Long-Term Impacts of Mentoring for Disadvantaged Youth*

Alex Bell, University of California - Los Angeles

Neviana Petkova, Office of Tax Analysis, US Treasury

February 21, 2023

Preliminary Conference Discussion Draft

Abstract

How do adult mentors shape kids' life trajectories? To evaluate this question, we leverage the microdata from a 1991 RCT that randomized disadvantaged children's eligibility for a popular mentoring program. Our re-analysis of the original short-run survey data suggests that kids' behaviors improved during the time they were with mentors. A linkage to later-life administrative records yields imprecise estimates of treatment effects on earnings, but significant evidence that the treatment group is on a better trajectory in several ways. For instance, members of the treatment group today are significantly more likely to have attended college, though standard estimates of the economic returns to education would be too small to detect in this experimental sample. Although our estimates suggest that mentoring programs will not fully equalize economic opportunities for disadvantaged youth, the program's relatively low costs and substantial benefits may place it among the most cost-effective interventions of its type to be evaluated.

*This research was conducted while one of the authors was an employee at the U.S. Department of the Treasury. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors and do not necessarily reflect the views or the official positions of the U.S. Department of the Treasury. Any taxpayer data used in this research was kept in a secured Treasury or IRS data repository, and all results have been reviewed to ensure that no confidential information is disclosed.

1 Introduction

A growing body of literature has uncovered how disparities in childhood environments – from schools to neighborhoods – contribute to earnings and achievement gaps in adulthood. One avenue by which childhood environments may shape kids' trajectories is via exposure to mentors. Because economically successful adults are in low supply in low-income and minority communities, not all youth have equal access to mentors. Yet evidence is mounting that the types of adults that kids are exposed to shape their career trajectories.[1] Of particular policy interest is whether artificially created relationships can achieve the same hypothesized effects of the natural mentors that privileged youth can more easily find in their homes and communities.

To answer these questions, we leverage micro-data from a randomized control trial conducted nearly 30 years ago, which we link with US administrative tax data to observe participants' long-run outcomes. The experiment randomized eligibility for mentoring during 18 months among a pool of primarily low-income and minority youth. The treatment was a purely social one, with no financial transfers or tutoring. Relationships were begun between ages 10 to 14 and typically lasted only a few years, amounting to some 220 hours spent with a volunteer mentor total. By the standard of information-based experiments designed to improve kids' outcomes such as Bettinger et al. (2012), this comes off as a fairly intensive treatment. In contrast, relative to the conceptual experiment of changing one's parents implied by adoption studies like Sacerdote (2007), the time the mentors spend with the kids is about 3% of what an average parent would spend with his or her child according to the American Time Use Survey.

The paper proceeds as follows. Section 2 provides further background on the mentoring and experiment and Section 3 outlines the data available for analysis. Section 4 briefly re-analyzes behavioral outcomes data from the original study with an eye toward modern-day techniques for multiple hypothesis testing, and 5 extends the analysis to long-run outcomes from administrative data. Section 6 concludes with a discussion of costs an benefits.

2 Background

The experiment we study introduced a randomized waitlist into the typical workflow of eight affiliates of Big Brothers Big Sisters of America between 1991 and 1993.

[1] A growing theoretical and applied literature has studied the effects of early childhood experiences, particularly social ones, on kids' longer-term outcomes (Heckman, 2006; Cunha, 2013; Heckman and Mosso, 2014). More specific to mentoring effects, Bell et al. (2019) have shown shown that kids who are exposed to more inventors in childhood are more likely to become inventors themselves, and this effect is particularly strong along same-sex lines. Other recent work by Chetty et al. (2020) has suggested that mentors play a role in economic mobility based on the pattern that neighborhoods with the highest rates of economic mobility for young black boys are those with more black fathers present as role models.

2.1 Big Brothers Big Sisters

Founded in 1904, Big Brothers Big Sisters is a national non-profit mentoring program organized in local agencies that match youth with volunteer mentors and provide professional support during the relationship. In the United States, the mentoring program is currently implemented by 279 affiliates operating in all 50 states.

The mission of Big Brothers Big Sisters of America is to "[p]rovide children facing adversity with strong and enduring, professionally supported one-to-one relationships that change their lives for the better, forever" (BBBSA, 2016). BBBS has served more than 2 million youth in the last decade (Klinger, 2018). The program dates back, in some form, to 1902. The BBBS model pairs volunteer mentors with youth who face some form of adversity. This paper focuses on the BBBS community-based mentoring program run in the United States under the supervision of BBBSA. Several affiliates now supplement the original community-based model with school-based or even SMS-based mentoring programs.

