AERA Open
January-December 2024, Vol. 10, No. 1, pp. 1–15
DOI: 10.1177/23328584241299346
Article reuse guidelines: sagepub.com/journals-permissions
© The Author(s) 2024. https://journals.sagepub.com/home/ero

Exploring the Impact of New York City's Gifted and Talented Program: A Matched Comparison Study

Katherine J. Strickland
University of Pennsylvania

Wendy Chan
University of Pennsylvania

Michael Gottfried
University of Pennsylvania

Jiexuan Huang
University of Pennsylvania

Daniel Hildreth
New York City Department of Education

The Gifted and Talented program in New York City is one of the largest and longest running programs for gifted students in the nation. Yet little is known about its effects on student outcomes. Using student-level administrative data of New York City public school students between the 2010–2011 and 2018–2019 academic years, we studied the effects of the Gifted and Talented program on students' test score performance. Estimates from a matched comparison study show significant gains in middle school English language arts and math proficiency, after controlling for cohort, ethnicity, and district. Our balanced treatment and control groups provided sufficient sample sizes with which to analyze the performance of underrepresented minoritized groups, and we found significant treatment heterogeneity, where Black and Hispanic students showed the largest increase in academic proficiency scores after participating in the Gifted and Talented program. Implications of these findings for education policy—particularly related to new developments in selective school admissions—are discussed.

Keywords: *gifted education, urban education, policy, observational research, quasi-experimental analysis, research methodology, statistics, propensity scores, gifted education, heterogeneous effects, urban education*

The New York City (NYC) Department of Education is home to one of the largest public-school programs for gifted students in the country—the Gifted and Talented (G&T) program. Founded in 2005, the G&T program operates as an admissions-based class option for students in kindergarten through fifth grade. Students who score within the 90th percentile of a standardized test that assesses giftedness are eligible to apply to the program, where they are placed in separate classrooms with other gifted students and a specially trained gifted educator and curriculum. Given that the number of applicants to the G&T program consistently exceeds the number of available spots, qualified applicants are chosen by a combination of random selection and geographic availability. These details, plus the expansiveness and diversity of NYC public schools, make the G&T program a cardinal case study in gifted education for local and national education policymakers to consider.

The debate around gifted education is multifaceted. Advocates for these programs point out that the nation's top students—those scoring in the highest 10% on the National Assessment of Educational Progress—have shown inconsistent growth over the past decade (Duffett & Farkas, 2009; Loveless et al., 2008). The insufficient development of high-achieving students could have significant downstream effects on national innovation, creativity, and groundbreaking discoveries (Wai & Lovett, 2021). This so-called excellence gap is particularly harmful to high-achieving students from low-income and underrepresented minoritized groups, who, despite early academic success, often lack access to the support structures needed to help them reach their full potential in later grades (Plucker et al., 2010). Compounding this issue are other socioeconomic challenges. For instance, low-income students are more likely to attend schools with higher teacher turnover, lower academic performance, and reduced social capital (Hanushek et al., 2004, Ascher & Fruchter, 2001; Fahle et al., 2023). These schools are also less likely to offer advanced class options (Yaluma & Tyner, 2018).

Equity and excellence often present conflicting ideals in education, particularly when they compete for the same limited resources and the attention of policymakers (E. F. Brown & Wisnhey, 2017). In resource-constrained environments, there is a tendency to prioritize support for low-achieving students under the assumption that high-achieving students are more likely to succeed independently. Although some research supports the idea that gifted students benefit from a differentiated curriculum, other studies suggest that these students can achieve similar levels of success in a regular classroom setting (Argys et al., 1996; Betts and Shkolnik, 2000, Steenbergen-Hu et al., 2020). The debate is further complicated by evolving definitions of *giftedness*. Contemporary approaches emphasize the malleability of growth and intellectual potential over measuring and ranking intelligence (Haimovitz & Dweck, 2017; Liu et al., 2012). This shift raises questions about the necessity of tracking and the appropriateness of separate classroom models based on ability.

A critical concern in the debate over gifted education, acknowledged by both supporters and critics, is the persistent and significant racial disparity within these programs. This issue is well documented: in 2012, for instance, White K-12 students in the United States participated in gifted programs at twice the rate of their Hispanic and Black peers (Card & Giuliano, 2016). If gifted programs are indeed effective at educating students, these benefits are clearly not reaching enough talented Black and Hispanic students (Wai & Worrell, 2020). One potential cause of this disparity is the limited availability of gifted education programs in low-income schools. Data from the Office for Civil Rights and the National Center for Education Statistics indicate that more than a third of children in the United States do not attend a school that identifies students as gifted and that higher-income schools are more likely to offer these programs (Peters et al., 2019; Worrell & Dixson, 2022). These disparities have spurred efforts to rethink how more minoritized students can be identified for gifted education. Proposals include implementing universal screening programs to test all students for giftedness, establishing local norms within certain communities to identify a greater proportion of gifted students, and expanding program options in low-income neighborhoods (Kaul et al., 2015; Morgan, 2020; Ferguson, 2022; Peters et al., 2019).

Despite the rigorous debate on both sides, there is little empirical research to match. Our study provides evidence on the effectiveness of gifted education programs by addressing two research questions:

1. What are the effects of participating in NYC's G&T program on academic outcomes?
2. Do the effects differ by race/ethnicity?

We choose to focus on these questions to offer empirical evidence for policymakers in the discussion of gifted education programs. This study analyzed grade-level proficiency scores because academic outcomes in the early grades are linked to a variety of positive outcomes, including economic mobility, long-term health, and civic engagement (Edgerton et al., 2011; John-Akinola, 2014). The second question addresses a major concern about gifted education programs—whether such programs exacerbate divisions along racial and ethnic lines. By examining academic outcomes across different racial and ethnic groups, this study explored whether gifted education programs mitigate educational disparities among underrepresented students or add to chronic educational inequities.

Background

Research that evaluates the causal impact of gifted programs on student outcomes remains inconclusive. One reason for this unresolved question is the variability in how gifted programs are implemented across different schools. Often determined by local education agencies rather than state-level policymakers, gifted education programs vary significantly in both program type and admission criteria. Examples of program types include within-classroom programs, pullout programs, and separate-class programs. Admission strategies also differ, ranging from percentile cutoff scores on school-administered exams to standardized test scores or a combination of quantitative and qualitative measures. This heterogeneity complicates research design. For instance, within-classroom programs may be more susceptible to spillover effects from the general education classroom than separate-class programs. Moreover, treatment assignment within a district may rely on teacher referrals rather than standardized test cutoff scores, leading to preexisting differences between students in the gifted program (treatment group) and those in the general education program (control group) based on the criteria used for admission.

Several descriptive and correlational studies (Aldrich & Mills 1989; Delcourt et al., 2007; Roberts et al., 1992; van der Meulen et al., 2014) have addressed gifted education programs through quantitative research methods, namely naive regression, and a handful of more recent studies estimate the effects of gifted programs using quasi-experimental methods. The latter group considers the nuances of treatment selection and program type to more faithfully estimate program effects.

In one of these quasi-experimental studies, Bhatt (2009) uses National Educational Longitudinal Survey data to estimate the impact of a gifted program on a sample of eighth grade students in the year 1988. She found that participation in a gifted program increased standardized test score performance and the probability of taking Advanced Placement classes. Although the study estimated the effects causally using an instrumental variables approach, the sample of 530 schools all had different eligibility criteria, and the types of gifted programs varied widely among schools. Similarly, Bui et al. (2014) studied a school district in Texas using two different research designs. The first, a regression discontinuity design, compared achievement for students who marginally qualified for the gifted program in sixth grade and those who did not. They found no discernible impacts of participation in the program for students on the margin. The second study capitalized on a lottery system, using random assignment as the treatment design. The authors compared the impact of lottery winners (attending a gifted magnet school) relative to the control group (i.e., attending a gifted program in other schools) (Bui et al., 2014). As in the paper by Bhatt (2009), the lack of significant differences may have been due to program implementation, given that the gifted programs were spread across 1,029 independent school districts.

