12-2021

# Essential Programs and Services Report of Findings: Gifted and Talented Education

Lisa Morris

Amy Johnson

# Essential Programs and Services Report of Findings:
# Gifted & Talented Education

## Report to the Maine Department of Education

**Lisa Morris**

**Amy Johnson**

**Maine Education Policy Research Institute**

**December 2021**

**EPS Component Review Report of Findings:**

**Gifted & Talented Education Funding**

## Table of Contents

**EPS Component Review Report of Findings:**

**Gifted & Talented Education Funding**

Lisa Morris
*lisa.morris@maine.edu*

Amy Johnson
*amyj@maine.edu*

## Report Overview

This review of the Gifted and Talented (G&T) funding component within the Essential Programs and Services (EPS) education cost model begins with a summary of our most recent (2019) analyses of the G&T funding patterns and trends in Maine. Next we provide updated data (Part I) about G&T program implementation and enrollment in Maine public schools in FY2020 and FY2021 to confirm that trends from prior studies persist. We then provide a summary of the available research literature on the rationale and effectiveness of G&T programming in order to revisit the underlying tenets of this aspect of Maine's funding model and evaluate whether it remains consistent with the EPS model's overall goals of adequacy and equity.

## Background

Maine provides an annual allocation in the EPS formula for districts that successfully apply and receive approval from the Maine Department of Education for their Gifted and Talented (G&T) program plans. In FY2021, the total allocation to the 156 Maine SAUs with approved G&T programs was $13.49 Million, with amounts ranging from $147 to $427,570 per district (averaging $86,457).

Once a district's program is approved, they are provided funding that is based on the amount actually spent on G&T in the most recent fiscal year. This expenditure-based approach means that the amount of funding is not based on pre-established criteria such as the number of students participating in G&T programs or a fixed ratio for hiring staff. Rather, the allocations are directly related to prior spending, so that districts who spend more on G&T services receive larger program allocations (and vice versa). In general, expenditure-based funding raises concerns about equity because it tends to disproportionately benefit districts with greater ability to raise funds through local property taxes (i.e. wealthier communities).

The findings in the most recent MEPRI review of the Gifted and Talented funding component (2019)[1] corroborated those concerns about equity. Students identified as Gifted and Talented are disproportionately white, female, and *not* economically disadvantaged compared to the general population of Maine students. The underrepresentation of economically disadvantaged students is particularly stark, with such students representing 45.0% of all Maine students but only 21.1% of G&T identified students in 2018.

This pattern is not unique to Maine as detailed in the national scan included in the 2019 component review. States are beginning to respond to concerns about equity and the potential unequal opportunities for high-achieving learners of all backgrounds to benefit from supplemental G&T programs. Currently, 16 states do not provide any G&T funding, including New York and all of the New England states except for Maine. Others have shifted in recent years to make G&T programming optional and/or reduce financial support.

Because this recent and comprehensive G&T component review was based on 2018 student and expenditure data, it was decided to shift the focus of the current study rather than repeat the same analyses. This is because it was not desirable to analyze data from FY2020 or FY2021 due to the likely impact of the pandemic on program expenditures. Repeating the 2018 analyses with FY2019 data – just one year more recent -- was deemed unlikely to be fruitful and not the highest priority for investigation. Thus the current study instead provides only a brief overview of G&T program enrollment in FY2020 and FY2021. The remainder of the report is a detailed analysis of the existing empirical (quantitative) national research on the overall outcomes and impacts of providing separate pull-out services and programs to student who are identified as Gifted and Talented. This is intended to examine the underpinnings of Maine's funding model in light of other emerging trends, such as using a Multi-Tiered System of Support (MTSS) framework to serve students whose academic needs may not always be adequately addressed solely through differentiated instruction in the regular classroom.

[1] https://www.maine.gov/doe/sites/maine.gov.doe/files/inline-files/GiftedTalentedAppendixA_FinalApril2019.pdf

## Part I: Analyses of Participation in Maine G&T Programs (FY2020 and FY2021)

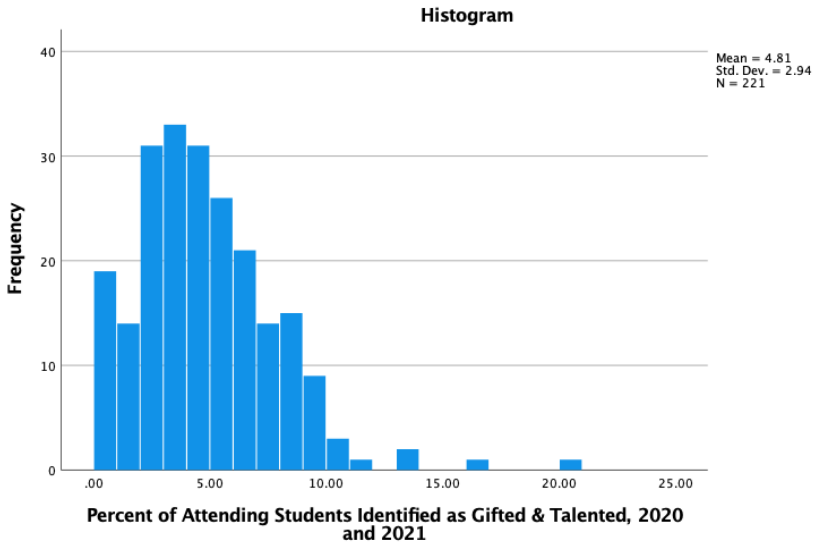
In academic years 2019-20 and 2020-21, just over half of Maine public school districts reported enrollment of students identified as Gifted and Talented (Table 1).

**Table 1. Descriptive Information on SAUs with Enrolled Attending G&T Students**

| | AY2019-20 | AY2020-21 |
|---|---|---|
| Number (%) of SAUs reporting G&T students | 112 of 207 (54%) | 109 of 206 (53%) |
| Total Number of G&T Students | 7,197 | 6,495 |
| Mean % Enrolled G&T students | 4.97% | 4.65% |
| Range of G&T Enrollment %s | 0.1% to 14% | 0.1% to 16% |

Table 1 also describes a large variation in the proportion of students identified as Gifted and Talented. Among districts that reported one or more G&T students, the overall identification rate was about 4.5% to 5% but rates in individual districts ranged from nearly zero (0.1%) to about 15%. Figure 1 depicts the G&T identification rates in districts with one or more reported G&T student.

**Figure 1. % of Attending Students Identified as Gifted & Talented, FY2020 and FY2021***




*Excludes SAUs that did not report any students identified as G&T

When analyzing the demographics of the districts that reported enrolled G&T students compared to those that did not report any students, we found similar trends to prior reports. Namely, there was a significant difference with respect to the proportion of students considered economically disadvantaged (Table 2).

**Table 2. Average Poverty Levels in Schools With and Without Enrolled G&T Students**

| | Average % Economically Disadvantaged Students | |
|---|---|---|
| | 2019-20 | 2020-21 |
| SAUs with reported G&T students | 41.7% | 39.1% |
| SAUs without reported G&T students | 49.3% | 45.6% |
| Statistically significant difference? | Yes (p=.004) | Yes (p=.009) |

These findings continue to raise concerns about the equitable allocation of funds through this component of the EPS cost model. The summary and conclusion section of this report provides additional recommendations for next steps based on these implications.

## Part II: Background and Existing Literature on Gifted & Talented Programs

### General background

Part II contains a comprehensive review of the existing quantitative research related to Gifted and Talented education programs in the United States. Guiding questions for the literature review included:

- Are gifted and talented programs effective?
- Do students assigned to gifted and talented programs achieve academic gains above and beyond what they would have achieved in regular educational programs?
- Do benefits of G&T participation depend on the type of G&T program? I.e. do some students benefit while others do not (by race, gender, economic dis/advantage, marginally gifted vs exceptionally gifted)?
- What are the socioeconomic characteristics of students selected for G&T programs?
- Do G&T programs help compensate for the lack of advantages and supports low-income families can provide their gifted children, or the lower expectations that some teachers have for economically disadvantaged students or students of color?

### *How G&T programs are theorized to work*

There is a lot of variety in how G&T programs are designed. Some schools create separate classrooms for students identified as gifted, some group students within the regular classroom and provide them more advanced content. Some are "pull-out" enrichment programs where gifted students are taken out of the regular classroom to participate in project-based, independent study. Other programs provide students with acceleration options, skipping a grade or taking more senior classes or college courses while still in high school.[2] G&T programs are presumed to help participating students offering more challenging and faster paced curricula commensurate with their abilities as well as more opportunities for independent work, creative and critical thinking. Students participating in G&T programs are also theorized to benefit from peer effects: their fellow G&T classmates are higher achieving, more academically motivated, etc. and they rise to the occasion, are more supported. Indirect effects of being in a class of

[2] https://www.nagc.org/resources-publications/resources/frequently-asked-questions-about-gifted-education

higher achieving peers result from the teacher teaching at a higher level, as long as the student can keep up.[3]

When budgets are tight there is a tendency to focus scarce resources on struggling students, the assumption being that gifted students are generally able to reach their full academic potential on their own, or that at least they will not fall irreparably behind. Because socioeconomically advantaged (white and/or rich) students are more likely to be identified as gifted, there's also the argument that their families will be able to provide compensatory supports and outside-of-school enrichment activities, and that G&T programs are merely adding to existing advantages as well as elitist and even segregationist.[4]

On the other hand, if all gifted students were identified as such and G&T programs are effective (and if they were to be particularly in lifting the academic achievement of disadvantaged students) then eliminating them will hit gifted poor, EL and racial minority students hardest because their families are less likely to have the means to provide compensatory support (tutors, enrichment activities, summer camps, etc.) or to send them to private schools with accelerated and advanced curricula.[5]

Proponents argue that if gifted students are not sufficiently challenged by regular curricula they will not reach their full academic potential, resulting not only in personal losses in earning power and life and career satisfaction but losses to society as well, the argument being that well-educated gifted students are more likely to go on to become exceptionally productive workers, effective leaders and innovators who invent valuable new products and services.[6]

"*We assert that aspiring to fulfill one's talents and abilities in the form of transcendent creative contributions will lead to high levels of personal satisfaction and self-actualization as*

---

[3] http://jhr.uwpress.org/content/early/2021/02/03/jhr.58.4.0920-11170R1.refs https://journals.sagepub.com/doi/10.1177/016235321103400302

[4] https://journals.sagepub.com/doi/10.1177/00169862211002535?icid=int.sj-abstract.similar-articles.2 High-Achieving Students in the Era of No Child Left Behind (brookings.edu) https://www.nytimes.com/2019/08/26/nyregion/gifted-programs-nyc-desegregation.html Gifted and Talented Programs Aren't the Problem - The Atlantic

[5] https://fordhaminstitute.org/national/commentary/we-are-squandering-talents-too-many-low-income-high-achievers How talented low-income kids are left behind - kappanonline.org https://www.nber.org/papers/w21519

[6] https://journals.sagepub.com/doi/10.1177/1529100611418056 https://www.nagc.org/resources-publications/gifted-education-practices/why-are-gifted-programs-needed https://www.nagc.org/myths-about-gifted-students

*well as produce yet unimaginable scientific, aesthetic, and practical benefits to society* (Subotnik, Olszewski-Kubilius, & Worrell, 2011)."[7]

The Maine Educators of the Gifted and Talented (MEGAT) website reads: "*The MEGAT Mission is to further the common good of gifted education in the State of Maine by supporting the development of programs for gifted and talented youth in Maine."*[8]

The belief that gifted and talented programs will yield social benefits and level the playing field by helping socioeconomically disadvantaged gifted students is a powerful policy argument even when educational resources are scarce.

