academic and Labor Market Outcomes. Working Paper.

McCrary, J. (2007). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, *142*, 698-714.

Quandt, R. (1972). A New Approach to Estimating Switching Regression Models. *Journal of the American Statistical Association*, *67*, 306–310.

Rubin, D. (1974). Estimating Causal Effects of Treatment in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. (1978). Bayesian Inference For Causal Effects. *Annals of Statistics*, *6*, 34–58.

Shaunessy, E. (2003). State Policies Regarding Gifted Education. *Gifted Child Today*, *26*, 16-21.

Thistlethwaite, D. & D.Cambell (1960). Regression-discontinuity Analysis: An Alternative to the Ex-Post Factor Experiment. *Journal of Educational Psychology*, *51*, 309–17.

van der Klaauw, W. (2002). Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression Discontinuity Approach. *International Economic Review*, *43 (4)*, 1249–87.

Vaughn, V. L., Feldhusen, J. F., & Asher, J. W. (1991). Meta-analysis and review of research on pull-out programs in gifted education. *Gifted Child Quarterly*, *35 (2)*, 92–98.

Zimmer, R. (2003). A New Twist in the Tracking Debate. *Economics of Education Review*, *22*, 307-15.

A The WIAT Subsample

Table 7 compares students with and without WIAT scores. Overall there are small differences between the two samples. We find that we are primarily missing WIAT scores for students in first and second grade. Moreover, the sample without WIAT scores has significantly fewer African American students. Two year retention is also higher for the subsample without WIAT scores. Our analysis of treatment effects shows that the results are qualitatively the same across samples.

Table 7: Comparison Between Students with and without WIAT Score

	With WIAT	Without WIAT	Two-sided p-value
Year	2005.63	2005.53	0.36
Grade	3.67	2.96	0.01
Male	0.50	0.54	0.42
African American	0.14	0.07	0.03
College	0.38	0.38	0.86
Income	41048.97	40657.46	0.80
Math	1633.33	1602.25	0.29
Reading	1515.81	1520.10	0.86
FSIQ	119.72	120.66	0.40
GAI	122.34	122.68	0.79
Max(FSIQ, GAI)	123.47	123.94	0.70
Gifted	0.58	0.62	0.36
1 year retention	0.88	0.91	0.34
2 year retention	0.78	0.86	0.05
Count	395	145	

B Standard RDD Estimates

Table 8 reports the first stage of the standard RDD estimator. We find that there exists a discontinuity at the admission threshold of 125. The estimates are significant in all specifications and have the correct sign in 11 of the 12 estimated models.

Table 8: First Stage of the Standard RDD

	Full Sample					
	Constant			Linear		
	BW4	BW5	BW6	BW4	BW5	BW6
Difference	0.626***	0.649***	0.662***	0.193***	0.343***	0.430***
St. Err	(0.072)	(0.061)	(0.055)	(0.026)	(0.019)	(0.015)
Obs.	146	175	218	146	175	218
	WIAT Sample					
	Constant			Linear		
	BW4	BW5	BW6	BW4	BW5	BW6
Difference	0.612***	0.643***	0.654***	-0.129***	0.146***	0.330***
St. Err	(0.088)	(0.073)	(0.064)	(0.038)	(0.028)	(0.022)
Obs.	104	127	163	104	127	163

Table 9 reports the results for the second stage of the standard RD estimator. Here the results are less promising. We find small positive, but insignificant, retention effects when we use the local constant estimator. That finding is robust across different bandwidths. When we use the local linear specification, the estimates are very sensitive to the specification of the bandwidth. This is not surprising since local manipulation implies that this estimator is not consistent.

Table 9: Second Stage of the Standard RDD

	Full Sample											
	One Year Retention						Two Year Retention					
	Constant			Linear			Constant			Linear		
	BW4	BW5	BW6	BW4	BW5	BW6	BW4	BW5	BW6	BW4	BW5	BW6
Gifted	0.103	0.092	0.034	-0.547	-0.039	0.186	0.072	0.039	0.051	-1.32	-0.235	-0.161
St. Error	(0.092)	(0.077)	(0.066)	(1.102)	(0.331)	(0.209)	(0.109)	(0.092)	(0.083)	(1.837)	(0.431)	(0.294)
p-value	0.264	0.234	0.605	0.619	0.907	0.373	0.511	0.671	0.542	0.473	0.586	0.584
Obs.	146	175	218	146	175	218	146	175	218	146	175	218
	WIAT Subsample											
	One Year Retention						Two Year Retention					
	Constant			Linear			Constant			Linear		
	BW4	BW5	BW6	BW4	BW5	BW6	BW4	BW5	BW6	BW4	BW5	BW6
Gifted	0.159	0.121	0.049	2.359	-0.468	0.260	0.12	0.055	0.072	3.422	-0.847	-0.304
St. Error	(0.121)	(0.093)	(0.077)	(4.880)	(1.068)	(0.355)	(0.139)	(0.110)	(0.099)	(6.717)	(1.802)	(0.510)
p-value	0.190	0.193	0.522	0.629	0.661	0.464	0.388	0.619	0.466	0.611	0.638	0.55
Obs.	104	127	163	104	127	163	104	127	163	104	127	163