Card and Giuliano (2014) used data from a large urban school district to study the impacts of separate gifted classrooms on three distinct groups of fourth grade students. This regression discontinuity design was based on an Intelligence Quotient (IQ) threshold. Card and Giuliano compared subgroups of students based on IQ thresholds and socio-economic factors. The authors found no effect on reading or math in the fourth grade for the overall group but found significant effects of separate classrooms on nongifted high achievers (students who missed the IQ threshold but scored high on a statewide achievement test in previous year). Participating in gifted programs increased fourth grade math and reading scores by 0.4 to 0.5 standard deviations, and these impacts were most pronounced for low-income and minoritized high achievers (Card & Giuliano, 2014). In a related follow-up study, Card and Giuliano (2016) analyzed the impact of a universal screening program in a large school district in Florida. The program screened all students on IQ testing and required that students achieve a minimum of 130 points on standard IQ tests to qualify for the gifted program. The study found that universal screening led to a significant increase in the number of poor and minority students who met the IQ standards for gifted status (Card & Giuliano, 2016).

G&T in NYC

Most gifted education programs in the United States are organized and implemented through locally developed policies rather than mandated by state-level requirements. New York City is no different, where decisions on the G&T program are made by the local NYC Department of Education rather than handed down by broader New York state policies. Operating in this local content, NYC's G&T program has, over the last decade and a half, become one of the largest and longest running programs in the state. Where only 11.38% of students in New York state attended schools that recognize gifted and talented youth in 2015–2016 (a significantly lower proportion than at least 40 other states), students in Title I schools in New York state were identified as gifted at a higher rate than those in non-Title I schools (Gentry et al., 2019). This positive representation of students in Title I schools is likely due to the consistency, spread, and popularity of NYC's distinctive G&T program.

To gain a deeper understanding of the G&T landscape in NYC, we collaborated with the NYC Department of Education to explore the workings of this unique district program. Since its inception in 2007, the K–5 G&T program has used a screened admissions process. Each fall, families request that their child be tested, and the results of those tests determine eligibility for applying to G&T program across the city. The program employs two tests: the Otis–Lennon School Ability Test and the Naglieri General Ability Test. These tests, administered under a contract with the testing vendor (Pearson), aim to measure students' abstract thinking abilities and identify giftedness. To qualify for the program, students must score within the 90th percentile on either of these exams (Gossett, 2022; Sewell & Goings, 2019).

The number of G&T classrooms in NYC has varied across years as new programs opened and others closed throughout the last decade. In 2010 through 2020, there were approximately between 80 and 90 programs available to families across the city. These programs are not equitably distributed across NYC's five boroughs, leading to significant disparities in where these programs were genuinely accessible options for families with young children. Moreover, the implementation and oversight of the individual G&T classroom varies widely among school leaders in different local contexts, further complicated by the lack of a standardized gifted curriculum.

The most common type of G&T program is the *district* G&T program, which is integrated within a larger community school. These programs include G&T classes at each grade level alongside general education classes that do not have screened admissions or specialized G&T instruction. Additionally, there are five *citywide* G&T programs, which are stand-alone schools exclusively admitting students through the G&T selection process. These schools are the most sought after and competitive, and they all include middle school grades, further enhancing their appeal as a long-term school option.

Admission criteria differ between these two types of programs. Students who score in the 90th percentile on the G&T assessment are eligible for district programs, whereas those who score in the 97th percentile or higher qualify for both district and citywide programs. Once families receive their child's test results, they can submit an application listing up to 12 preferred program choices, ranked by preference. Placements are then centrally managed by the Department of Education's Office of Student Enrollment based on seat availability and admissions priorities. Priority is given to siblings and residents of the geographic school district for district programs.

A critical aspect of the G&T landscape in NYC is that demand far exceeds the availability of seats in the programs. To provide context, in 2019, 15,185 incoming kindergarteners—representing 20% of all kindergarten students in NYC—were tested, with 3,690 students (24%) qualifying by scoring within the 90th percentile on gifted identification exams. Of those, 2,871 families submitted applications, and 2,222 students received an offer to join a G&T program. Consequently, the likelihood of securing a spot after qualifying and applying is relatively high, at 78%. However, this situation leaves two notable groups of students in the general education pool who are comparable to those in the G&T program: (1) students who successfully applied but did not secure a seat and (2) students who never applied or tested but may have scored within the 90th percentile had they been given the opportunity to take the test.

This Study. Our study estimates the effects of NYC's G&T program using a matched comparison design. Other studies have estimated the effects of ability tracking in NYC high schools (Abdulkadiroğlu et al., 2014; Dobbie & Fryer, 2014) and explored test-taking gaps in the G&T program (Lu & Weinberg, 2016), but the city lacks empirical research that causally identifies the effects of gifted programs on academic outcomes such as math/English language arts (ELA) performance and attendance.

Our study also adds to the limited number of studies using quasi-experimental methods to estimate program effects of gifted education programs, specifically presenting the method under two key design constraints. For one, NYC's G&T program is not run by a true lottery admissions process. This removes the possibility of lottery selection in the design, as in Bui et al. (2014). Therefore, students admitted into the G&T program may differ systematically from those not admitted (and who do not qualify), leading to a selection bias in the analysis. To control for this selection bias, quasi-experimental methods such as matching (on a variety of student and school-level covariates) may be preferred. The second design constraint is the absence of exact pretreatment admissions scores (scores on the gifted identification exams) in the dataset. Estimating the effects of admission-based programs in the absence of admissions scores requires stronger assumptions than those used in the work of Card and Giuliano (2014, 2016). The lack of pretreatment scores is not necessarily a severe limitation, however. Here we use scores from the New York State Testing Program (NYSTP) as a proxy for incoming academic achievement. We also make use of the single cutoff for admissions (i.e., 90th percentile) that remained the same over the time by estimating effects separately for students who score within the 90th percentile of their incoming state-level scores. These students represent students who would likely have qualified for the G&T program had they tested in the first place.

Methods

NYC Student-Level Data

The study used administrative data files from the NYC Department of Education collected from academic years 2005–2006 through 2018–2019. We brought these files together using a randomized student ID and built a dataset that captured information on all students who attended a NYC school for at least some point in grades K–12 in the years 2005–2006 through 2018–2019. The full dataset contained information on 3,129,857 unique students.

From here we built a curated dataset including only certain students with a minimum standard of recorded information. We did not include students with a missing district borough number. We also removed students with missing test score information in grades 3 through 8. We did not include students in a given school year when their demographics were missing for that entire academic year. For students who had repeated ethnicities listed and those who had other repeated demographics, we identified their demographics features at grade 3. This brought the final analytic sample of students who were K–12 students in the NYC public schools for at least 1 year in the years 2005–2006 through 2018–2019 with completed demographic information who attended at least 50 days of school to 2,219,586 unique students.

Class identification of the G&T program began in the 2010–2011 academic year, so we built a subsample of the data (2010–2011 through 2018–2019) to analyze the baseline differences in the G&T program. The sample size of this sample was 548,646 students. Each cohort in this group (a cohort is identified as the year the student was in grade 3) had grade 3, 4, and 5 students in the G&T program. Defining the treatment as those who were ever in the G&T program in grades K–5 in this group, 24,924 students, or 4.54% of the total school population, were ever in the G&T program. The demographic breakdown of the G&T group was 34% White, 12.8% Hispanic, 14.8% Black, 35.2% Asian, 2.52% multiracial, and 0.613% Native American. For the general education group, the demographic breakdown was 23.4% Black, 42.7% Hispanic, 16.2% Asian, 0.893% Native American, 16% White, and 0.784% multiracial. The average grade 3 proficiency scores for the general education and G&T groups were 2.66 and 3.94 for math and 2.44 and 3.65 for ELA,