*However, the empirical evidence in support of G&T program benefits is best described as mixed.*

While there are a bunch of studies showing a positive correlation between G&T program participation and academic outcomes including higher test scores, grades, and educational attainment it is not clear that these outcomes are due to G&T programs themselves.[9]

Participation in gifted programs is not randomly assigned. Students identified as gifted and assigned to G&T programs are by definition a selected group. The observed positive outcomes produced by correlational studies could be due to G&T programming (causation) or to higher levels of ability, motivation, confidence, socioeconomic advantages, and family support (selection).

Is their higher achievement the result of their participation in a G&T program or is it due to a mix of factors including natural ability, motivation and drive, confidence, exposure to high expectations (e.g., teacher presumptions regarding their ability), access to support (e.g., teacher and parent encouragement and assistance, tutors), participation in extracurricular enrichment activities (e.g., summer camps, internships) and lower levels of socioeconomic stress and distractions (e.g., stable housing, food security, parents with less work-family conflict, fewer responsibilities for younger siblings)?[10]

---

[7] https://journals.sagepub.com/doi/10.1177/1529100611418056

[8] http://www.megat.org/about.html

[9] https://ies.ed.gov/ncee/edlabs/regions/northeast/AskAREL/Response/72 https://www.nagc.org/resources-publications/gifted-education-practices/why-are-gifted-programs-needed https://files.eric.ed.gov/fulltext/EJ746290.pdf https://medium.com/age-of-awareness/do-gifted-and-talented-programs-work-aaee7dcaaa30

[10] https://www.nagc.org/blog/no-child-just-born-gifted-creating-and-developing-unlimited-potential

In other words, it is plausible that students who were selected to participate in G&T programs had higher academic skills, were more motivated, and/or provided with more support. These characteristics would have helped them to prosper in any school program.

If all students were tested and all gifted children were accurately identified and assignment to the treatment (G&T program) was random (i.e., there was no systematic difference between the two groups in terms of ability, motivation, access to support, etc.) then we could validly determine program effects by comparing the differences in academic outcomes between the two groups.

In the absence of random assignment researchers need to use statistical techniques to try and isolate causal effects by reducing endogeniety bias caused by omitted variables (unobserved and thus uncontrolled confounding variables), simultaneity (when a predictor variable is correlated with both assignment to G&T programming *and* subsequent academic achievement) and measurement error.

Even using lots of control variables, regular regression techniques (OLS) will produce biased results. If students assigned to G&T programming are of higher ability and more highly motivated, supported, encouraged, etc. the bias will be upward, making G&T program participation look more beneficial than it actually is.

While we found a couple of studies that benefited (more or less) from random assignment, most of the studies investigating the impact of G&T program participation used statistical techniques to try and control for bias. In the absence of random assignment to the treatment, researchers typically used the following approaches to reduce selection bias– instrumental variables, regression discontinuity models, propensity score matching and student fixed effects.

***Primer on quantitative research methods and their limitations***

When there is a test (IQ or achievement) or some other type of score (i.e., a continuous index of tests and other factors) used to select which students are assigned to G&T programming researchers are able to use regression discontinuity techniques to try and isolate causal program effects. This approach compares students just above the eligibility cut-off to those who are just

https://www.journals.uchicago.edu/doi/abs/10.1086/444275 https://us.sagepub.com/sites/default/files/upm-assets/38607_book_item_38607.pdf

below the cut-off, the assumption being that the two groups of students are very similar in terms of both ability as well as other unobserved/able characteristics and factors that impact subsequent academic outcomes. Researchers must decide how far out from the eligibility cut-off to go: If they use observations from more points out from the cut-off their sample is larger and therefore, they are more likely to find a statistically significant difference but the further out the move the more they risk introducing bias (i.e., the two groups of students become less similar). Parametrically, researchers use standard linear regression with the dependent variable the subsequent academic outcomes of interest (e.g., post-participation achievement test score, % who went on to college) with the predictor variables of interest being the index score (i.e., the test or index score used to assign students to G&T) and a dummy variable that indicates whether the student's score is above the cut-off (the coefficient on this variable is the estimate of the G&T program effect).

It is often the case that not all the students who score above the cut-off participate in the G&T program (and some who score below do). Participation in G&T programs is voluntary, and some students opt out. In some instances, there are limited G&T program slots and so not all qualifying students are able to participate. In some cases, students who score below the cut-off are able to participate because their parents or teachers advocate on their behalf. In this case researchers use what is called "fuzzy" regression discontinuity design (RDD), a two-stage estimation technique in which in the first stage produces and estimates the likelihood the student participated based on their test or index score and a dummy that indicates whether their score is above the cut-off. The estimated likelihood is then used as a predictor in the second stage where you estimate the effects of program participation on achievement outcomes.

To test the validity of using RDD, researchers should provide a comparison of the two groups of students using the available data to see if there are "discontinuities" between the two groups in terms of characteristics and factors prior to the G&T program participation (or not). The only discontinuity you want to see between the two groups is on the dependent variable (in which case you can conclude there is a causal program effect). In other words, they should compare the two groups in terms of observable data like student demographics and prior achievement test scores. If there are no statistically significant differences (or no discontinuities in graphical plots) then the researchers can conclude the two groups are similar enough to isolate an estimate of causal effects of G&T participation. If they find significant differences, they need

to at least include these variables as controls in their estimation models. If they find too many differences between students just above and just below the eligibility cut-off, then RDD is not appropriate.

While RDD is generally more robust than instrumental variables techniques (described below) a possible source of bias with RDD occurs when there are spillover effects (which "contaminate" the comparison group). Negative spillover occurs when the comparison group (the students below the cut-off) suffer in ways that negatively impact their academic performance (e.g., upset about not making the cut-off; teachers in regular classrooms teach to lower ability students once G&T students are removed from the classroom and these higher ability students left behind in the regular classroom suffer from not being challenged). Negative spillover effects will inflate the estimated effects of G&T class participation (i.e., the comparison group will be made worse off making the students above the cut-off in the G&T classroom look better even if there were no positive program impacts). There is also the possibility of positive spillover effects which will have the opposite effect on estimated program effects (i.e., will make the G&T program look less effective than it actually is). This might occur if students below the cut-off benefit from now being the top students in the regular class because the slightly higher achieving students went to the G&T class (confidence boost; teacher adjusts his/her pace and level in ways that better suit their needs; their parents provide more support – tutors, outside school enrichment - perhaps in hopes they make the cut next year or to at least make up for the fact that they aren't getting extra from the G&T program). If average achievement scores for the students just below the cut-off decline that is a sign there may be negative spillover effects; if they increase, there may be positive spillover effects.

The most obvious limitation of RDD is that it examines program effects only on the marginal students (i.e., those just above the eligibility cut-off); it does not enable accurate analysis of the effects on more highly gifted students (because the further you move away from the eligibility cut-offs, the less similar become your comparison groups).

Another way of addressing the potential for endogeneity bias is to use *instrumental variables.* Instead of using a G&T participation (yes/no) variable to estimate the program effect, researchers use some other variable – referred to as the instrumental variable (IV) - that is correlated with the treatment (G&T program) but not directly correlated with the outcome

(subsequent academic achievement), i.e., the IV affects academic achievement only indirectly through G&T program participation. The estimation technique is two-staged. In the first stage the IV is used to predict the probability of being placed in a G&T program. The estimated probability of program participation is then used as a predictor variable in the second stage where you estimate the effects of program participation on achievement outcomes. The rub with this technique is that good IVs are hard to come by and there is no direct way to measure if they are good or bad (i.e., there is no way to statistically test whether the IV is both a strong predictor of G&T program participation and only correlated with subsequent academic outcomes via the G&T program). Choice of IVs is based on theory and logic. At best, researchers in certain cases can use statistics (like the F statistic or R-squared) associated with the first stage equation to assess the strength of the IV in predicting program participation. Weak IVs can lead to both biased and imprecise estimation of program effects. Note, however, a strong IV could still be biased (i.e., it may be strongly correlated with the participation but also correlated with the achievement outcomes used to assess program effectiveness).[11]

*Propensity score (PS) matching* is another technique used to try and reduce endogeneity bias. Using logistic or probit regression (usually), a propensity score (the probability a student participated in a G&T program) is estimated for all students using variables correlated with program participation available in the data. The resulting score is used to create matched comparison groups (i.e., very similar in terms of their PS) of G&T program participants and non-participants. There are a number of matching techniques (nearest neighbor closet score; with and without replacement, etc.) and researchers evaluate the quality of the matching by comparing the two groups on the covariates (differences of means t-tests, comparing of distributions graphically, etc.). Program effects are estimated by comparing the means on the outcome variable between the two groups, using regression to control for the variables that were "unbalanced" (i.e., on which the two groups did not match well) and in some cases including the PS as a covariate. The limitation of this technique is that is that the two groups are matched based on observable factors (i.e., variables in the available data); there remains a risk of bias

[11] https://www.statisticshowto.com/instrumental-variable/ https://www.nber.org/system/files/working_papers/t0284/t0284.pdf http://economics.mit.edu/files/15326

related to unbalanced unobserved/unmeasured confounding factors (e.g., family support, motivation, etc.).[12]

Finally, when researchers have access to longitudinal data with repeated measures on the same students (i.e., panel data) they can use student-fixed effect analysis. The analysis of repeated measures on the same student allows for the control of bias resulting from time-invariant unobserved confounding factors. Instead of comparing participating and non-participating students, this approach basically uses the individual student as their own control, comparing achievement outcomes during the years they received G&T services to the years during which they did not receive G&T services. To the extent that unmeasured confounding factors like motivation and parent support remain fixed (time-invariant), selection bias is controlled. The obvious limitation of this approach is that some unobserved confounding factors are not time-invariant. For example, wealthy parents may invest more time and resources to provide a child with outside school support once the child has been assigned to a G&T program (or maybe when they do not make the cut-off, hoping to help them do so in the next school year). On the other hand, parents with less resources may provide less support once their child is assigned to G&T (resources are scarce and they assume the G&T program will meet their child's educational needs). Another limitation of this technique is that the reduction in bias (by estimating effects of repeated within-student measures) may come at the expense of precision (i.e., it will be harder to achieve statistical significance), particularly if there is little change in observed program participation over time (i.e., if the observation period is relatively short and most of the students in the sample are either always in G&T or always out of G&T).

**Summary and Critique of Available, Rigorous Studies**

We conducted a search for studies examining the effects of G&T program participation academic achievement. Below is a summary of these studies. We do not include correlational studies that do not attempt to control for bias and isolate causal effects. We begin by describing the more rigorous studies, those that were able to benefit from random assignment, use RDD or panel data and student fixed effects. After that we include those using the less robust methods: instrument variables and propensity score matching. Overall, the results are mixed with some

[12] https://link.springer.com/article/10.1007/s12350-017-1012-y

studies finding positive effects and others finding no effects, some finding benefits for disadvantaged students, others finding net gains only for advantaged students.

The most rigorous U.S.-based studies were conducted by Bui, Craig and Imberman (2014) and Card and Guiliano (2014 and 2016). Both use data from one district (one concern about studies using national data is that program variability is masking program effects from higher quality G&T programs), employ RDD and run an extensive set of validity checks and sensitivity analyses. Bui et al also are able to confirm the results obtained using RDD by analyzing data from a lottery that more or less replicates random assignment.