Although some affiliates enroll youth at ages as young as 6 or as old as 18, almost all youth in this study were aged between 10 and 14 at time of application (a few as old as 16). The study youth are 60% male and more than half belong to a racial minority group, with virtually all youth living either with one parent or with other guardians. The program targets youth with some form of adversity, but ideally not yet with emotional problems so severe that make working with a mentor difficult; an old service delivery manual for a Boston affiliate states "because of the large number of boys applying to our program, we are unable to accept those who are doing well." Upon intake, the youth and guardian are interviewed by caseworkers. Staff elicit information on family histories including trauma and developmental problems, preferences for mentor characteristics, and overall suitability of the youth for the program.

Volunteer mentors, who are often well-educated young professionals, are screened by staff interviews and through criminal background checks as well as character references from friends, family, and employers. In addition to screening for red flags such as a history of violent crime, a goal of the screening process is to determine whether the volunteer seems reliable. Personal information about the volunteer's family history and relationships is also collected for matching purposes.

Staff make matches subjectively based on a variety of quantitative and qualitative data they record on activities of interest, preferences, shared life experiences, and geography. During the study, matches are only same-sex. The matching is one-to-one, and youth and volunteers are typically not re-matched after completion. When all parties agree to a match, the caseworker facilitates an introduction either in the agency's office or the youth's home. Both sides are asked to commit to stay matched for at least one year, with an expectation that if the relationship is beneficial it will continue for longer. The time commitment

varies slightly across affiliates and over time, but for the matches in the study, the requirement was typically one four-hour outing per week. Most affiliates allow for some lessening of this commitment after the first year. The match is declared closed when the necessary frequency of meetings can no longer be met or both sides otherwise agree to call the match closed.

The types of activities that matches engage in are usually social or cultural, and often outdoors. Volunteer mentors are instructed not to serve as tutors to the youth and also not to buy the youth excessively valuable gifts. Common activities reported by matches in the study include eating at a restaurant, going to the movies or mall, playing games, and athletic activities like biking. Many affiliates also offer annual picnics or holiday parties for matches. Affiliates typically provide matches with information about local cultural opportunities and are sometimes able to provide matches with discounted admissions to events.

While the match is open, Big Brothers Big Sisters commits to provide match support to volunteers, youth, and their parents. Most often, this entails regular phone conversations with a caseworker referred to as a match support specialist, who usually has some training or experience in social services or related fields. The match support specialist evaluates the relationship and gives guidance, and can liaise with all parties.

2.2 Experimental Design

2.2.1 Background

From 1970 to 1990, the percent of kids growing up without a father rose from 13% to 27%. By 1990, nearly 60% of black boys did not live with a father (Bureau, 2017). This rise in single-headed households was followed by a proliferation of mentoring programs for at-risk youth. Internal study documents reflect that these demographic trends and proliferation of competing mentoring programs was the impetus for BBBSA to commission a randomized control trial of its effect in 1991. The trial was implemented by Public/Private Ventures (P/PV), a now-defunct evaluation firm, with additional contract work provided by Mathematica Policy Research.[2]

The academic report on the experiment, published by Grossman and Tierney (1998), highlighted significant effects on drug/alcohol use, violence, school attendance, and family relationships. Although no formal pre-analysis plan was submitted, prior to randomization researchers had outlined five broad sets of out-

[2]This trial was one of two large-scale RCT's commissioned by Big Brothers Big Sisters of America. This one evaluated the Community-Based Mentoring (CBM) program, whereas Herrera et al. (2011) evaluated School-Based Mentoring (SBM). CBM pairs a youth with an older mentor from the community to do activities outside of school, whereas SBM pairs youth with mentors who meet with them for shorter periods of time during or after school, such as at lunch. This paper investigates the CBM program because it has been in effect for much longer and has more established support in the research community. Two other smaller-scale studies involving BBBS include one in Canada by De Wit et al. (2007) that found few significant results among 71 families and another evaluation by the U.S. Dept of Justice (2011) of a program serving children of incarcerated parents. Other work in the mentoring field includes the meta-analysis of 73 studies by DuBois et al. (2011) and qualitative work by Spencer (2007) on dynamics of mentoring relationships. In economics, Rodriguez-Planas (2012) has evaluated long-term impacts of a randomized trial of a different national program with a mentorship component, finding a strong positive effect of the program on earlier high school completion and college attendance but negligible long-run effects on employment.

comes on which they hoped to see effects. In brief, they were: social and cultural enrichment, self-concept, relationships, school, and antisocial activities. More information on these hypotheses is in Appendix A.