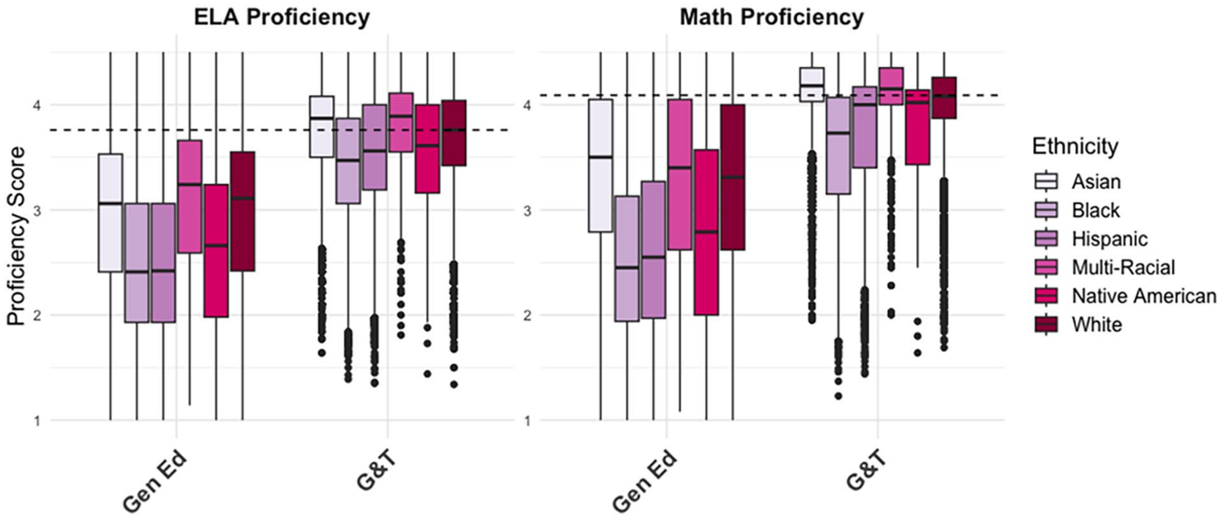

FIGURE 1. *Boxplots of Grade 3 Proficiency Scores.*
Note. Dotted line refers to the 90th percentile score.

respectively. Boxplots of these grade 3 math and ELA proficiency scores are plotted in Figure 1.

The analytic sample used in this study required that students be tested in grades 3 through 8. This restriction (requiring that the student is in NYC schools from grades 3 through 8) brought this analytic sample to 128,464 students. These students were necessarily in cohorts 2010, 2011, 2012, and 2013 due to data availability.

The available demographic features were sex, English language learner status, students with disabilities status, poverty status, and ethnicity. Poverty was a binary variable that captured whether or not the student qualified for free or reduced-price lunch or was eligible for Human Resources Administration benefits. Attendance referred to the number of days a student was present or absent throughout the academic year. The grade 3 through 8 math/ELA proficiency score referred to a student's scores on the NYSTP, a standardized test given to all grade 3 through 8 public-school students each academic year. The scaled scores on the NYSTP range from 148 to 423. These scores were used to determine a student's performance level on a four-point scale, which were denoted as follows: level 1: below proficient; level 2: partially proficient; level 3: proficient; and level 4: exceeds proficiency. The proficiency rating showed where a student fell within a particular performance level. Proficiency ratings ranged from 1.0 to 4.5. Students were identified as being in the G&T program from a separate class list provided by the NYC Department of Education only for the years 2010 through 2019.

Analysis Approach

Because admission into the G&T program is not a random process, estimating the causal impact of participating in the program requires quasi-experimental methods. Here we used propensity score methods (Rosenbaum & Rubin, 1983). Propensity scores model the conditional probability of a unit being in the treatment as a function of observable covariates. Propensity scores have made important contributions to both the observational studies and generalization literature (Stuart et al., 2011) by providing a way to derive bias-reduced estimates in the absence of randomization and experimentation. In the context of our study, students in the G&T program are likely different from students not in the G&T program in observable and unobservable ways. For example, students in the G&T program likely have different grade-level proficiency scores, but they also could be from a different socioeconomic makeup of students. Propensity scores can be used to create matched groups between the two groups of students so that remaining differences in the matched groups are no longer systematic. Additionally, propensity scores have the advantage of being balancing scores in which matching by the propensity score is equivalent to matching on all the covariates used to estimate the propensity score (Rosenbaum & Rubin, 1983). Throughout the remaining sections, we refer to students in the G&T program as the *treatment group* and students not in the G&T program as the *comparison* or *control group*.

To estimate the propensity scores, assume that a vector **X** of covariates is observed for each student in the sample. Z represents the treatment-assignment variable, where $Z = 1$ indicates that a student is in the G&T program and Z = 0 indicates that the student is the control group. The propensity score $s(X)$ is defined as

$$s(X) = \Pr\ (Z = 1|\, X) \tag{1}$$

In practice, it is common to estimate the propensity score using a logistic regression model based on p covariates $\mathbf{X} = (X_1, X_2, \ldots, X_p)$:

$$\log \left\{ s(X) / \left[1 - s(X) \right] \right\} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \qquad (2)$$

Propensity scores are used to address the selection bias in estimates of the average treatment effect. Using matching, treatment individuals are paired with one or more control individuals based on the propensity score. Matching reduces the influence of selection bias by making the treatment and control groups as similar as possible on the observable covariates, mimicking the qualities of a randomized trial (Dehejia & Wahba, 2002; Ho et al., 2007; Rosenbaum & Rubin, 1985).

The validity of propensity score methods depends on several assumptions. First, the stable unit treatment value assumption (Rubin, 1986) must hold. Under this assumption, there is no interference between the two groups of students. Second, the propensity score model must include all moderators of the association of the G&T program on academic outcomes, the so-called strong ignorability assumption (Rosenbaum & Rubin, 1983). Whether or not strong ignorability holds is often controversial because it requires researchers to speculate, in our study, about the factors that moderate the effects of gifted education on academic outcomes. In practice, it is common to include variables that are related to the outcome of interest—namely the variables that predict changes in academic outcomes in this study (Tipton & Olsen, 2018).

An important question is whether the assumptions for propensity scores are satisfied in the context of the study. We argue that there is empirical evidence of their validation. Participating in a G&T program is a fixed feature: A student either does or does not participate. Under strong ignorability, we assume that there is no interference among students in either program. Although many public schools in NYC are co-located due to location constraints and the same building may house two or three different schools, it is unlikely that student experiences in a G&T classroom would affect those of students in a different classroom. With respect to the second assumption on the inclusion of moderators, we acknowledge that the covariates selected for analysis are unlikely to explain all differences between students in the G&T program and those not. However, the data in the study suggest that the chosen covariates are both predictive of G&T participation and academic outcomes and that this relationship held over multiple years.

To obtain the average treatment effect of the G&T program on grade 6 through 8 math/ELA state proficiency, consider three possible comparisons that are based on student's state proficiency exam scores. The first reason for using the state proficiency exam scores as the basis of comparison is that we lacked access to gifted identification test scores. The second is a noted discrepancy in the percentile rates between the gifted identification test and state proficiency exam scores for many students. Although the cutoff for G&T program admission was scoring within the 90th percentile on the gifted identification tests (Otis–Lennon School Ability Test and Naglieri General Ability Test), the grade 3 scores on state proficiency exams (NYSTP) for many students in the G&T program did not fall within the 90th percentile. This could mean that the gifted identification tests are truly measuring something different from what the state proficiency exams measure. Indeed, the former purports to measure giftedness and the latter purports to measure grade-level achievement. Another possibility is that students test differently based on motivation, comfort, or engagement in the testing situation. The gifted identification tests are administered out of school, by a third-party testing agency, and the state proficiency exams are administered in school, by a standardized state testing agency.

Given the absence of exact scores on the gifted identification exam, we estimated effects separately for students who were within the 90th percentile of the grade 3 state standardized exams, using this as a proxy for giftedness. Students who scored within the 90th percentile of grade 3 ELA and math proficiency exams were deemed "high achievers," and the remaining group were deemed "low achievers." In other words, if admissions to the G&T program were only based on 90th percentile of the grade 3 ELA and math state proficiency exam scores and not on 90th percentile of the gifted identification scores, low achievers would not have qualified for the G&T program. The distribution of these incoming proficiency scores are mapped in Figure 1. This high-achieving group can be thought of as those who would have been eligible but did not test and those who did pass the test but did not get a seat from the lottery. Outside this group, the data do not provide information on those who tested and did not test, so we are unable to look directly at those who never tested and those who qualified and did not get a spot.