***Card and Giuliano (2016) Can Tracking Raise the Test Scores of High-Ability Minority Students?***[13]

**Summary:** Except for studies employing random assignment or benefiting from lotteries, this is the strongest study methodologically. It is published in one of the top peer-reviewed journals in economics by a Nobel prize-winning economist. Taking advantage of extensive student level information, they use two different analytic approaches to test for the effects of assignment to separate G&T classrooms, RDD and a between school/cohort analytic design. Using data on multiple cohorts of students from one large urban district, they estimate the effect of being assigned to a separate G&T classroom on high-achievers (students who did not pass the IQ test cut-off but who were top-ranked according to the previous year's achievement tests). Using RDD, which examines the effects of G&T program participation on marginal, just-above-the-cut-off students as compared to just below the cut-off students who remain in regular classrooms, they find strong positive effects for students of color but not for white students (poor or not). Their findings persist when they use a between school/cohort analytic design as well as to a battery of alternative model specifications, validity checks and sensitivity analyses, including tests for potentially biasing spillover effects. They conduct additional analysis using student level data to show that racial minority students have lower achievement scores in the 3rd grade (prior to participating in the G&T program) than white students with the same cognitive ability (as measure by NNAT in 2nd grade) and that placement in a G&T class all but eliminates the achievement gap. They speculate that the reasons the program has valued-added impacts for minority students and not white students relates to minority students obtaining more support and higher teacher expectations in the G&T classroom: higher ability minority students do better in G&T classrooms because they are more supported (additional analysis also showed a reduction in absences and suspensions), have higher ability and more female peers (less flack for "acting white") and that teacher expectations are higher (in another paper they showed that before universal screening was adopted, teachers in the same district systematically under-refer black and Hispanic students for G&T programs) and that students rise to the occasion.

[13] https://www.aeaweb.org/articles?id=10.1257/aer.20150484

- **Data, Method, Sample:** The data used in this study come **from a large urban school district** in the U.S. Starting in 2004 the District required elementary schools to set up separate classrooms for gifted students (identified using IQ tests: non-disadvantaged students with IQ scores ≥130; subsidized lunch participants and English language learners with IQ scores ≥116) in fourth and fifth grades, with any open seats allocated to students with the highest scores on the previous year's achievement tests. Since most schools have only a handful of gifted children per grade, and class sizes are maintained at 20-24 pupils, most gifted classrooms in the district contained a mixture of high IQ students and high-achievers (the students who didn't make the IQ test cut-off but they scored in the top ranks of achievement tests). In this paper, the researchers focus on the effects on high achievers - the students who were assigned to the separate G&T classrooms based on their achievement test scores (in another study, described below, they look at the effects on students assigned to G&T classrooms based on their IQ tests). The researchers conduct two types of analyses: (1) **using regression discontinuity design (RDD)** based on the eligibility rules for the 4$^{th}$ grade G&T classroom, which estimates the effect of being in a separate G&T classroom by comparing students just above the achievement score cutoff (eligible for the G&T class) to those just below (remain in regular classrooms). They estimate separate models for white and minority students and examined whether there was evidence that teacher quality and peer characteristics impacted the estimates of program effects. (2) **they used a between school/cohort design to create a counterfactual scenario in order to test for spillover effects** (i.e., do students ranked just below the cutoff do worse when there is a separate G&T classroom in their school to which the higher ranked students are assigned compared to when there is no separate G&T classroom and the higher ranked students remain in the regular classroom). Note: it is important to check for spillover effects; negative spillover will inflate estimated effects of G&T class participation (i.e., the comparison group will be made worse off making the students above the cut-off in the G&T classroom look better even if there were no positive program impacts); positive spillover effects will produce a downward bias on estimated effects (if students below the cut-off benefit from being the top students in the regular class because the slightly higher achieving students went to the G&T class). The school/cohort design also enables them to

test how different ranks of students performed on 4th grade tests when there is a G&T classroom option and when there is not. This approach enabled them to test whether effects differ by student rank (they ranked students according to their 3rd grade test scores, 1-20 were high ranked; 25-44 were low ranked) and as a confirmation of the RDD analysis.

- They tracked all students (N=4,144) who completed 3rd grade in the years from 2008 to 2011 and entered the 4th grade the following year at one of the district's 140 elementary schools. They follow the students into the 5th and 6th grades as long as they remain in the district. To construct their RDD sample they selected 4th graders in schools with a separate G&T classroom (i.e., the school had at least one student identified as gifted using IQ tests) and restrict their sample of students to those who scored within 10 points on either side of the eligibility cut-off. For the between-school analysis (schools with and without separate G&T classrooms), there were 4,767 students who they ranked 1-20 (who were likely to move to a G&T classroom if their school had one) and 5,016 who ranked 25-44 (who were likely to remain in the regular class regardless of the availability of a separate G&T class). All models include student controls (age, gender, race, median income for the zip code in which they live) as well as school and year fixed effects (dummies for all years and schools).
- **Results:** They first confirm the validity of using RDD by showing that students just below and above the achievement score cut-off are very similar (according to 3rd grade reading score, 3rd grade math score, 2nd grade NNAT score). They also show that there are no differences in attrition from their sample (a threat to their analysis would result if students assigned to the separate G&T class were more or less likely to remain in the district). Their RDD results show positive and statistically significant effects from being assigned to a separate G&T classroom on 4th grade reading and math test scores (but not on writing test results); the magnitude of the estimated effects are quite similar with and without student controls (gender, age, race and median household income for their zip code) or school fixed effects (school dummy variables). They estimate the net positive effects to be in the range of 0.3 standard deviations. They estimate effects separately for white and racial minority students and find that almost all the positive effects are

accruing to students of color: they find no statistically significant differences in test scores between white students above and below the cut-off. They check whether this is the result of white students above the cut-off topping out on exams – i.e., scoring so high in 3rd grade they can't go any higher – but find no indication of this (only 2% of white students achieved the top score in reading and 10% earned the top score in math). As a double check on the topping out possibility they estimate program effects using Tobit regression models (which account for censoring that might happen if a high enough number of white students were topping out) and confirm the linear RDD results. They found similar sized impacts for poor and non-poor students of color, and somewhat larger effects for male students compared to female students. When they compared poor and non-poor white students, they found no significant effects for either group. The positive effects of program participation on minority students appears to persist: in the 5th grade the effects of G&T classrooms on reading scores are positive but insignificant, larger for math test outcomes though only marginally significant; in the 6th grade, the results are more clearly positive, with marginally significant impacts on reading test scores of about 0.2 standard deviations and about 0.4 standard deviations for math. They also run a bunch of models looking into the effects of classroom-level differences: While they found no significant differences in teacher effects, they did find that students in G&T classes had a higher percentage of female peers, peers with higher average scores, and peers slightly fewer suspensions and that these differences had small effects: no effects on reading outcomes and small effects on math achievement.

- The results of their between school/cohort analysis indicate that the presence of a separate G&T class at the school has a positive impact on the top 20 students (ranked 1-20 based on 3rd grade test scores) and no effect on students ranked lower (24-44) and that the effects of being in a G&T classroom are smaller for highly ranked students and larger for those who are lower ranked (just above the cut-off), from which they conclude that students in G&T classes are not suffering from mismatch or invidious comparison. They also find no evidence of negative or positive spillover effects: lower ranked students in schools with a separate G&T class (i.e., they get left behind because their higher ranked peers are moved out of the regular classroom) do not do any better or worse than lower

ranked students in schools without a separate G&T classroom (i.e., they are together with the higher ranked students in the regular classroom).

- Additional analyses using student level data show that minority students have lower achievement scores (3rd grade) than white students with the same cognitive ability (measure by NNAT in 2nd grade) and that placement in a G&T class all but eliminates the achievement gap.
- **Discussion:** This is a very rigorously conducted study. It is published in one of the top peer reviewed journals in economics by a Noble prize-winning economist. Taking advantage of extensive student level information, they use two different approaches, and both show positive program effects but only for minority students. They also provide compelling evidence that the program closes the achievement gap between white and racial minority students. They speculate that the reasons the program has valued-added impacts for minority students and not white students relates to minority students obtaining more support and higher teacher expectations in the G&T classroom: higher ability minority students do better in G&T classrooms because they are more supported (additional analysis also showed a reduction in absences and suspensions), have higher ability and more female peers (less flack for "acting white") and that teacher expectations are higher (in another paper they showed that before universal screening was adopted, teachers in the same district systematically under-refer black and Hispanic students for G&T programs) and that students rise to the occasion.

***Card and Giuliano (2014) Does gifted education work? For which students?[14]***

**Summary:** A more extensive version of the study conducted for the National Bureau of Economic Research (NBER) reports no positive benefits from being placed in a separate G&T classroom for students who were assigned based on IQ testing. Positive program effects only accrue to racial minority students assigned based on the previous year's achievement test results. They find no significant impact on subsequent math and reading achievement tests for the students assigned to the separate G&T classrooms based on their IQ scores – neither economically advantaged students or FRPL eligible or EL students (although they found a marginally significant positive benefit on the writing test scores for boys and students attending schools with high rates of FRPL eligibility). The authors speculate that students selected to participate in G&T classroom on the basis of previous standardized test scores may have a combination of cognitive abilities and non-cognitive traits (i.e., not captured IQ tests) that cause them to do even better on achievement tests after participating in the G&T program (traits that high IQ students do not have, such as a willingness to meet social expectations, attention to task).

- **Data, Method, Sample:** The data used in this study come from the same large urban school district in the U.S. used above. Starting in 2004 the District required elementary schools to set up separate classrooms for gifted students in fourth and fifth grades, with any open seats allocated to students who didn't make the IQ cut-offs (see below) but were among the top-ranked according to the previous year's achievement tests. Using the same methods in their published paper (described above) - RDD and a bunch of alternative specifications to conduct sensitivity analyses - they examine the effects of being placed in a separate G&T classroom on three groups of students: Plan A: 2,679 non-disadvantaged students with IQ scores ≥130; Plan B: 4,472 subsidized lunch participants and English language learners with IQ scores ≥116; and Plan C: 4,144 students who miss the IQ thresholds but scored highest among their school/grade cohort in state-wide achievement tests in the previous year, 2,098 of whom are not poor and 2,046 of whom are poor.
- **Results:** They find no significant impact on subsequent math and reading achievement tests for the students assigned to the separate G&T classrooms based on their IQ scores –

[14] https://www.nber.org/papers/w20453

neither economically advantaged students or FRPL eligible or EL students (although they found a marginally significant positive benefit on the writing test scores for boys and students attending schools with high rates of FRPL eligibility). But, as described in the published paper (see above), they do find positive effects for minority students placed in G&T classrooms based on previous achievement test scores.

- **Discussion:** They provide possible explanations – for why they find no program effects for students identified as gifted using IQ testing: (1) the statewide annual standardized test may not be capturing the effects of G&T participation on these students (because these students are already performing well academically, teachers in G&T classrooms may be teaching other stuff that's not measured by the achievement tests); (2) students selected to participate in G&T classroom on the basis of previous standardized test scores may have a combination of cognitive abilities and non-cognitive traits (i.e., not captured IQ tests) that cause them to do even better on achievement tests after participating in the G&T program (traits that high IQ students do not have, such as willingness to meet social expectations, attention to task); (3) students placed in G&T classes based on their IQ tests results may be topping out on achievement tests while those placed based on achievement tests still had room to grow (this argument works for the economically advantaged Plan A students but not for the disadvantaged Plan B students, whose 3$^{rd}$ grade achievement test scores were not as high as Plan A's and so they had room to grow by their 4$^{th}$ grade achievement test) (4) small fish deeper pond effect (marginal students – those with IQs just above cut-off – may experience invidious comparison effects – they went from being the top of their class to the bottom of the G&T class).
- They also used survey data that measured student responses to: my teacher(s) believe I can succeed; my teacher(s) answer my questions in a way I can understand, I enjoy learning at my school, I am accepted and feel like I belong at this school. To the extent that 4$^{th}$ graders self-report reliably (a possible limitation here)….again using the RDD to the compare students on either side of the cut-off they find that among Plan A students (not poor, IQ>129) just above the cut-off were more satisfied with their learning environment compared to those below the cut-off and left back in the regular classroom while Plan B students (poor and/or IQ>115) above the cut-off were no different in their

average responses compared to those just below the cut-off. Plan C students (missed IQ but scored high on 3rd grade achievement tests) just above the cut-off expressed about the same level of satisfaction with their learning environment as the comparison students just below the cut-off and in regular classrooms.