2.2.2 Sample Construction

Researchers selected eight affiliates to participate with the goals of geographic diversity and large caseloads. The sample cities were as follows: San Antonio, TX; Columbus, OH; Houston, TX; Minneapolis, MN; Philadelphia, PA; Rochester, NY; Wichita, KS; and Phoenix, AZ. Most youth applying to BBBS in these locations in October of 1991 through February 1993 were included in the research sample.[3] Families that agreed to participate in the study faced an equal chance of the agency attempting to match them immediately or being put on an eighteen-month waitlist. Those who declined to participate faced an automatic twelve-month waiting list. As part of this modified agency intake process, 1,138 families gave consent and only 32 declined to participate in the randomization scheme. Within a few days of the intake interview, researchers notified the applicants of their experimental group. Randomization was implemented so as to be balanced in agency by gender by minority cells, and siblings were randomized independently. Researchers began follow-up telephone interviews with youth and their parents in April 1993, ultimately reaching 959 families. The final analysis sample consisted of 487 treated youth and 472 control youth.

2.2.3 Treatment & Control Dilution

This experiment randomized eligibility. Not all members of the treatment group were matched during the period of the 18-month study; typically this was due to inability to find a suitable volunteer to match with. By the time of the follow-up survey, 78% of the treatment group had been matched with a mentor through BBBS. Most of the unmatched youth were boys, due to the low supply of male volunteers relative to over-demand by boys for mentors.

Whereas many youth from the treatment group did not get treatment, it was relatively rare for controls to receive mentors. Although members of the control group were ineligible during the study to be matched with BBBS, they technically could have applied for a mentor through a competing organization. But in practice, only 5% of the control group reported participating in any other type of mentoring program by the time the study concluded.

When we proceed to analyze the long-run effects of the study, it is important to note that all of the control group would have been eligible for a BBBS mentor upon conclusion of the study. Throughout the whole paper, we only ever report intent-to-treat estimates, which should be a conservative lower-bound on the effect of being matched with a mentor. Unfortunately, the study did not track how many youth from

[3]Applicants younger than 10 or older than 16 were excluded, as were families that were unable to complete a telephone interview in English or Spanish. Most affiliates also excluded from the research sample youth that were referred from certain social service agencies that they were contractually obligated to serve. Each affiliate stopped sample enrollment when it reached its quota from the researchers.

either the treatment or control group eventually got matched during their lifetimes, but anecdotally this number is thought to be quite low among the control group. A recent phone survey by DuBois et al. (2018) estimated that only 8.5% of control youth were ever in a BBBS relationship lasting a year or longer, relative to 56.7% of treatment youth. These estimates are based on survey responses of 296 of the original study participants and partial match history information from three of the eight affiliates involved in the study.

3 Data Sources

3.1 Mentoring Data

For our primary analysis, we link the analysis sample of 959 youth randomized by researchers in 1991 to tax records. Linking on name and date of birth, we obtain a 92% linkage rate that does not differ by treatment status. The main analysis sample for long-run outcomes thus consists of 883 youth from the original study for whom who we also observe administrative records. The original study dataset contains detailed survey information about the youth and their parents, but very little information on the mentors, who were not part of the study. Table 1 presents baseline descriptive statistics for the youth in the analysis sample; no statistically significant differences appear between treatment and control at baseline.

After 18 months, the researchers conducted interviews with the youth and their parents. The interviews, which were primarily done over the phone but sometimes in person, lasted approximately 30 minutes with each youth and 10 minutes with the parent. The result was several hundred survey outcome variables collected. Researchers asked parents of youth who received mentoring about quality of the match and their subjective opinions on whether their children's performance improved on several dimensions. Researchers asked youth a number of objective questions on whether they engaged in certain activities, as well as a variety of more subjective assessments of their self-concept, attitudes, and relationships with parents and friends.