As mentioned earlier, the analytic subsample consists of students in NYC schools from grades 3 through 8. This restriction ensures that students did not leave the district at any time during their proficiency testing years, preventing treatment contamination from outside district schools. Given the data availability of proficiency scores, this subsample comprised only cohorts 2010, 2011, 2012, and 2013. The 2013 upper boundary is due to available proficiency outcomes: The outcomes for cohort 2014 (eighth graders in 2019) are outside the scope of this analysis because outcome data were only available up to the 2018–2019 school year.

We define the treatment as those students who were in the G&T program (in either a citywide or district program) in any year in grades K through 5. Control students were those who were never in the G&T program in any of those grades. Because cohorts are defined based on when the student was in grade 3 and G&T class data were only available beginning in 2010, the lower bound of cohorts is 2010, to ensure full treatment information. This leaves the chance that some of the control students in cohorts 2010, 2011, and 2012 did receive G&T in grades K through 2, however. But based on the exploratory analysis of the full sample, which showed that 81.60% of students remained in the G&T program from entrance year through grade 5, we assume that the control group is unlikely to both have received G&T services in earlier years and remained in the NYC school system through grade 8. Nonetheless, possible treatment contamination in these early, pretesting grades may be at play.

The final analytic sample included 3,522 students who were in the G&T program for at least 1 year in grades K through 5 and 124,942 students never in the G&T program in grades K through 5. The cohorts are similar in treatment sample size ($N_{T,2010}$= 1,058; $N_{T,2011}$= 853; $N_{T,2012}$= 769; $N_{T,2013}$ = 842). The treatment group comprises 36.30% White, 19.00% Black, 28.50% Asian, and 13.20% Hispanic individuals, with 2.44% multiracial and 0.73% Native American representation. In contrast, the control group consists of 16.0% White, 21.9% Black, 16.3% Asian, and 43.4% Hispanic individuals, along with 1.11% multiracial and 1.25% Native American representation. These demographic distributions align with those of the broader sample population in both the G&T and general education programs.

Matching. Using the analytic sample of G&T program and general education students, we performed matching to estimate the causal effect of participation in the G&T program on academic outcomes. Students were matched within each cohort with nearest-neighbor matching (Stuart, 2010) at a ratio of 3:1 on eight covariates: grade 3 math/ELA proficiency, poverty status, disability status, English language learner status, ethnicity, sex, and grade 3 absences. A total of 3,522 G&T program students were matched with 10,566 general education students.

These variables were selected because they had a significant relationship with G&T program and grade 6 proficiency in multiple regression models, indicating that they are variables that explain both treatment selection and outcome. Omitting important covariates related to both the intervention and the treatment results in bias (Steiner et al., 2010). Average proficiency levels on the NYSTP are different from year to year. A sixth grader's scores in 2011 are not comparable with a sixth grader's scores in the 2012. Despite work by the NYC Department of Education to norm the test from year to year and create comparable proficiency standards, annual variability in the test scores is evident. For this reason, students were matched on the above-mentioned eight covariates within each cohort.

In the unmatched groups, 18.74% of the treatment group are high achievers and 1.27% of the control group are high achievers. In the matched group, 18.74% of the treatment group are high achievers and 12.84% of the control group are high achievers. High achievers scored in the 90th percentile in both math and ELA state proficiency exams. In some cases, high achievers and low achievers were in the same subclass, but this is so only because the qualifying distinction requires both ELA and math proficiency to be above the 90th percentile. Note that we also performed 3:1 ratio matching, which includes both ELA and math grade 3 proficiency. Under this matching, we still observed subclasses with similar state proficiency scores, although some of these students with high proficiency may not have officially met the high-achieving distinction, which is meant for future comparison rather than for direct matching requirements.

We estimate the effect of the G&T program on academic outcomes using two distinct models. The first model is a complex model that measures the effect of G&T program participation on math/ELA proficiency (Y) in grades 6, 7, and 8. The model incorporates fixed effects for ethnicity (X_1), cohort (X_2), and district (X_3), ensuring that any differences observed are not due to variations in these factors. Additionally, the model includes the additive effects of grade 3 math/ELA proficiency ($X_{4,5}$) and grade 3 absences (X_6). to account for prior academic performance and attendance as potential predictors of proficiency in other grades. The model also accounts for additional factors such as sex (X_7), English language learner status (X_8), disability status (X_9), and poverty status (X_{10}). Thus:

$$\begin{aligned} Y = \alpha_0 + \beta_1 GT + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 \\ + \beta_5 X_4 + \beta_6 X_5 + \beta_7 X_6 + \beta_8 X_7 \\ + \beta_9 X_8 + \beta_{10} X_9 + \beta_{11} X_{10} + e \end{aligned}$$

The second model investigates the interaction between ethnicity and treatment effects on the outcome and presents these interaction effects across comparison groups. The treatment sample size for the high-achieving group in certain ethnicities is too low to explore interaction effects. For this reason, we define the high-achieving group in the second model as students who scored in the 90th percentile of either the math or ELA proficiency exams. Under this qualifying distinction, 48.18% of the treatment group are high achievers and 6.92% of the control group are high achievers. Although this is now a larger proportion of the overall treatment group (48.18% compared with 18.74% in Model 1) that meet the high-achieving distinction, both Model 1 and Model 2 still achieve balance in their representation of qualifying students (48.18% high achievers in treatment and 45.30% in control in Model 2 compared with 18.74% high achievers in treatment and 12.84% in control in Model 1).

Multiple robustness checks were run on the data to evaluate the sensitivity of the model to the method of matching and propensity score estimation. To evaluate the influence of the matching method, two new matching methods were explored. In the first, the matching ratio was increased from 3:1 to 5:1. This allowed for more general education students to be matched with students in the G&T program, which increased the control group sample size and potentially the precision of the estimates. In the second robustness check, students were matched within ethnicity (rather than within cohort groups) at a ratio of 5:1, using cohort fixed effects in the outcome model.

Results

Matching Diagnostics

Table 1 presents balance diagnostics from the 3:1 within-cohort nearest neighbor matching model. Differences between the G&T program and general education control groups reflect historical disparities in the program, where students in the G&T program are less likely to be in poverty, to be classified as English language learners, or to be a student with a disability. Likewise, students in the G&T program are more likely to be White and Asian than the general education classrooms. In our matched sample, the standardized mean differences (the difference in means of each covariate between treatment groups, standardized on the same scale for all covariates) are close to zero for all covariates within each matched cohort group. This suggests that matching was successful in improving the covariate balance between the two groups (Stuart et al., 2013).

Effects on Grade 6 Math

Table 2 presents the results of all coefficients from the first model, which predicts grade 6 math proficiency. The coefficients represent the estimated effect of each predictor variable on grade 6 math proficiency while controlling for other variables in the model. Participating in the G&T program was associated with a 0.180 standard deviation increase in grade 6 math proficiency. Several demographic variables showed significant associations with grade 6 math proficiency. Being Black (−0.494), Hispanic (−0.388), or Native American (−0.436) was associated with lower math proficiency compared with other ethnicities. Similarly, students identified as having a disability (−0.153) or those with higher grade 3 absences (−0.009) tended to have lower grade 6 math proficiency. Furthermore, the district of attendance also played a significant role in grade 6 math proficiency. Several districts showed significant negative associations with math proficiency, such as Districts 3, 13, and 16, indicating that students from those districts tended to have lower math proficiency compared with others. Overall, this model explains ~57.80% of the variation in grade 6 math proficiency.