***Is Gifted Education a Bright Idea? Assessing the Impact of Gifted and Talented Programs on Students (Bui, Craig, Imberman, 2014)***[15]

**Summary:** This is another rigorously conducted study using a lot of student level data from one large urban school district in the southwestern U.S. It's actually two studies in one. The first study employs a RDD, which examines the effects of G&T program participation on students just above the eligibility cut-off to non-participating students just below the cut-off. The second study tests for program effects by comparing achievement outcomes of students randomly assigned to a specialized G&T magnet school compared to those who remain behind in G&T programs in regular schools. Neither study finds evidence of positive gains to program participants, even the magnet school lottery which targets more highly gifted students. Their initial findings persist under a number of alternative model specifications and sensitivity analyses. While their data come from one district there is variation across schools in terms of how their G&T programs is implemented (some offer separate G&T classrooms; some offer pull-out enrichment activities; some offer advanced course material called Vanguard, etc.). They conduct additional analysis to see if G&T participation effects differ by type of G&T program (whether or not the student attends a G&T magnet, whether or not their school offers the more advanced curriculum classes vs receiving additional materials within the standard AP classroom) and find no evidence of heterogeneous effects by G&T treatment type. They also examine whether estimated program effects differ in terms of the "intensity" with which a school provides G&T programming and find no clear pattern by program intensity. They also test whether program effects differ by student ability (to check if it's the very marginal students just above the cut-off who aren't making any gains) and find no difference in achievement results between the lower ability G&T participants and students just below the cut-off who remained in regular classrooms (i.e., does not appear that their finding of no program benefits is the result of very marginal students suffering from invidious comparison issues). They also test whether program effects differ by gender, race, economic status, or prior gifted status and find no statistically significant program effects by student type. A limitation of the RDD study is that they didn't rule out spillover effects; positive spillover effects accruing to students just below the eligibility cut-off – perhaps they do better when G&T participating students leave the regular classroom - could

[15] https://www.nber.org/system/files/working_papers/w17089/w17089.pdf
https://www.aeaweb.org/articles?id=10.1257/pol.6.3.30

downwardly bias estimates of program effects, although given that there is variation in G&T program type across schools it is less likely that positive spillover is masking a global program effect. The lack of additional benefits from attending a G&T magnet school may be the result of weak treatment effects (i.e., the G&T program at the G&T magnets may not have been that much more intensive than what lottery losers received back in the regular schools).

- **Data, Sample and Method:** This study is really two studies in one. The first study uses a **regression discontinuity design (RDD)** and compares students just above and below the G&T program eligibility cut-off. The second study takes advantage of a **lottery** that **randomly assigned** gifted students to specialized G&T magnet schools and compares achievement outcomes of lottery winners to lottery losers (gifted students left behind in regular G&T programs). The data they used are from **one large urban school district in the southwestern U.S.** where all 5th graders are evaluated to determine eligibility for G&T services starting in the 6th grade. The district generates a matrix score for each student based on standardized test results (SAT, Aprenda, NNAT), average course grades, teacher recommendations, and indicators for socio-economic status (5 extra points for being poor, 3 extra points for students of color). Students above the cut-off are eligible for G&T services. The regular **G&T program varies** across schools but, according to administrators surveyed by the researchers, most schools provide a more advanced curriculum (called Vanguard) that delves deeper into subjects and are geared towards developing creative and critical thinking as well as analytical skills although some provide students with other options (e.g., additional material within the regular AP class). In some schools G&T students are in their own classrooms; in others they are in mixed classrooms.
    - **The RDD study:** Because not all students above the cut-off participate in the G&T program and some below the cut-off do, they use a "fuzzy" RDD. After restricting their sample to a 15-unit band above and below the cut-off there are 4,055 students in the 7th grade who were evaluated for G&T in the 5th grade (1,509 were eligible for G&T and 2,546 were not). They test for the validity of RDD: they test to see how similar students above and below the cut-off are in observable characteristics (including race, gender, EL status, special education status, FRPL eligibility, achievement test scores from 5th grade in math, reading,

science and social studies, any missing data, teacher characteristics, school size, %FRPL) and find only one statistically significant difference at the student level ($5^{th}$ grade math scores). They "fix" this by adding $5^{th}$ grade scores as a control variable in their models. They make sure students above the cut-off are actually receiving "treatment" and find that they were more likely to attend a G&T magnet school, take advanced Vanguard courses and their peers are higher achieving (their estimates show that G&T students were not more likely than the students just below the cut-off to have a G&T certified teacher or a teacher with an advanced degree). They find no difference in achievement test scores in reading, language, social studies, science, or math using Stanford Achievement Test between students just above the eligibility cut-off (G&T participants) and those just below the cut-off (non-participants). They conduct additional analysis to see if G&T participation effects differ by type of G&T program (whether or not the student attends a G&T magnet, whether or not their school offers the more advanced curriculum classes vs receiving additional materials within the standard AP classroom) and find no evidence of heterogeneous effects by G&T treatment type. They also examine whether estimated program effects differ in terms the "intensity" with which a school provides G&T programming (measured in terms of achievement scores of classmates, percentage of students that take advanced curriculum classes and percentage of students identified as gifted). They find no clear impact of program intensity (G&T program effects are mixed – some positive, some negative – at both high and low intensity schools). Their findings stand up to additional sensitivity analyses including school fixed effects (to control for unobserved school level factors), alternative cut-offs for G&T eligibility, and using different sized bandwidths around the cut-off. They test for ceiling effects (because if the G&T students have no room to improve program estimates will be biased downward and make it look like the G&T program is less/not effective) and conclude this is not what is causing them to find no positive program effects (the majority of G&T students had room to improve). Their results do not change if they include a baseline control (prior standardized test scores in the tested subject). They also test whether program effects differ by

gender, race, economic status, or prior gifted status and generally find no statistically significant program effects by student type (except for non-poor students for which there is a marginally significant small negative effect on math scores; and students who were previously identified as gifted, for whom there is also a small negative effect on math scores).

- **The lottery study:** Conditional on meeting the district-wide GT eligibility requirements and completing an application, students were randomly offered admission to the district's premier magnet schools (versus staying behind in regular schools and receiving regular G&T services). In this part they are able to examine the effects of G&T program participation over the full range of G&T students (at least those that apply to get into the lottery, anyway), not just marginal students just above the cut-off (in fact, they mention that the students applying to magnet schools were more exceptionally gifted, scoring significantly higher on pre-tests than other gifted students). They compare the achievement test results of students who win the lottery and attend one of the magnet GT programs to those who lose the lottery and either attend a neighborhood GT program in the district, a magnet school based on a different specialty, or a charter school. In order to try and control for selection bias related to whether or not an eligible student applied or not, they use a 2SLS model to estimate program effects conditional on applying for admission to magnet program. Control variables included race, gender, special education status, EL status, FRPL eligibility and the baseline 5$^{th}$ grade achievement test score. They find no differences in math, reading, social studies scores between the lottery winners attending G&T magnet schools and lottery losers who remained behind in regular schools; they do find a positive effect on science achievement test scores (around 0.28 standard deviations higher) but it does not always hold its significance under different model specifications and sensitivity tests. To control for higher attrition among lottery losers (perhaps parents change schools when their child does not make it into the G&T magnet) they run weighted regressions (weighted by the inverse of the estimated probability of remaining in the data) and their initial results stand.

- **Discussion:** This a rigorous study. The researchers used two different approaches – RDD and lottery sample – and find, for the most part, confirmation of no program effects. All their baseline OLS models (which don't control for endogeneity bias) estimated positive and statistically significant program effects but the size of their RDD estimated program effects are all much smaller and never statistically significant. This indicates that achievement benefits observed using regular OLS are primarily explained by unobserved student traits and other factors (motivation, support, etc.) and reflect upward bias caused by selection. While RDD studies are inherently limited in that they only look for program effects for marginally gifted students (those just above the eligibility cut-off) the researchers note that the marginal students in the RDD sample still represented a range of giftedness: students right at the eligibility threshold ranged from 45 to 97 in national percentile rankings in reading and between 55 and 98 percentiles in math. Also, their lottery sample included a broader range of gifted students and skewed towards the higher end of gifted and presumably academically motivated (i.e., motivated enough to apply to get into the G&T magnet in the first place). They confirm significant "treatment" effects for their RDD sample (students above the cut-off were more likely to attend a G&T magnet and/or take advanced courses, they did have higher-achieving classmates) to make sure their results were not being drive by no real differences in "treatment". The treatment effects for their lottery sample appear smaller: students attending the G&T magnet schools did have stronger peers but that they were as likely as lottery losers to take advanced (Vanguard) courses. The authors speculate as to why they did not find much by way of significant program effects: (1) if parents of students who did not make the cut-off invested in more support for their children (perhaps hoping they'd test higher and get in next year), for example, by hiring a tutor or outside school enrichment activities (2) The school district may have set eligibility cut-off too low (research shows getting into a G&T program keeps parents from removing their children from the school), which would mean that more marginal students would be in the G&T program and they either bring down the achievement scores and/or cause teachers to teach a less rigorous curriculum (3) since they found that G&T participating students had academically stronger peers (at least those in separate classrooms or at the G&T magnet), maybe the marginal just above-the-cut-off students suffered from invidious comparison (their confidence is shaken; they were the top students in the regular classroom and now they are the bottom students in

the G&T program) or the marginal students suffer because the teachers teach to the average or higher ranked students' level and they can't keep up. I would add to this the possibility of weak treatment effects, especially for the lottery sample: while students attending the G&T magnet schools had higher achieving peers than the lottery losers, they weren't more likely to take advanced Vanguard courses.

***Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya (Dufflo, Dupas, and Kremer, 2011)*[16]**

**Summary:** While this study uses data from Kenya which has a much different education system than the U.S. it is included because of its rigorous methodology and use of experimental data. It is also important because it evaluates a specific type of G&T programming: tracking. In addition to regular regression techniques (which would be sufficient because they have experimental data), they also conduct RDD comparing two students – the one just above the eligibility cut-off and the one just below - from each of the 60 randomly selected schools that employed tracking. Their results suggest that all students benefit from tracking, both those in the upper and lower parts of the ability distribution. They found that these gains persisted one year later, after the experimental tracking program was concluded. They also show that the results were similar for boys and girls (although girls got a bit of a larger boost from tracking on their math test results). They also find interesting evidence suggesting that teacher type matters: They show differences in the benefits of tracking were influenced by whether the student had a contract teacher (no job security) or a civil service teacher (tenured; promotion based on seniority not performance), with all students benefiting if their teacher was a contract teacher and only initially high scoring students benefiting from tracking if their teacher was a civil service teacher (they provide additional analysis and some speculation that suggests that this is because tracking induced civil service teachers to increase their effort when they were assigned to teach upper track students but not when they were assigned to teach lower track students while contract teachers – hoping to be hired as civil service teachers - exerted high effort no matter what types of students they were assigned to teach). Since this is a top tier journal, they do a bunch of sensitivity analyses and model specifications to make sure their results are robust.