Effects on Grade 6 Through 8 Math and ELA Proficiency

Table 3 presents the treatment effects across three different grade levels (grades 6 through 8) and two subject areas (math and ELA), controlling for ethnicity, cohort, grade 3 district, grade 3 academic scores, grade 3 attendance, poverty status, disability status, English language learner status, and sex. The effects of the G&T program were positive and significant across all grade levels and subject areas, although the magnitude of this range varied by subject area, grade level, and comparison group. The smallest gains (0.052 and 0.059 standard deviations in grade 8 ELA and math proficiency) were made by the high achievers (those who scored within the 90th percentile of the gifted identification exam but not within the 90th percentile of grade-level exams). The largest gains were made by the low achievers in grade 6 math (0.213) and grade 7 math (0.188). Gains made in math generally were larger than those made in ELA, although only by a percent of a standard deviation in most cases. Effects gradually faded out over time, again by only a small magnitude. For instance, the effect of the G&T program on grade 8 ELA was still 60% of the effect of the G&T program on grade 6 ELA. Other fade-out effects were less pronounced, especially in math, where the effect only faded out by 20% in some cases.

Interaction Effects of Ethnicity

Table 4 presents the results from Model 2, which incorporated an interaction effect of the treatment with ethnicity. In the overall group, the effects of participating in the G&T program on the most proximal outcome (grade 6 math proficiency) were highest for Black and Hispanic students. Black and Hispanic students who participate in the G&T program saw gains of 0.240 and 0.214 standard deviations in grade 6 math proficiency. The gain for Asian students was less than half of this (0.102 standard deviations increase in grade 6 math). Across later grade levels, Black and White students saw consistent gains of at least 0.150 standard deviations in math proficiency. Distal outcomes were eventually greatest for White students, who saw the largest effect on grade 8 math scores after participating in the G&T program. For ELA proficiency, the trend is similar: The G&T effect on academic outcomes is driven by Black, Hispanic, and White students rather than by Asian students.

Differential effects for minority groups were even more pronounced among the high achievers. Within this group, the G&T program effect was greatest for Hispanic students in grade 6 math. Hispanic students who were high achievers gained an average of 0.349 standard deviations in math proficiency, which was the largest gain of all ethnicity groups across all measured outcomes. Across later grade levels, Hispanic high achievers continued to see the greatest gains in math achievement among all ethnicities. For ELA achievement, the effect of the G&T program among the high achievers was greatest for Black students in grades 6 and 8, where they gained at least 0.220 standard deviations in each grade level. Hispanic high achievers realized the greatest gains in grade 7 ELA. The effects for White and Asian high achievers were lower than for Black and Hispanic students across grade levels.

In the low-achieving group, White students saw the greatest effect of G&T program participation in math and ELA proficiency. White students gained anywhere from

TABLE 1

Results of Nearest Neighbor Matching for a Sample Cohort (Cohort 2011)

	All students			Matched students		
Factor	Means treated	Means control	SMD	Means treated	Means control	SMD
Asian	0.22	0.158	**0.15**	0.22	0.226	−0.014
Black	0.22	0.241	**−0.05**	0.22	0.223	−0.006
Hispanic	0.156	0.452	**−0.817**	0.156	0.166	−0.028
Multiracial	0.004	0.002	**0.027**	0.004	0.002	0.02
Native American	0.006	0.005	**0.007**	0.006	0.005	0.01
White	0.394	0.141	**0.517**	0.394	0.377	0.034
English language learner	0.009	0.187	**−1.842**	0.009	0.011	−0.02
Students with disabilities	0.027	0.191	**−1.015**	0.027	0.021	0.036
Poverty	0.407	0.72	**−0.637**	0.407	0.415	−0.018
Grade 3 ELA	3.596	2.815	**1.689**	3.596	3.593	0.006
Grade 3 Math	3.794	3.054	**1.569**	3.794	3.776	0.039
Grade 3 Absences	6.292	8.728	**−0.378**	6.292	6.285	0.001
Male	0.496	0.522	**−0.052**	0.496	0.480	0.031

Note. SMD refers to the standardized mean difference. It is the difference in covariate means between G&T program and general education students divided by the pooled standard deviation. All differences between treatment and control students are significant (in bold) at the 0.05 level, tested with the χ^2 test of independence (ethnicity and sex), Wilcoxon rank-sum test (sex, poverty, English language learners, and students with disabilities), and *t* test (grade 3 ELA/math and absences). ELA, English language arts

TABLE 2

Results of Model 1 for Grade 6 Math Proficiency

Estimate	SE	*t* Value	*p* Value	Significance
(Intercept)	0.216	0.055	3.891	0.001***
Grade 3 math	0.444	0.008	58.979	0.001***
Grade 3 ELA	0.153	0.007	20.984	0.001***
Treatment	0.18	0.013	13.667	0.001***
Black	−0.494	0.021	−23.984	0.001***
Hispanic	−0.388	0.02	−18.952	0.001***
Multiracial	−0.151	0.065	−2.319	0.020*
Native American	−0.436	0.089	−4.895	0.001***
White	−0.135	0.015	−8.886	0.001***
Poverty	−0.064	0.012	−5.199	0.001***
Student with disability	−0.153	0.037	−4.14	0.001***
English language learner	−0.058	0.057	−1.021	0.307
Grade 3 absences	−0.009	0.001	−10.418	0.001***
Grade 3 cohort (2011)	0.003	0.015	0.184	0.854
Grade 3 cohort (2012)	0.272	0.016	17.16	0.001***
Grade 3 cohort (2013)	0.243	0.015	15.761	0.001***
Male	−0.019	0.011	−1.724	0.085
District 2	0.044	0.058	0.759	0.448
District 3	−0.305	0.08	−3.815	0.001***
District 4	−0.151	0.081	−1.857	0.063
District 5	−0.289	0.086	−3.351	0.001**
District 6	−0.065	0.078	−0.838	0.402
District 7	−0.476	0.094	−5.043	0.001***
District 8	−0.217	0.066	−3.314	0.001**

(continued)

TABLE 2 (CONTINUED)

Estimate	SE	t Value	p Value	Significance
District 9	−0.069	0.075	−0.912	0.362
District 10	−0.247	0.064	−3.879	0.001***
District 11	−0.023	0.061	−0.373	0.709
District 12	−0.228	0.083	−2.743	0.006**
District 13	−0.431	0.072	−5.983	0.001***
District 14	−0.72	0.086	−8.368	0.001***
District 15	−0.013	0.063	−0.202	0.84
District 16	−0.579	0.083	−6.982	0.001***
District 17	−0.218	0.064	−3.394	0.001**
District 18	−0.149	0.067	−2.234	0.025*
District 19	−0.28	0.068	−4.115	0.001***
District 20	0.042	0.056	0.758	0.449
District 21	0.079	0.058	1.364	0.173
District 22	0.018	0.057	0.324	0.746
District 23	−0.24	0.092	−2.601	0.009**
District 24	−0.002	0.06	−0.04	0.968
District 25	−0.036	0.059	−0.614	0.539
District 26	−0.102	0.063	−1.604	0.109
District 27	−0.291	0.063	−4.611	0.001***
District 28	−0.007	0.067	−0.099	0.921
District 29	−0.161	0.062	−2.6	0.009**
District 30	0.028	0.059	0.47	0.638
District 31	−0.082	0.056	−1.463	0.143
District 32	−0.234	0.089	−2.624	0.009**

Note. Adjusted $R^2 = 0.577$. All groups matched 3:1 G&T program to general education standardized coefficients. A total of 3,522 G&T program students are matched with 10,566 general education students.
$*p < 0.05$; $**p < 0.01$; $***p < 0.001$.

TABLE 3
Matched Data, Complex Model Effects, Controlling for Ethnicity, Cohort, and Grade 5 District

Comparison	Subject	Grade 6	Grade 7	Grade 8
G&T vs. general education ($N_{\text{Treatment}} = 3{,}522$)	Math	0.18 (0.013)***	0.158 (0.013)***	0.131 (0.014)***
G&T vs. general education	English language arts	0.157 (0.015)***	0.135 (0.014)***	0.098 (0.015)***
High achievers ($N_{\text{Treatment}} = 660$)	Math	0.07 (0.018)***	0.057 (0.019)**	0.059 (0.024)*
High achievers	English language arts	0.092 (0.025)***	0.101 (0.024)***	0.052 (0.024)*
Low achievers ($N_{\text{Treatment}} = 2{,}862$)	Math	0.213 (0.015)***	0.188 (0.015)***	0.149 (0.016)***
Low achievers	English language arts	0.178 (0.017)***	0.15 (0.016)***	0.117 (0.017)***

Note. Adjusted $R^2 > 0.55$ for all models. All groups matched 3:1 G&T program to general education standardized coefficients. Includes fixed effects of district, ethnicity, and cohort as well as effects of poverty, students with a disability, English language learner, grade 3 absences, and grade 3 math and English language arts proficiency. A total of 3,522 G&T program students are matched with 10,566 general education students. High achievers are those who scored in 90th percentile of grade 3 math and English language arts proficiency.
$*p < 0.05$; $**p < 0.01$; $***p < 0.001$.

0.219 to 0.317 standard deviations in math and ELA proficiency across grade levels. The magnitude was smaller for Black students among the low achievers, and the gains for Hispanic students were, in some cases (grade 7 and 8 ELA) and all of math, lower than gains made by Asian students in the program.

TABLE 4

Interaction Effects of Ethnicity with Treatment, Controlling for District and Cohort

Factor	Race/ethnicity	*N* (treatment)	Grade 6	Grade 7	Grade 8
G&T program vs. general education			Math		
	Asian	987	0.102 (0.024)***	0.05 (0.023)*	0.035 (0.026)
	Black	732	0.24 (0.024)***	0.217 (0.023)***	0.173 (0.026)***
	Hispanic	479	0.214 (0.024)**	0.148 (0.023)*	0.12 (0.026)
	Other	47	0.143 (0.024)	0.146 (0.023)	0.119 (0.026)
	White	1,277	0.197 (0.024)**	0.214 (0.023)***	0.187 (0.026)***
			English language arts		
	Asian	1,027	0.086 (0.027)**	0.085 (0.026)***	0.046 (0.028)
	Black	714	0.197 (0.027)**	0.147 (0.026)	0.118 (0.028)
	Hispanic	499	0.159 (0.027)	0.153 (0.026)	0.077 (0.028)
	Other	114	0.138 (0.027)	−0.036 (0.026)	−0.019 (0.028)
	White	1,322	0.189 (0.027)**	0.164 (0.026)*	0.136 (0.028)*
High achievers			Math		
	Asian	632	0.057 (0.025)*	0.014 (0.025)	0.013 (0.028)
	Black	169	0.284 (0.025)***	0.206 (0.025)***	0.191 (0.028)**
	Hispanic	133	0.349 (0.025)***	0.285 (0.025)***	0.252 (0.028)***
	Other	30	0.145 (0.025)	0.112 (0.025)	0.241 (0.028)
	White	733	0.133 (0.025)*	0.14 (0.025)***	0.145 (0.028)***
			English language arts		
	Asian	632	0.059 (0.03)*	0.06 (0.028)*	0.026 (0.029)
	Black	169	0.255 (0.03)**	0.237 (0.028)**	0.222 (0.029)**
	Hispanic	133	0.225 (0.03)*	0.273 (0.028)**	0.189 (0.029)*
	Other	30	0.057 (0.03)	−0.089 (0.028)	0.009 (0.029)
	White	733	0.113 (0.03)	0.103 (0.028)	0.088 (0.029)
Low achievers			Math		
	Asian	355	0.192 (0.044)***	0.125 (0.042)**	0.075 (0.046)
	Black	563	0.226 (0.044)	0.222 (0.042)	0.163 (0.046)
	Hispanic	346	0.165 (0.044)	0.104 (0.042)	0.073 (0.046)
	Other	17	0.13 (0.044)	0.171 (0.042)	−0.066 (0.046)
	White	544	0.291 (0.044)	0.317 (0.042)***	0.242 (0.046)**
			English language arts		
	Asian	355	0.114 (0.047)*	0.112 (0.045)*	0.088 (0.05)
	Black	563	0.181 (0.047)	0.123 (0.045)	0.092 (0.05)
	Hispanic	346	0.125 (0.047)	0.107 (0.045)	0.033 (0.05)
	Other	17	0.318 (0.047)	0.084 (0.045)	−0.043 (0.05)
	White	544	0.304 (0.047)**	0.265 (0.045)**	0.219 (0.05)*

Note. Adjusted $R^2 = 0.63$. All groups matched 3:1 G&T program to general education standardized coefficients. "Other" is Native American and multiracial. High achievers are those who scored in the 90th percentile of grade 3 math or English language arts proficiency.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Discussion

This analysis follows from recent studies (Bhatt, 2009; Bui et al., 2011; Card & Giuliano, 2014, 2016) that causally estimated the effects of Gifted & Talented programs on academic outcomes. Our study estimated the effects of a single district-level Gifted & Talented program, as Card and Giuliano (2014, 2016) did in a district in Florida. By using matching, we ensured comparable treatment and control groups with similar observable characteristics, helping to reduce bias and increase the precision of the treatment effect.

The student-level dataset focused on the individual-level treatment effect, considering student-level demographics and test scores. First, we used a host of pretreatment covariates to create matched groups. These covariates controlled for socioeconomic factors that have been shown to vary with both treatment selection and the outcome of academic performance. We then showed that the effects of the G&T program on academic achievement were strongest for students whose incoming grade 3 math/ELA scores were not within the 90th percentile. But after accounting for ethnicity, the gains were strongest and most consistent among the Black and Hispanic high achieving subgroup. Results from robustness checks (included in the online Appendix) further support and confirm the observed treatment effects on student outcomes. These additional analyses provide evidence that the findings presented here are robust and hold under different specifications or assumptions.

We estimate that the G&T program increases academic performance in middle school math/ELA proficiency by anywhere from a 0.1 to 0.3 standard deviations. These positive treatment effects are nearly consistent across subject areas and grade levels. No comparison resulted in a detrimental effect on student test scores. The findings are consistent with those of Card & Giuliano (2014, 2016), who found positive effects on achievement for nongifted high achievers who filled remaining seats in the class.

In our view, the larger academic impact that the G&T program had on the low achievers compared with high achievers brings up two important considerations. For one, the ceiling effect likely impacts the difference. Low-achieving students have more room to improve because their starting scores (grade 3 proficiency scores) are further from the maximum proficiency level. High-achieving students may make gains that are better realized through other academic outcome metrics, such as internal motivation or creative thinking. One mechanism behind this change could be a greater sense of academic self-concept and motivation that low achievers gain from participating in the G&T program. Because the low achievers scored within the 90th percentile on the gifted identification exam but not within the 90th percentile on the state proficiency exam, this score gap suggests a lack of academic motivation or self-concept rather than a lack of potential that the G&T program helps to untap. High achievers, in contrast, may have always fostered a high sense of academic self-concept. Although the G&T program offers them more challenge, it may not drastically shift their motivation and orientation toward challenge as it would for a low achiever.

It can then be argued that the larger gains made by Black, Hispanic, and White students compared with Asian students may be a result of the ceiling effect. Baseline grade 3 scores in the descriptive analysis showed that Asian students had higher incoming test performance than the other ethnicities. The larger academic gains made by Black, Hispanic, and White students partially reflect this ceiling effect.

However, the gains for Black and Hispanic students who were high achievers were greater than those of Black and Hispanic students who were low achievers. This suggests that a high-achieving Black or Hispanic student in grade 3 who did not get identified as gifted or who did not get a seat in the program would not reach as high or levels of middle school math and ELA proficiency as another grade 3 high-achieving Black or Hispanic student, similar on all the characteristics we controlled for (i.e., district, attendance, sex, poverty status, English language learner status, and disability status). High-achieving Black and Hispanic students might be left behind in ways that Asian and White high-achieving students are not when left out of G&T education programs. For this reason, the magnitude of the gains for Black and Hispanic students, especially the high-achieving students, is particularly noteworthy. Given that these students were the most underrepresented in the G&T program, the counterfactual for each student who made these middle school gains after participating in the G&T program is a high-achieving Black and Hispanic student who might be left unchallenged or unmotivated in a general education classroom.