- **Data, sample, methods:** In 2005, 140 primary schools in western **Kenya** received funds to hire an extra first grade teacher. Of these schools, 121 had a single first-grade class, which they split into two sections, with one section taught by the new teacher. These are the schools they use for their study. **In 60 randomly selected** schools, students were assigned to sections based on initial achievement (tracking schools: those scoring in top

[16] https://www.aeaweb.org/articles?id=10.1257/aer.101.5.1739

half of the distribution go into one class, those scoring in the bottom half into another). In the remaining 61 schools, **students were randomly assigned to one of the two sections** (non-tracking schools). After students were assigned to sections, the teachers by type (contract vs civil-service) were randomly assigned to their sample includes 589 students in tracking schools, and 549 in non-tracking schools. They compare end of year achievement test scores for students in the tracking schools to those in the non-tracking schools; they also conduct regression analysis (using all students and a dummy variable indicating whether the student attended a tracking or non-tracking school) and control variables include the baseline achievement test scores, age and gender, teacher type (contract vs civil servant, the latter has union-like protections and tenure like job security), school size, student-teacher ratio. They compare effects based on whether the student was in the bottom or top half of the distribution (so these aren't just marginal students like in the RDD studies). They test for peer effects by estimating models with the average baseline test score of classmates and school fixed effects (school level dummy variables to control for school conditions not observable in the data like school culture, setting, etc.). In addition to regular regression, they also employ RDD using the students in tracking schools to assess peer effects and possible spillover effects comparing students just above the cut-off who got tracked into the class of high scoring students to those just below who got tracked into the class of lower scoring students. Because they have 60 different schools with tracking, they have 60 different discontinuities in their dataset and instead of comparing students further out from either side of the cut-off (the further you get the more dissimilar students are) they compare the two students just to either side of the cut-off from each of the 60 schools. Since this is a top tier journal, they do a bunch of sensitivity analyses and model specifications to make sure their results are robust. They do additional analysis (via random surprise classroom visits, etc.) to get a handle on how contract vs civil service teachers differed in their behaviors and teaching.

- **Results:** Their results suggest that all students benefit from tracking, both those in the upper and lower parts of the distribution. Students in tracking schools scored 0.14 standard deviations higher than students in non-tracking schools (the effect increased to 0.18 standard deviations when controls were included in the model). They found that

these gains persisted one year later, after the experimental tracking program was concluded. They also show that the results were similar for boys and girls (although girls got a bit of a larger boost from tracking on their math test results). They show differences in the effects of tracking were influenced by whether the student had a contract teacher or a civil service teacher, with all students (both initially low scoring and initially high scoring) benefiting if their teacher was a contract teacher and only initially high scoring students benefiting from tracking if their teacher was a civil service teacher (they provide additional analysis that suggests that this is because tracking induced civil service teachers to increase their effort when they were assigned to teach upper track students but not when they were assigned to teach lower track students while contract teachers exerted high effort no matter what types of students they were assigned to teach). Using the significant variation in student levels in non-tracked schools (classrooms are heterogeneous with students from across the full distribution, based on baseline testing), they also provide evidence of both direct and indirect peer effects. Their RDD results: despite the big gap in average peer achievement (with those assigned to lower level classes having on average lower scoring peers and vice versa), the marginal students' final test scores do not seem to be significantly affected by assignment to the bottom section; they speculate that this is the result of teachers assigned to the lower level class teaching to the top students within that group and so the marginal student does not suffer too much (i.e., there are no negative spillover effects).

***Redding and Grissom (2021) Do Students in Gifted Programs Perform Better? Linking Gifted Program Participation to Achievement and Nonachievement Outcomes***[17]

**Summary:** This study uses data from a nationally representative panel study and student-fixed effects models to evaluate the effect of receiving G&T services on achievement tests and teacher reported absences and student reported engagement. They find positive effects on both reading and math tests (smaller effect sizes than what Card and Guiliano found). But I'm a little wary of the robustness of their results because their results are so closely aligned with socioeconomic advantage (white and non-poor students are estimated to make bigger gains from G&T, at least in reading; Asian students do better in math). Student fixed effect techniques can control for selection and simultaneity bias only to the extent that unobservable factors related both to being assigned to G&T programming and subsequent achievement outcomes are time-invariant. This approach basically uses the individual student as their own comparison (comparing outcomes in the years they received G&T services to the outcomes in the years they did not). To the extent that student motivation, support, etc. remain unchanging, selection bias is controlled. The authors themselves caution that if unobserved factors like parents providing additional outside-of-school supports are more likely to occur during the years students are assigned to G&T then they would upwardly bias estimated program effects. They don't say this but differences in family resources could explain why they find larger benefits for socioeconomically advantaged students: the parents of socioeconomically disadvantaged students may tend to provide less outside of school support once their child was assigned to G&T (i.e., if resources are tight and they assume the extra programming would sufficiently cover their child's educational needs) while parents of socioeconomically advantaged students might tend to do the opposite (i.e., because they have the resources they may increase supports to ensure their child excels once identified as gifted). They also find differential program effects on students who are reported to move in and out of G&T programming and students who once they are reported to be in G&T are always reported to be receiving G&T services. They report that these two groups of students are similar on most of the observable measures available in the dataset. However, when they run their models separately on these two sub-samples, they find that it was the sample who *persisted in G&T once assigned* that

[17] https://journals.sagepub.com/doi/10.3102/01623737211008919

made the gains; for the students who switched out in at least one year, the estimated program effects were negative but not significant. This suggests there may be some unobservable factors that are at least biasing if not driving the estimated effects. For example, perhaps the students that moved in and out of G&T programming had lower levels of teacher and parent support and encouragement. Other concerns I have about this study relate to how G&T service status is measured: whether a child receives G&T programming is based on teacher answers to rather vague questions that changed a bit over time (there may be some measurement error). Even if their results are solid and not contaminated by bias, the fact that their estimated effects are quite small and accrue to socioeconomically advantaged students pose the policy question as to whether it's worth it to fund programs that produce such small effects and primarily for advantaged students whose parents can probably make up the difference if they aren't getting enough in regular classrooms and curricula.

- **Data, Sample and Method:** They use data from the nationally representative U.S.-based Early Childhood Longitudinal Study (2010-2011) Kindergarten Cohort (ECLS-K:2011) which tracks students from kindergarten (2011) through the 5$^{th}$ grade (n=18,170). The ECLS contains a lot of information about students, including information collected from their parents about their family life and a bunch of information collected from teachers about the students and their schools. They restrict the sample to public school students. Their analytic sample includes 37,980 student-year observations. Outcome variables include math and reading achievement test scores, teacher-reported student absences, student-reported engagement (do well in school, work hard in class, participate in class discussions, pay attention in class, listen carefully in class) and student mobility (whether student stayed or left the school within that year). The primary independent variable is whether the student received G&T services that year. This information is based on teacher report. Each year teachers are asked if the student received G&T instruction in math and/or reading. In year 3 an additional category was added that indicates participation in G&T programming in general. For students with IEPs, teachers are also asked if they receive G&T services as part of their plan. They use the information from all these questions to identify students participating in a G&T program. Their preferred statistical model is a **student fixed effect model**. This approach basically uses the individual student as their own comparison (comparing outcomes in the years they

received G&T services to the outcomes in the years they did not). To the extent that student motivation, support, etc. remain unchanging, selection bias is controlled (i.e., in these student fixed effects models the time invariant variables drop out of the model). They estimate the effect of participating in G&T program on the change in the outcome (e.g., test score) from the previous year. Controls include a bunch of student (race, gender, SES, EL status, disability status, whether English is the primary language at home, parent's report of child's health), teacher (race, years of experience, indicators of whether they have a Master's degree, certification status) and school (size, % FRPL, locale) level variables. They look for differential effects across student groups (race, SES, etc.) by including interactions between these demographic variables and whether the student participated in G&T programming that year.

- **Results:** They find relatively small but positive and significant effects on reading test scores (students scored 0.065 standard deviations higher when they were receiving G&T services: the typical student who ever receives G&T services scores at the 78$^{th}$ rank in years when they don't get G&T and in the 80$^{th}$ when they do) and only a very small and marginally significant (p=0.08) on math test scores (0.019 standard deviations: the average student scores in the 76$^{th}$ percentile when they didn't receive G&T and in the 77$^{th}$ when they do). They look at differential effects by student race (everyone is compared to white students) and FRPL status (everyone is compared to lowest SES) and find only a handful of statistically significant results (on interaction terms): G&T participation has smaller effects on reading test scores for Black students (0.177 SDUs less than for white students), larger effects on math tests for Asian students (0.09 SDU higher than for white students) and the most affluent students benefit more from G&T compared to the least affluent but only in reading (0.099 SDUs more). They found little evidence that G&T participation impacted the non-achievement outcomes (student absences, engagement in school, or whether a student leaves or stays in a school). They run a number of different model specifications (drop students with gaps in reported G&T participation, drop cases with missing values/imputed, use only those who switched in and out of G&T, student and school fixed effects) and their initial results generally stand except the estimated effects on math test scores sometimes lose statistical significance.

- **Discussion:** I'm a little wary because their results are closely aligned with socioeconomic advantages (white and non-poor students make bigger gains from G&T; Asians students do better in math); might these be the results of advantages and not G&T program effects? Student fixed effects can control for selection and simultaneity bias only to the extent that unobservable student level factors related both to being assigned to G&T programming and subsequent achievement outcomes are time-invariant. The authors themselves caution that if unobserved factors like parents providing additional outside-of-school supports are more likely to occur during the years students are assigned to G&T than they would upwardly bias estimated program effects. They don't say this but differences in family resources could explain why they find larger benefits for socioeconomically advantaged students: the parents of socioeconomically disadvantaged students may tend to provide less outside of school support once their child was assigned to G&T (i.e., if resources are tight and they assume the extra programming would sufficiently cover their child's educational needs) while parents of socioeconomically advantaged students might tend to do the opposite (i.e., because they have the resources they may increase supports to ensure their child excels once identified as gifted). Another possibility is that the G&T programing received by students differs by race and poverty status. For example, if minority students are more likely to be assigned to less intensive G&T programming or if they are more likely to attend schools with fewer resources devoted to G&T programming (perhaps they receive add-on instruction within the regular classroom instead of more advanced curriculum in a separate class). Also, they report (footnote 6) that 60% of their sample consisted of students who once they were identified as receiving G&T education, continue to be so and 40% were reported by their teachers as participating in one year and not in at least one subsequent year. They report that these students are similar on observable variables available in the dataset (except that those who stayed in once in were more likely to be in poor, rural schools, located in the South). In footnote 9 they report that they ran their models separately on these two sub-samples and found that it was the sample who persisted in G&T once assigned that made the gains; for the students who switched out in at least one year the estimated program effects were negative but not significant. They also report (in footnote 10) that when they include prior student achievement as a control the estimated program effects (in both

math and reading) are smaller. Taken together this sounds to me like there's unobservable things about the students could be at least biasing if not driving the estimated program effects. For example, perhaps the students that moved in and out of G&T programming had lower levels of teacher and parent support and encouragement. Another concern I have about this study relates to how G&T service status is measured: whether a child receives G&T programming is based on teacher answers to rather vague questions that changed a bit over time (there may be some measurement error). Even if there results are not biased, their estimated effects are quite small and accrue to socioeconomically advantaged students, which of course, raises the policy question: is it worth it to fund programs that produce such small effects and primarily for advantaged students whose parents can probably make up the difference if they aren't getting enough in regular classrooms and curricula?

***Enriching Students Pays Off: Evidence from an Individualized Gifted and Talented Program in Secondary Education (Booij, Haan, & Plug, 2016)***[18]

**Summary:** This study examines the effect of an enrichment type "pull-out" type G&T program (students get to trade regular classroom time to work on a self-designed independent project) offered at one prestigious secondary school in the Netherlands. It is included because of its statistical rigor and unlike the RDD studies above, which estimate program effects on marginal students (i.e., those just above the eligibility cut-off), this one uses RDD but explores the effect of G&T treatment on a group of exceptionally gifted students. In addition to focusing on one program at one school another advantage of this study is researchers had access to IQ test scores, standardized achievement test scores (exit exam from primary school) and a (presumably validated) test that measures motivation as well as supplement survey data on students' perceptions of the level of support they received from teachers and parents and their level of work effort, self-confidence, motivation, and academic self-esteem. This gives them the ability to control for selection and other bias beyond what most researchers have been able to do. Their estimates consistently show positive program effects on student achievement in the range of at least 0.30 standard deviations (larger for math and other subjects, smaller for reading). Based on analysis of their survey data they are able to rule out bias caused by spillover effects. They also show that the program did not spur students to work harder or increase their motivation or general self-esteem but that program participants did report an increase in academic self-esteem. This leads them to conclude that one of the mechanisms by which the G&T program works is that merely labeling students "gifted" raises their academic self-esteem and they rise to the occasion. The authors note that the students in this study are not your typical students and that the strong program effects may reflect that fact: they score high both on IQ tests and achievement test (students in this study were selected to participate based on their IQ test score cut-offs but only high achievers - based on their primary school exit achievement exam - are permitted into this school in the first place) and so they might be exceptionally exceptional in that they have both higher cognitive ability and stronger academic skills.

[18] https://www.econstor.eu/bitstream/10419/141516/1/dp9757.pdf

- **Data, sample, method**: This study examines the effect of an enrichment type "pull-out" type G&T program (students get to trade regular classroom time to work on a self-designed independent project) offered at **a prestigious secondary school in the Netherlands**. It is included because unlike the RDD studies above, which estimate program effects on marginal students (i.e., those just above the eligibility cut-off), this one uses RDD but explores the effect of G&T treatment on a group of **exceptionally gifted** students. (Dutch secondary education is tracked into three levels. The top level is further tracked into two types of schools and the school under study is among the most selective). Because all the students at this school are very high achievers, those who are then identified as "gifted" and eligible to participate in the G&T programs are exceptionally gifted students (even those just above the cut-off). Students at this school qualify for participation in the G&T program based on a standardized cognitive aptitude test. Students are tested during their first year and those who make the cut-off are eligible to participate for the next 6 years they are at the school. The cut-off is typically set as 1 standard deviation above the mean (but higher if G&T capacity is tight). The test is, however, only one factor the school considers for acceptance into the G&T program; while it is the main factor of determination, a committee makes the final determination (not all high scoring students get in; some students with lower are accepted). Because the cut-off is not strict, they **use a "fuzzy" regression discontinuity design (RDD).** Their sample includes 3,127 students, of which 785 students are assigned to the GT program. Their dataset includes student demographics (gender and age), primary education exit exam scores (CITO), GT program assignment status, scores on an intelligence test (IST) and a test that measures motivation (FES). The academic outcomes measures include grade retention, cumulative grade point averages (GPA) for math, languages, and other school subjects (all grades), and three indicators of choosing an advanced curriculum in the final two years (the number of exam subjects, the number of science subjects, and taking advanced math). They also had access to post-secondary data on university enrollment, field of study, whether the student switched majors, and the average starting salary that corresponds to field of study. The first stage in their two-stage RDD model estimates G&T program assignment predicted by a binary indicator for having an IST (intelligence test) score above the cut-off. The second stage estimates the effect of being assigned to the G&T program on academic outcomes controlling for gender, age (at the IST test), score of the exit exam at the end (CITO) or

primary school and FES (motivation) test score. They also conducted surveys to gauge students' level of work effort, self-confidence, motivation, and academic self-esteem as well as they level of support they perceived receiving from teachers and parents.

- **Results:** Their estimates consistently show positive program effects on student achievement in the range of at least 0.30 standard deviations (i.e., the marginal student who is barely admitted to the GT program has a cumulative GPA at least 0.30 standard deviations higher than the group of students just below the cut-off). Specifically, their estimated effects show indicate the G&T program raises cumulative grade point averages in math by 0.38 SD and language scores by 0.30 SD (and up to 0.44 SD in other subjects). They also find that male students work more on math and science related independent projects, and female students work more on language related projects and that male students experience the largest gains in math grades, and female students in language grades. They also find that students assigned to the G&T program are more likely to follow a more science intensive curriculum (particularly girls), and tend to report stronger beliefs about their academic abilities. They find evidence that program effects persist into university, where G&T participants chose more challenging fields of study with, on average, higher wage returns. They also use survey data to try and understand how the program works. Based these results they conclude that the program did not encourage students to work harder, boost their general self-confidence, or raise their motivation to learn. The program did, however, improve the G&T program participants' academic esteem.
- **Discussion:** An advantage of this study – in addition to its statistical rigor - is that they had access to IQ test scores, standardized achievement test scores (exit exam from primary school) and a (presumably validated) test that measures motivation as well as supplement survey data on students' perceptions of the level of support they received from teachers and parents and their level of work effort, self-confidence, motivation, and academic self-esteem. This gives them the ability to control for selection and other bias beyond what most researchers have been able to do. First, they show that students just above and below the cut-off do not differ in pre-treatment test scores, providing confidence that the treatment and comparison groups are not significantly different in academic ability and that any program effect can be interpreted to be causal and not driven by selection effects. Another possible

limitation is related to the fact that their primary outcome - cumulative GPAs for each subject during the secondary school grades - may suffer from some measurement error (teachers may be upwardly biased in grading when they know a student has been assigned to the G&T program). They test for this possibility using a nationwide standardized test students take in their final year: they rerun their RDD models using the results of this externally validated standardized exam instead of cumulative GPAs (by subject) and find effect estimates that are as large, if not larger (i.e., the positive gains in standardized tests were larger than the positive gains in GPAs). The authors note that the size of their estimated program effects are comparable to what Card and Giuliano (2014, described above) find for high achievers. They also note the fact that Card and Giuliano found no positive program effects for students assigned to G&T programming based on their IQ test results while they do. They posit that this is because the gifted students in their study are comparable in that they were also high achievers (students in this study were selected to participate based on their IQ test score cut-offs but only high achievers - based on their primary school exit achievement exam - are permitted into this school in the first place). They also test to make sure their positive program results weren't driven by spillover effects - the just below cut-off comparison group being disappointed they weren't selected for G&T (if the students just below the cut-off are negatively impacted and their grades suffer the RDD results are biased upward). They find no decline in grades for students just below the cut-off. They also asked all students not assigned to the G&T program how disappointed they were about not be selected and found that over 80% said they were not disappointed (and that only 5 students said they were seriously disappointed). They also survey students to see if a chance in support from parents or teachers might explain their effects. If parents and teachers know that some of the children are gifted and assigned to the G&T program, they may treat these children differently. Students were asked six questions on whether they were helped, encouraged, or pushed by their parents and teachers. They find positive effects (i.e., students assigned to the G&T program reported more of this support than the students below the cut-off) but the differences were not statistically significant. Finally, their survey also asked students whether they think of themselves as a good learner and their regression results showed positive, significant, and substantial program effects on self-assessed measures of academic esteem. This leads them to

conclude that one of the mechanisms by which the G&T program works is that merely labeling students "gifted" raises their academic self-esteem and they rise to the occasion.

**Dobbie and Fryer (2011) Exam High Schools and Academic Achievement: Evidence from New York City[19]**

**Summary:** I hesitate to even include this study… the only reason I do is it gets cited a few times by others. While they use the relatively robust RDD, the study has a couple of limitations (which may explain why it was never published), the most serious being that its comparison group is likely contaminated: they use administrative data from three of the 9 NYC "exam" schools (more rigorous curricula, higher achieving peers, more resources than typical public schools). Students who don't end up making the cut for one of the three exam schools under study may have in fact attended one of the other 6 exam schools or a private elite school with similarly rigorous standards or participated in the G&T program at a regular school (i.e., their below the cut-off comparison group is likely a terrible control and may explain why they find few positive effects). In addition, their data are more limited than other studies and so they are not able to do as much additional analysis to try and get a handle on the possibility of spillover effects or other endogenous biasing effects. Finally, they don't report on the extent of robustness checks they conducted.

- **Data, sample, method:** they use student level **data from NYC** to test whether enrollment in one of the city's three "exam" schools produces net benefits. Exam schools tend to have higher achieving peers, more rigorous instruction, and additional resources compared to regular public schools. Students compete to be enrolled into one of the three exam schools by taking the Specialized High Schools admissions test (SHSAT). The test is broken into a math and verbal section. The sample is restricted to NYC public school students in the 2002 through 2013 high school cohorts. **They employ a regression discontinuity design** to compare students who score just above the admissions cut-off to those who score just below.
- **Results:** Attending an exam school increases the rigor of high school courses taken and the probability that a student graduates with an advanced high school degree. Attending an exam

[19] https://www.nber.org/papers/w17286

school has little impact on Scholastic Aptitude Test scores, college enrollment, or college graduation.

- **Discussion:** While they use the relatively robust RDD, the study has a couple of limitations (which may explain why it was never published), the most serious being that its comparison group is likely contaminated: they use administrative data from three of the 9 NYC "exam" schools (more rigorous curricula, higher achieving peers, more resources than typical public schools). Students who don't end up making the cut for one of the three exam schools under study may have in fact attended one of the other 6 exam schools or a private elite school with similarly rigorous standards or participated in the G&T program at a regular school (i.e., their below the cut-off comparison group is likely a terrible control and may explain why they find few positive effects). In addition, their data are more limited than other studies and so they are not able to do as much additional analysis to try and get a handle on the possibility of spillover effects or other endogenous biasing effects. Finally, they don't report on the extent of robustness checks they conducted.

***Adelson, McCoach, amd Gavin (2012) Examining the Effects of Gifted Programming in Mathematics and Reading Using the ECLS-K[20]***

**Summary:** This study employs propensity score matching, which is generally less rigorous than RDD, but the researchers use a very rich dataset and a more thorough matching approach. Also, PSM evaluates program effects for all G&T participants, not just those just above the eligibility cut-off. Their sample comes from the nationally representative Early Childhood Longitudinal Study, Kindergarten Class of 1988-1989, which tracks students from kindergarten through to 5th grade and contains a ton of student, family, and school level information. To account for the fact that students are clustered in schools, they use hierarchical linear models (HLM) and to deal with non-random assignment they use propensity score matching to create comparison groups of not just students but schools as well. They maximized the use of this especially rich dataset and used up to 300 variables (that were both theoretically important and had a bivariate association with achievement or academic attitudes) to estimate the student level propensity scores and up to 82 variables (based on theory and whether they had a bivariate association with whether the school had a G&T program or the school's mean achievement scores) to estimate the school level propensity scores. They estimate G&T program effects at both the school and student levels. They find no effects at the school level – average reading and math scores were about the same regardless of whether a school offered G&T programming. They find no effects at the student level – average math and reading scores are the same for gifted students who attended a school with a gifted program and gifted students who attended a school w/o a gifted program. They also found no effects on students' reported attitudes about reading and math. There is of course still the possibility of selection bias because propensity score matching uses only what is observable (i.e., available in the data set). But they went further than other studies to limit selection effects. They matched students on 300 variables, including many often associated with endogeniety and selection effects that often go unmeasured in other studies. They also limit selection bias related to unobserved factors by comparing gifted students at schools with G&T programs to gifted students at schools that did not have G&T programs (since it wasn't even an option the comparison group contains both students who would have participated if they could have and

[20] https://journals.sagepub.com/doi/abs/10.1177/0016986211431487

those who would have not). Finally, selection is generally assumed to inflate estimated program effects and since they find none, it doesn't seem like selection is a big problem. The study is limited in that it does not test for heterogeneous program effects by student demographics (race, FRPL status, etc.)