Limitations

We focus our limitations on the unobserved covariates likely at play. One limitation of propensity score matching, as a quasi-experimental method, is that it relies on the assumption that all relevant confounding variables are observable and have been accounted for. In our study, while we achieved good overlap between the treatment and control groups, ensuring comparability and enhancing generalizability, unobserved variables could introduce bias into our estimates. Unlike randomized, controlled trials, which control for both observed and unobserved factors, propensity score matching only adjusts for what is observed.

This is particularly crucial in our study, where factors such as the quality of middle schools or access to test-preparation courses may have influenced both the likelihood of entering the G&T program and subsequent academic outcomes but were not captured in the dataset. Our models did not control for the quality of the middle school in the outcome model. Although we controlled for differences in grade 3 district (when the student received the treatment), it is likely that longitudinal middle school outcomes are also influenced by the differences in these schools. Simply controlling for grade 6, 7, and 8 district might not capture all the nuances in middle school quality, however, especially because students in NYC have the option to apply to selective middle schools. A middle school fixed effect at the school level, rather than the district level, might have helped to tease out some of these differences, although, again, the sample size restricted by the treatment assignment would not allow for a large enough treatment sample in each school. As another note, the dataset did not provide information on which G&T program classrooms were citywide or districtwide, preventing us from exploring the differences between these programs.

Additionally, the study identified treatment heterogeneity by ethnicities, but we hypothesize that treatment heterogeneity by district also was likely at play. For one, the district fixed effects in Model 1 showed differential academic performance in some districts over others. It is likely that these differences were related to socioeconomic factors such as percentage of minority students in a school, parental involvement, or economic need of a district. Even more, the interaction between treatment and ethnicity only appeared when we added district-level fixed effect. This suggests that the effect of the G&T program on academics varies across subgroups within different districts. The greater effects for Black and Hispanic students might be heterogeneous between specific districts, especially if socioeconomic factors interact with the G&T effect. In this study, our stringent treatment identification requirement did not yield a large enough sample size with which to evaluate a district-by-treatment interaction to compare with the district-by-ethnicity interaction.

The last limitation is the absence of exact scores on the gifted identification tests. Because the exploratory analysis revealed that many students in the G&T program did not score within the 90th percentile on the grade proficiency exams, the gifted identification tests seem to truly measure something outside of grade-level proficiency. In contrast, critics of G&T programs point to the fact that some students have access to test-prep courses to prepare for the admissions test, which might explain the gap between grade proficiency exams and gifted identification tests. Indeed, we would expect 10% to be in the 90th percentile on the gifted identification test, but in 2019, 20% of students reached the 90th percentile. These probabilities (perhaps due to self-selection of who takes the test) make the case for universal screening.

With or without access to scores on the gifted identification exam, we would have had to find a way to proxy for giftedness in the general education student group because these high-achieving students who might have scored within the 90th percentile never had a chance to take the test in the first place. For this reason, we feel that this is not necessarily a severe limitation and instead differentiates this analysis from other regression discontinuity approaches to gifted education effects, which focus on the cutoff scores of the test.

Implications and Future Directions

As discussions about balancing equity and excellence continue among educators and policymakers, empirical research measuring the effects of G&T education programs will become vital. Our study presents evidence of the positive treatment effects of participating in NYC's G&T program on academic outcomes, particularly for Black and Hispanic students. School districts across the nation also recognize persistent demographic and racial disparities in G&T education programs. Our study demonstrates that the G&T program yields differential positive gains for Black and Hispanic students—the very students who are less likely to receive these services in the first place.

A host of mechanisms might explain the G&T program effect. The gains made by students in the G&T program could be a result of a specific instruction style or quality of the teacher. Or they might be due to the simple environment change of being placed in a separate environment with like-minded peers. The outward labeling of gifted and the separate classroom environment could reasonably increase academic self-concept and motivation for students lacking a challenge in the general education classroom. Understanding the context-dependent nature of these programs, through qualitative work, is a necessary net step in teasing out the mechanisms that produce these gains. On our end, we surmise that district-level differences interact with student-level characteristics such as ethnicity to influence the effectiveness of educational interventions such as the G&T program. We specifically hope to capture local contextual factors (e.g., teaching quality, school climate, and community support) in future work to understand how the treatment interacts with ethnicity to influence academic outcomes.

Conclusion

The effect of the G&T program on student outcomes in NYC was positive and consistent over the course of the years 2010 through 2018. The 2018–2019 school year was the last year of analysis in this study because academic proficiency scores were not collected by the NYSTP in the 2019–2020 school year due to the COVID-19 pandemic, which resulted in widespread learning loss for students in Grades 3-8. Notably, this learning loss disproportionately affected students from low-income, predominantly minority communities (Kuhfeld et al., 2022; Fahle et al., 2023). Additionally, 2020 marked the last year the NYC Department of Education (DOE) offered testing to determine G&T program eligibility. The following year, the Panel for Education Policy, a governing body responsible for approving major contracts for the DOE, voted against renewing its contract with Pearson. This required the DOE to develop a new process to determine G&T program eligibility. The new process, beginning in 2021, required pre-K teachers to nominate students for G&T eligibility based on reviewing individual applicants against a list of gifted behaviors and characteristics.

The effects of the G&T program on those admitted after 2020 remains to be seen, but the results of this study suggest that, especially for the same group of students particularly hard hit by the pandemic, participating in the G&T program results in positive and consistent effects on academic achievement.

Acknowledgments

We acknowledge the NYC Department of Education, particularly the Research and Policy Support Group, for their invaluable collaboration in providing the data and supporting research on NYC public schools. Their encouragement and assistance were instrumental in the completion of this study.

Declaration of Conflicting Interests

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Open Practices

The data and analysis files for this article can be found at https://doi.org/10.3886/E209077V1.

ORCID iDs

Katherine J. Strickland https://orcid.org/0009-0003-9066-7510

Michael Gottfried https://orcid.org/0000-0002-4396-0576

References

Abdulkadiroğlu, A., Angrist, J., & Pathak, P. (2014). The elite illusion: Achievement effects at Boston and New York exam schools. *Econometrica*, *82*(1), 137–196. https://doi.org/10.3982/ecta10266

Aldrich, P. W., & Mills, C. J. (1989). A special program for highly able rural youth in grades five and six. *Gifted Child Quarterly*, *33*(1), 11–14. https://doi.org/10.1177/001698628903300102

Argys, L. M., Rees, D. I., & Brewer, D. J. (1996). Detracking America's schools: Equity at zero cost?. *Journal of Policy Analysis and Management*, *15*(4), 623–645. https://doi.org/10.1002/(SICI)1520-6688(199623)15:43.0.CO;2-J

Ascher, C., & Fruchter, N. (2001). Teacher quality and student performance in New York City's low-performing schools. *Journal of Education for Students Placed at Risk*, *6*(3), 199–214. https://doi.org/10.1207/S15327671ESPR0603_3

Betts, J. R., & Shkolnik, J. L. (2000). The effects of ability grouping on student achievement and resource allocation in secondary schools. *Economics of Education Review*, *19*(1), 1–15. https://doi.org/10.1016/S0272-7757(98)00044-2

Bhatt, R. R. (2009). The impacts of gifted and talented education. *Andrew Young School of Policy Studies Research Paper Series*, (09–11). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1494334

Brown, E. F., & Wishney, L. R. (2017). Equity and excellence: Political forces in the education of gifted students in the United States and abroad. *Global Education Review*, *4*(1), 22–33. https://files.eric.ed.gov/fulltext/EJ1137995.pdf

Bui, S. A., Craig, S. G., & Imberman, S. A. (2014). Is gifted education a bright idea? Assessing the impact of gifted and talented programs on students. *American Economic Journal: Economic Policy*, *6*(3), 30–62. https://doi.org/10.1257/pol.6.3.30

Card, D., & Giuliano, L. (2014). *Does gifted education work? For which students?* (NBER Working Paper w20453). National Bureau of Economic Research. https://www.nber.org/papers/w20453

Card, D., & Giuliano, L. (2016). Universal screening increases the representation of low-income and minority students in gifted education. *Proceedings of the National Academy of Sciences*, *113(*48), 13678–13683.