- **Data, Sample, Method:** They use data from the nationally representative Early Childhood Longitudinal Study, Kindergarten Class of 1988-1989, which tracks students from kindergarten through to 5th grade and contains a rich set of student level, family level and school level variables. Teachers report whether students participated in gifted and talented programming separately for math and reading. They selected and analyzed reading and math samples separately. They selected only those students who remained in the same school through to 5th grade and who were consistently in or not in a gifted program (all or nothing). The reading sample included 5,630 students in 850 different schools and the math sample included 2,740 students in 720 schools. They used multiple imputation to deal with missing information. They checked to make sure their partial sample looked like the full nationally representative sample according to student and family demographics as well as school level demographics (% FRPL, % minority). They combined information from administrators (whether a school offers G&T programming or not; the number of PT and FT G&T teachers) and student level information collected from teachers as to whether they received G&T services to identify which schools provided a G&T program and which did not. They use propensity score matching to produce matched pairs of schools (those that provide G&T programs and those that do not) and students (those that participate in G&T and those that do not). They used up to 300 variables (that were both theoretically important and had a bivariate association with achievement or academic attitudes) to estimate the student level propensity scores and up to 82 variables (based on theory and whether they had a bivariate association with whether the school had a G&T program or the school's mean achievement scores) to estimate the school level propensity scores. Their outcome variables include achievement test scores in math and reading and academic attitude (based on survey questions asked of students in 3rd and 5th grade as to their perceptions of their grades, the difficulty of their schoolwork, and their interest and enjoyment in the subject). They test for G&T program effects at both the school and student level. At the student level instead of just

comparing students who participated in G&T programming to those who did not they compare gifted students at schools with G&T programs to gifted students at schools that did not have G&T programs.

- **Results:** They find no effects at the school level – average reading and math scores were about the same regardless of whether a school offered G&T programming. They find no effects at the student level – average math and reading scores are the same for gifted students who attend a school with a gifted program and gifted students who attend a school w/o a gifted program. They found no effects on students' reported attitudes about reading and math either.
- **Discussion:** This appears to be a more rigorously designed PSM study than most (although I'm generally not a fan of this method so I don't see a lot of it). These researchers had access to a very rich data set with information about students, their families, and the schools they attend. They used up to 300 different variables to match students and up to 82 variables to match schools. Among the variables they used to match participating and non-participating students are a number of factors that are associated with endogeniety and selection bias effects but often go unmeasured in other studies (e.g., self-control, attention, cooperation, the level of their parents' participation in their education). They also do a better job of limiting selection bias related to unobserved motivation by comparing gifted students at schools with G&T programs to gifted students at schools that did not have G&T programs (since it wasn't even an option the comparison group contains both students who would have participated if they could have and those who would have not). That said, if there are unobserved factors that relate to school choice – if, say, rich parents remove their child from schools w/o G&T programs and move them to ones with G&T programs (there's research that shows this happens[21]) and if this means that more supported and encouraged students are attending the G&T schools leaving behind the less supported or encouraged students in the comparison schools then their results could still be biased (i.e., if rich parents who also have the means to provide outside of school supports and enrichment activities are more likely to remove their children from schools without G&T programs then those schools will

[21] https://www.jstor.org/stable/23646325?seq=1#metadata_info_tab_contents

contain more gifted but poor and presumably less supported students). This would inflate the estimated effects of G&T programs. Since they don't find any statistically significant effects if inflation is happening it's small. Finally, they confirm their results by looking for G&T program effects at both the school and student levels. The study is limited in that it does not test for heterogeneous program effects by student demographics (race, FRPL status, etc.)

***Park, Lubinski and Benbow (2012) When Less is More: Effects of grade skipping on Adult STEM productivity among Mathematically Precocious Adolescents.***[22]

**Summary:** This next study is included because it looks at the effects of acceleration (grade skipping is a cost-effective G&T type program…) and focuses on the exceptionally gifted students. Their sample is drawn from three cohorts of the Study of Mathematically Precocious Youth, a panel survey that has been tracking students for 40 years (3 cohorts include 1972-74, 1976-79, and 1980-83) and includes 3,467 students all in the top 1% or higher (based on results of SAT math tests) by age 13. They use a combination of exact (gender, number of grades previously skipped before identified as precocious by the study) and propensity matching (SAT scores, measures of subject interest, class standing, parent occupation and educational attainment, number of siblings, birth order) to generate a treatment and comparison group. They find that grade skippers were more likely to pursue advanced degrees, advanced degrees in STEM (unless they were female, in which case they were more likely to get a medical or law degree), and author peer reviewed articles, earned their degrees and authored peer reviewed articles at a younger age, have more citations and highly cited publications by age 50. A major limitation with any propensity score matching study is that matching is only on observed variables which leaves room for bias related to unobserved factors. Also, they have access to very few variables for matching (compared to a study described below) and so bias due to unobserved factors is likely even more of an issue here, meaning the differences they find could actually be due to these unobserved factors and not grade skipping.

- **Data, Sample, Method:** The type of G&T programming investigated here is acceleration (grade skipping). Their sample is drawn from three cohorts of the Study of Mathematically Precocious Youth, a panel survey that has been tracking students for 40 years (3 cohorts include 1972-74, 1976-79, and 1980-83) and includes 3,467 students all in the top 1% or higher (based on results of SAT math tests) by age 13. They use a combination of exact (gender, number of grades previously skipped before identified as precocious by the study) and propensity matching (SAT scores, measures of subject

[22] https://my.vanderbilt.edu/smpy/files/2013/02/Park-Lubinski-Benbow-2013.pdf

interest, class standing, parent occupation and educational attainment, number of siblings, birth order) to generate a treatment and comparison group. After matching grade skippers and non-skippers exactly on sex and number of previous grades skipped, matches were further improved by matching on these other covariates (SAT scores, measures of subject interest, class standing, parent occupation and educational attainment, number of siblings, birth order) by matching to the nearest in propensity score. Their analytic sample includes 363 grade skippers matched to 657 non-skippers. They compared the two groups straight up and then using logistic regression to control for the slight differences in covariates between the two groups.

- **Results:** Grade skippers were more likely to pursue advanced degrees, advanced degrees in STEM, and author peer reviewed articles, earned their degrees and authored peer reviewed articles at a younger age, have more citations and highly cited publications by age 50. They do find differences by gender: female grade skippers were actually less likely than non-skippers to get STEM PhDs, but they were more likely to than their matched controls to get PhDs in general; while female grade skippers were less likely to pursue STEM PhDs compared male grade skippers, female skippers tended to pursue medical degrees and law degrees (these outcomes mean they also had fewer STEM pubs and citations, of course).
- **Discussion:** A major limitation of any propensity score matching study is that matching is only on observed variables which leaves room for bias related to unobserved factors. Also, they have access to very few variables for matching (compared to a study described above) and so bias due to unobserved factors is likely an issue here, meaning the differences they find could be due to these unobserved factors and not grade skipping.

***Bhatt, R. R. (2009). The impacts of gifted and talented education. Andrew Young School of Policy Studies Research Paper Series.23***

**Summary:** This study was never published in a peer review journal, likely because of its unusual findings and weak instrumental variables….but, it gets cited fairly often so I include it here. Bhatt uses data from National Education Longitudinal Survey, a nationally representative, longitudinal study of 8th graders begun in 1988 that tracks students all the way through high school and includes a lot of information about students, their families and the schools they attend. Her analytic sample includes 5,265 8th graders attending 850 schools across the U.S. offering G&T programs. She examines the effect of G&T participation during the 8th grade on a number of outcome variables including scores on math and reading standardized tests, whether they took advanced placement classes in high school, enjoyed school, felt challenged, and took the college entrance exams. She employs instrumental variables regression techniques to try and control for selection bias. While Bhatt uses a rich dataset and was able to control for lots of other factors related to academic performance (student level: race, gender, average GPA from grades 6 and 7; family level: whether they get the newspaper, have encyclopedias at home, whether at least one parent works, parents' highest level of education, family SES; school-level: 8th grade attendance rate, student-teacher ratio, teacher salary, %FRPL, %remedial, % racial minority, urban-rural location, type of G&T admissions criteria used), her instrumental variables are weak and probably do not actually control for selection bias. This may explain why her estimated effects are almost three times larger (0.89 SD) than what the more rigorous studies above report (typically in the range of 0.30 SD). Additional evidence that her IVs might be invalid are the pattern of her results: it is generally assumed that the effect of unmeasured heterogeneity between participating and non-participating students will result in OLS estimates being biased upward (because the program participation variable is capturing both program effects and unmeasured stuff like greater motivation, family support, etc.) and so models that control for selection will thus produce smaller estimates of program effects. Bhatt finds the opposite, at least with math achievement scores. She suggests that maybe it is because lower ability students are often allowed to participate in G&T programs and her IV's capture this (parents push, teachers

[23] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1494334

adjust scores of below-cut-off students, etc.). Instead of positive selection bias (where estimates make the program look more effective than it is because OLS does not control for the fact that higher motivated students are assigned to G&T) Bhatt's posits that her OLS estimates suffered from negative selection bias (i.e., lower ability student's pulled subsequent test scores down). However, this occurs only with the math outcomes; the estimated gains in reading test scores are all smaller when using the IV model compared to those produced by the OLS model (indicating there was positive selection bias with the OLS model). Her argument would be stronger if she found evidence of negative selection bias in both cases. Another limitation of this study is that it does not examine whether the effects of G&T participation differ by student type (race/ethnicity, gender, FRPL status).

- **Data, Sample and Method:** Bhatt uses data from National Education Longitudinal Survey, a nationally representative, longitudinal study of 8th graders begun in 1988 that includes assessment scores in reading, social studies, mathematics, and science from the 8th, 10th, and 12th grades and high school and post-secondary transcripts (types of courses taken, grades) with additional data collected using surveys of students, parents and teachers (school environment, neighborhood environment, home life, support, attitudes, aspirations, extracurricular activities, etc.). Her analytic sample includes 5,265 $8^{th}$ graders attending 850 schools across the U.S. offering G&T programs. She examines the effect of G&T participation during the $8^{th}$ grade on a number of outcome variables including scores on math and reading standardized tests for $8^{th}$, $10^{th}$ and $12^{th}$ grades, whether they took advanced placement classes in $10^{th}$ grade, enjoyed school ($8^{th}$ grade), felt challenged ($10^{th}$ grade), and took the college entrance exams (PSAT, SAT) in the $10^{th}$ or $12^{th}$ grade. She employs **instrumental variables regression** techniques to try and control for selection bias. She employs two different 3-way interaction terms as her instruments: (1) the school uses past GPA to assign students to G&T program, (2) child's average GPA in 6th and $7^{th}$ grades, and (3) percentage of remedial students in the child's school. The other IV she uses is the following 3-way interaction: (1) whether the school uses race as a criterion for admission into G&T (yes/no), (2) whether the student is a racial minority (yes/no) and (3) the % of students in the student's school who are racial minorities. The first stage uses these one of these 3-way interactions plus both the 2-way

and the individual variable so that selection is off the 3-way variables (she's hoping that means they will be predictive of whether a student gets assigned to G&T program but not directly correlated with the academic outcomes she uses to evaluate the program effects) to predict whether the student is assigned to a G&T program. The second stage estimates the effect of being assigned to a G&T program. She shows both the base model results (OLS, which doesn't attempt to control selection bias) and IV 2SLS results. All models control for a lot of student, family, and school variables.

- **Results:** She finds that students who reported being in a G&T program in 8th grade have significantly higher standardized math scores (0.89 standard deviations) at the end of 8th grade (but no difference in ELA scores) compared to students who did not participate in G&T programs (she finds no significant differences in math scores in later grades); she also finds that G&T participating students are significantly more likely to take AP classes in 12th grade. She finds no other positive effects of G&T program participation (i.e., no significant differences between G&T participants and non-participants in their reports of enjoying school or feeling challenged, whether they agreed that it was important to their friends to get good grades, whether they knew schoolmates who dropped out, whether they took the SATs).
- **Discussion:** While Bhatt uses a rich dataset and was able to control for lots of other factors related to academic performance (student level: race, gender, average GPA from grades 6 and 7; family level: whether they get the newspaper, have encyclopedias at home, whether at least one parent works, parents' highest level of education, family SES; school-level: 8th grade attendance rate, student-teacher ratio, teacher salary, %FRPL, %remedial, % racial minority, urban-rural location, type of G&T admissions criteria used), her instrumental variables are weak and probably do not actually control for selection bias. This may explain why her estimated effects are larger than what the more rigorous studies above report (typically in the range of 0.30 SD). Recall that a strong IV is one that is highly correlated with program participation but not with subsequent academic achievement. Her 3-way interactions might be invalid if, for example, motivated parents of less-gifted students seek out schools in which their child would have a better chance of getting in. Or if highly motivated but-less gifted students knowing that

the school uses 6th and 7th grade GPA and that the competition in their school is high work harder to increase their GPA. However, even if her IVs are valid, they likely suffer from weak instrument bias. A test for the strength of the IVs is the F test in the first stage predicting participation in the G&T program: a rule of thumb is that the F statistic should be 10 or higher. Bhatt's F tests ranged from 2.3- 4.4. A weak instrument means the second stage estimate of the effect of participating in the G&T program will be imprecise and probably biased. Additional evidence that her IVs might be invalid as well as weak are the pattern of her results: it is generally assumed that the effect of unmeasured heterogeneity between participating and non-participating students will result in OLS estimates being biased upward (because the program participation variable is capturing both program effects and unmeasured stuff like greater motivation, support, etc.) and that IV models will produce smaller estimates of program effects (because they are separate out program effects from higher levels of motivation and support among program participants). Bhatt finds the opposite, at least with math achievement scores: the estimated program effect sizes are larger with the IV model compared to estimates produced using OLS. She suggests that maybe it is because lower ability students are often allowed to participate in G&T programs and her IV's capture this. Instead of positive selection bias (where estimates make the program look more effective than it is because OLS does not control for the fact that higher motivated students are assigned to G&T) Bhatt's posits that her OLS estimates suffered from negative selection bias (i.e., lower ability student's pulled subsequent test scores down). However, this occurs only with the math outcomes; the estimated gains in reading test scores are all smaller when using the IV model compared to those produced by the OLS model (indicating there was positive selection bias with the OLS model). Her argument would be stronger if she found evidence of negative selection bias in both cases. Another limitation: Bhatt does not examine whether the effects of G&T participation differ by student type (race/ethnicity, gender, FRPL status).

***Does Sorting Students Improve Scores? An Analysis of Class (Collins and Gan, 2013)***[24]

**Summary:** This study examines whether sorting students by ability (or G&T status) leads to academic gains. They use student-level **data from the Dallas Texas school district** including standardized test scores, a student's identification as gifted and/or special needs or EL and their demographics (race, gender and FRPL status). They had a classroom ID so they were able to link students to a particular class and use student level achievement test scores to develop a an index indicating how homogeneous (sorted) or heterogeneous (not sorted) a student's class was relative to the other same grade classes in the school (they do the same to measure sorting by G&T status). They use instrumental variables to try and isolate a causal effect of sorting. The estimated effects of sorting from the 2SLS IV model are positive and significant for both math and reading scores, and generally larger in magnitude than the OLS estimates, suggesting there was a downward bias caused by selection in the base model (i.e., OLS with a variable indicating whether the student attended a sorting school or not). These results hold across various specifications—for both level scores and score gains, and when outliers are excluded. They also ranked students according to their previous year testing score and estimated separate models for high and low scoring students. While the results suggest slightly larger effects for high scoring students, they still find large, positive, and significant results for the low scoring group. While they find positive effects for students in classes that are sorted (homogeneous) by ability, they do not find any significant effects for G&T students in schools where G&T students are sorted. Like Bhatt (described above) there IV results suggest that selection was causing downward bias on estimated program effects. Unlike Bhatt they find this with both math and reading outcomes. Their IV seems stronger and more plausibly valid than Bhatt's, although of course we can't know for sure. As discussed above there's no way to statistically assess whether the IV is exogenous and valid (i.e., whether it is both correlated to assignment to a tracked classroom but not correlated with subsequent academic achievement). This study is also not published in a peer-reviewed journal which may be because the robustness of results produced by IV models are tough to evaluate. Other limitations of this study are related to its investigation as to whether being in a homogenous G&T classroom improves achievement among G&T identified students

[24] https://www.nber.org/system/files/working_papers/w18848/w18848.pdf

(compared to those in unsorted classrooms). The authors point out that they had no information on what types of G&T programs the schools provide. Their finding of no effects by G&T sorting could be because G&T students in unsorted classes are pulled out of class for certain subjects or projects (if this positively impacts their 4th grade test scores it will make any gains made by G&T students in sorted classes relatively smaller). They also did not explore whether there are differential effects by student gender, race or FRPL status.

- **Data, Sample and Method:** This study examines whether sorting students by ability (or G&T status) leads to academic gains. They use student-level **data from the Dallas Texas school district** including standardized test scores, a student's identification as gifted and/or special needs or EL and their demographics (race, gender and FRPL status). They had a classroom ID so they were able to link students to a particular class and use student level achievement test scores to develop a an index indicating how homogeneous (sorted) or heterogeneous (not sorted) a student's class was relative to the other same grade classes in the school. They construct a separate sort index using math and reading scores. They do the same with the gifted, EL and Special education status variables). Their sample includes all third grade students in the 2003-2004 school year who become fourth graders in 2004-2005, a total of 9,325 children from 135 different schools. They examine the impact of being in a sorted class on 4th grade achievement test scores (math and reading) and on changes between scores between 3rd and 4th grade. To try and isolate the effect of sorting by test scores from sorting on unobserved student characteristics they use **instrumental variables.** Their IV is a measure of whether the school sorts students in the 5th grade, the assumption being that if a school sorts students into (more or less) homogeneous classes in the 5th grade they probably also sort 4th graders (but that whether the school sorts in 5th grade shouldn't affect a student's 4th grade outcomes). The first stage model uses a school-level 5th grade sorting indicator as a predictor of whether a student is in a sorted or not class in 4th grade; the estimate from this first stage is entered into the second stage equation as a predictor in estimating the effect of being in a sorted classroom on math and reading scores on achievement tests taken at the end of the 4th grade. Control variables include: student's 4th grade math(reading) score, gender, race, EL, G&T and special education status; teacher's experience and the school's average

teacher salary, the class size, the school's average math and reading scores, enrollment and enrollment-squared.

- **Results:** The estimated effects of sorting from the 2SLS IV model are positive and significant for both math and reading scores, and generally larger in magnitude than the OLS estimates, suggesting there was a downward bias caused by selection in the base model (OLS with a variable indicating whether the student attended a sorting school or not). These results hold across various specifications—for both level scores and score gains, and when outliers are excluded. They also ranked students according to their previous year testing score and estimated separate models for high and low scoring students. While the results suggest slightly larger effects for high scoring students, they still find large, positive, and significant results for the low scoring group. While they find positive effects for students in classes that are sorted (homogeneous) by ability, they do not find any significant effects for G&T students in schools where G&T students are sorted.

- **Discussion:** Like Bhatt, who also uses 2SLS IV methods, their results indicate that selection effects were producing a negative (downward) bias (their base model – OLS with an indicator variable measuring the degree of sorting in the child's 4$^{th}$ grade classroom – produced smaller positive estimated effects of sorting). Unlike Bhatt they find this pattern with both math and reading outcomes. As is the case with all IV models, it is hard to know for sure if the instrument is controlling selection effects and allowing researchers to isolate causal program effects. They provide evidence that their IV is at least moderately strong (they report correlations of 0.37 to 0.57 between 5$^{th}$ and 4$^{th}$ grade sorting indices, with sorting by G&T status the lowest and sorting by math scores the highest). However, as discussed above there's no way to statistically assess fully assess the quality of an IV (whether it is both correlated to assignment to a tracked classroom but not correlated with subsequent academic achievement). It is plausible like the authors say that their IV is not biased, but as with all IVs, this cannot be directly measured. It seems logical that the way 5$^{th}$ grade students are assigned to classrooms would have no impact on the academic performance of 4$^{th}$ graders (unless, say, schools that sort in 5$^{th}$ grade also provide extra help to 4$^{th}$ graders to prepare for their achievement tests; or more

motivated or supported students at the school know that their 4th grade test results will determine whether they get into a homogenously high scoring classroom and act upon that knowledge).

- **Other limitations:** The authors point out that they had no information on what types of G&T programs the schools provide. Their finding of no effects by G&T sorting could be because G&T students in unsorted classes are pulled out of class for certain subjects or projects (if this positively impacts their 4th grade test scores it will make any gains made by G&T students in sorted classes relatively smaller). They also did not explore whether there are differential effects by student gender, race or FRPL status.

## Summary and Conclusions

Maine's Gifted and Talented (G&T) programs, and the subsequent funding that follows, appear to be unevenly distributed. Higher poverty SAUs are less likely to have approved G&T programs and to report students identified as G&T. This raises concerns about equitable opportunities to participate, if in fact G&T programs are an evidence-based intervention that should therefore be available to any student who would benefit. This leads to the more basic question of whether the costs of such programs are worth the investment, or whether Maine should consider following the precedent set by other New England states and reallocate these funds to other purposes.

The empirical evidence on the impacts of participation in G&T programs is decidedly mixed. In the absence of universal testing and random assignment, rigorous research on G&T program participation is hard to do–because of selection bias, confounding factors, simultaneity—and because there is so much variation on the type, quality and intensity of G&T programming. If there are net gains from G&T programs, they are more likely in the following circumstances:

- Socioeconomically disadvantaged students may benefit the most from G&T programs (Card and Guiliano); these benefits might come as much from indirect effects (more supportive classroom environment, classmate peers with stronger academic performance, higher teacher expectations) as they do from a more advanced curriculum. Yet in Maine, socioeconomically disadvantaged students are far less likely to participate. Universal screening may help to identify more students from disadvantaged backgrounds.
- More exceptionally gifted students—those outliers with the highest achievement—might benefit more than others identified (generally in the top 5%).

In conclusion, it is unclear whether the benefits of pull-out G&T programs are large or discernible enough in Maine to warrant the current level of investment. We recommend continued exploration of a personnel ratio in the EPS model to provide MTSS learning specialists who have the capacity to support any student with academic learning needs—at both ends of the spectrum—that are beyond the range of what can feasibly supported by the general classroom teacher through differentiated instruction.