Delcourt, M. A., Cornell, D. G., & Goldberg, M. D. (2007). Cognitive and affective learning outcomes of gifted elementary school students. *Gifted Child Quarterly*, *51*(4), 359–381. https://doi.org/10.1177/0016986207306320

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, *84*(1), 151–161. https://doi.org/10.1162/003465302317331982

Dobbie, W., & Fryer, R. G., Jr. (2014). The impact of attending a school with high-achieving peers: Evidence from the NYC exam schools. *American Economic Journal: Applied Economics*, *6*(3), 58–75. https://doi.org/10.1257/app.6.3.58

Duffett, A., & Farkas, S. (2009). *Growing pains in the advanced placement program: Do tough trade-offs lie ahead?* Thomas B. Fordham Institute. https://files.eric.ed.gov/fulltext/ED505527.pdf

Edgerton, J. D., Roberts, L. W., & von Below, S. (2011). Education and quality of life. *Handbook of social indicators and quality of life research*, 265–296. https://doi.org/10.1007/978-94-007-2421-1_12

Fahle, E. M., Kane, T. J., Patterson, T., Reardon, S. F., Staiger, D. O., & Stuart, E. A. (2023). *School district and community factors associated with learning loss during the COVID-19 pandemic*. Center for Education Policy Research at Harvard University. https://edopportunity.org/docs/seda2023_documentation_20240130.pdf

Ferguson, L. M. (2022). *Factors predicting identification of giftedness resulting from universal screening* [Doctoral dissertation]. Liberty University. https://digitalcommons.liberty.edu/doctoral/3524

Gentry, M., Gray, A., Whiting, G. W., Maeda, Y., & Pereira, N. (2019). *Gifted education in the United States*. Purdue University. https://education.purdue.edu/geri/new-publications/gifted-education-in-the-united-states/

Gossett, E. C. (2022). *WHO IS ANOINTED? The psychological and social justice implications of gifted and talented programs in the United States*. Bard Digital Commons. https://digitalcommons.bard.edu/senproj_s2022/154

Haimovitz, K., & Dweck, C. S. (2017). The origins of children's growth and fixed mindsets: New research and a new proposal. *Child Development*, *88*(6), 1849–1859. https://doi.org/10.1111/cdev.12955

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Why public schools lose teachers. *Journal of Human Resources*, *39*(2), 326–354. https://www.jstor.org/stable/3559017

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199–236. https://doi.org/10.1093/pan/mpl013

John-Akinola, Y. O., & Nic-Gabhainn, S. (2014). Children's participation in school: a cross-sectional study of the relationship between school environments, participation and health and well-being outcomes. *BMC Public Health*, *14*, 1–10. http://www.biomedcentral.com/1471-2458/14/964

Kaul, C. R., Johnsen, S. K., Witte, M. M., & Saxon, T. F. (2015). Critical components of a summer enrichment program for urban low-income gifted students. *Gifted Child Today*, *38*(1), 32–40. https://journals.sagepub.com/doi/10.1177/1076217514556533

Kuhfeld, M., Soland, J., & Lewis, K. (2022). Test score patterns across three COVID-19-impacted school years.

Educational Researcher, *51*(7), 500–506. https://doi.org/10.3102/0013189X221109178

Liu, S., Rovine, M. J., & Molenaar, P. (2012). Selecting a linear mixed model for longitudinal data: Repeated measures analysis of variance, covariance pattern model, and growth curve approaches. *Psychological Methods*, *17*(1), 15. https://doi.org/10.1037/a0026971

Loveless, T., Parkas, S., & Duffett, A. (2008). *High-achieving students in the era of NCLB*. Thomas B. Fordham Institute. https://files.eric.ed.gov/fulltext/ED501703.pdf

Lu, Y., & Weinberg, S. L. (2016). Public pre-K and test taking for the NYC gifted-and-talented programs: Forging a path to equity. *Educational Researcher*, *45*(1), 36–47. https://doi.org/10.3102/0013189x16633441

Morgan, H. (2020). The gap in gifted education: Can universal screening narrow it? *Education*, *140*(4), 207–214. Retrieved from https://files.eric.ed.gov/fulltext/EJ1304434.pdf

Peters, S. J., Gentry, M., Whiting, G. W., & McBee, M. T. (2019). Who gets served in gifted education? Demographic representation and a call for action. *Gifted Child Quarterly*, *63*(4), 273–287. https://doi.org/10.1177/0016986219833738

Plucker, J. A., Burroughs, N., & Song, R. (2010). *Mind the (other) gap! The growing excellence gap in K–12 education*. Center for Evaluation and Education Policy, Indiana University. https://files.eric.ed.gov/fulltext/ED531840.pdf

Roberts, C., Ingram, C., & Harris, C. (1992). The effect of special versus regular classroom programming on higher cognitive processes of intermediate elementary aged gifted and average ability students. *Journal for the Education of the Gifted*, *15*(4), 332–343. https://doi.org/10.1177/016235329201500403

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38. https://doi.org/10.2307/2683903

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, *4*(1), 87–94. https://doi.org/10.2307/1391390

Sewell, C. J., & Goings, R. B. (2019). Navigating the gifted bubble: Black adults reflecting on their transition experiences in NYC gifted programs. *Roeper Review*, *41*(1), 20–34. https://doi.org/10.1080/02783193.2018.1553218

Steenbergen-Hu, S., Olszewski-Kubilius, P., & Calvert, E. (2020). The effectiveness of current interventions to reverse the underachievement of gifted students: Findings of a meta-analysis and systematic review. *Gifted Child Quarterly*, *64*(2), 132–165. https://doi.org/10.1177/0016986220908601

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical science: a review *journal of the Institute of Mathematical Statistics*, *25*(1), 1. https://doi.org/10.1214/09-STS313

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *174*(2), 369–386. https://doi.org/10.1111/j.1467-985X.2010.00673.x

Stuart, E. A., Lee, B. K., & Leacy, F. P. (2013). Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology*, *66*(8), S84–S90. https://doi.org/10.1016/j.jclinepi.2013.01.013

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516–524. https://doi.org/10.3102/0013189X18781522

van der Meulen, R. T., van der Bruggen, C. O., Spilt, J. L., Verouden, J., Berkhout, M., & Bögels, S. M. (2014, June). The pullout program day a week school for gifted children: Effects on social–emotional and academic functioning. *Child & Youth Care Forum*, *43*(3), 287–314. https://doi.org/10.1007/s10566-013-9239-5

Wai, J., & Lovett, B. J. (2021). Improving gifted talent development can help solve multiple consequential real-world problems. *Journal of Intelligence*, *9*(2), 31. https://doi.org/10.3390/jintelligence9020031

Wai, J., & Worrell, F. C. (2020). How talented low-income kids are left behind. *Phi Delta Kappan*, *102*(4), 26–29. https://doi.org/10.1177/0031721720978058

Worrell, F. C., & Dixson, D. D. (2022). Achieving equity in gifted education: Ideas and issues. *Gifted Child Quarterly*, *66*(2), 79–81. https://doi.org/10.1177/00169862211068551

Yaluma, C. B., & Tyner, A. (2018). *Is there a gifted gap? Gifted education in high-poverty schools*. Thomas B. Fordham Institute. https://eric.ed.gov/?id=ED592389

Authors

KATHERINE J. STRICKLAND is a researcher and lecturer at the University of Pennsylvania Graduate School of Education. Her research centers on the application of quasi-experimental research methods to education policy.

WENDY CHAN is an assistant professor in the Human Development and Quantitative Methods Division at the University of Pennsylvania Graduate School of Education. She specializes in applied statistical methods to improve generalizations from small studies in education.

MICHAEL GOTTFRIED is a professor of education policy at the University of Pennsylvania Graduate School of Education. He conducts research aimed at improving data-driven decisions in education, with a focus on absenteeism, early childhood programs, and students with disabilities.

JIEXUAN HUANG is a Master's student in the Statistics, Measurement and Research Technology program at the University of Pennsylvania Graduate School of Education. She is interested in using applied statistics to inform education policy.

DANIEL HILDRETH is the director of elementary school admissions at the New York City Department of Education. His work focuses on improving access to and assessing the impact of public-school programs in large urban districts.