To provide additional context for the experimental group, we also link to tax data administrative information on 9,000 youth and 30,000 volunteer mentors who applied to Big Brothers Big Sisters of Massachusetts Bay between 1991 and 2010. The Boston affiliate was selected because its digital records reach back unusually far, which is necessary for the analysis of long-run outcomes. However, the Boston affiliate differs from the eight cities selected by Grossman and Tierney (1998) in that, due to organizational features, it served almost exclusively boys during the relevant time-period.

3.2 Long-Run Outcomes

To reduce the number of hypotheses tested, we group the long-term outcomes that we construct from the tax records into two broad categories. We use the term "economic self-sufficiency" to refer to various

Table 1: Baseline Descriptives & Balance

	Baseline Variable	Overall Mean	Treatment Diff.	p
Tax Characteristics	Linked to Tax Data	0.92	0.021	0.22
	Parent's HH Income 1996-2000	32436.8	-1051.6	0.76
	Parent's Wage Income 1996-2000	24160.5	-532.8	0.73
Youth's Baseline Characteristics	Male	0.63	0.0079	0.81
	Age	12.2	0.024	0.8
	Minority	0.56	-0.022	0.52
Youth's Home Environment	Currently in counseling	0.23	0.023	0.43
	Family receiving cash welfare payments	0.42	0.0091	0.79
	Family history of domestic violence	0.29	0.027	0.38
	Family history of substance abuse	0.39	0.014	0.66
	Parent/guardian never married	0.24	-0.024	0.41
Parent & Case Manager Assessment	Few opportunities to do things	0.88	-0.019	0.38
	Not thinking well of him/herself	0.73	-0.00052	0.99
	Underachiever in school	0.52	0.0048	0.89
	Poor social skills	0.44	-0.0017	0.96
	Few friends	0.44	-0.029	0.38
Parent's Education Level	None	0.0011	0.0022	0.32
	Less Than High School	0.2	-0.0011	0.97
	High School Diploma	0.31	-0.021	0.5
	High School Equivalent	0.063	-0.00015	0.99
	Vocational/Technical/Business	0.044	-0.027+	0.056
	Some College	0.27	0.035	0.24
	Associates (2 Years)	0.038	-0.0037	0.77
	Bachelors (4 years)	0.056	0.0092	0.55
	Masters	0.017	0.0062	0.48
	Doctorate/PhD/JD/MD	0.0046	-0.00018	0.97

Notes: Sample is 883 youth matched with the tax data, with the exception of the first row which has a sample of 959.
+ p<0.1 * p<0.05 ** p<.01 *** p<.001

Table 2: Summary Statistics of Long-Run Outcomes

Variable	Sample		
	Women	Men	Full
Wages, age 25-30	13973.46	16526.25	15586.66
Log(wages, age 25-30)	8.965603	9.044668	9.014219
Above FPL	0.4738462	0.4892473	0.4835787
College, ever	0.6246154	0.5268817	0.5628539
Married, ever	0.4369231	0.3870968	0.405436
Divorced (full sample)	0.24	0.1756272	0.1993205
Divorced (if married)	0.5492958	0.4537037	0.4916201
Non-employment ever during ages 25-30	0.4492308	0.4964158	0.4790487
Unemployment Insurance	0.4492308	0.453405	0.4518686
EITC	0.8676923	0.7258065	0.7780294
Social Security	0.16	0.1577061	0.1585504
Incarceration	0.0061538	0.109319	0.0713477
Teen birth	0.4769231	0.172043	0.2842582
Deceased	0.0184615	0.0430108	0.0339751
Social Index	0.32	0.4139785	0.3793885
Social Index (dropping college)	-0.3046154	-0.1129032	-0.1834655
Economic Index	-0.5846154	-0.6182796	-0.605889

Notes: Sample is 883 youth matched with the tax data.

indicators of a person's financial well-being. In contrast, we also track a group of "social" or "behavioral" outcomes that are not direct measures of the subject's labor market performance, but may still be of interest to policymakers. Some of the social outcomes, such as college attendance, could also be viewed as inputs to economic self-sufficiency. Means of these outcomes are presented in Table 2.

Social/Behavioral Outcomes

College Attendance. We code college attendance based on whether a 1098-T form was ever present for the individual. These forms are filed by institutions receiving Title IV funding on behalf of all tuition-paying students.[4] These institutions would include some vocational and technical post-secondary programs that may not ordinarily be referred to as "college." Because this form is present only from 1999 to present, we do not observe the full sample during all typical college-going years; in our sample, the modal age in this year would have been 19, but subjects would have ranged from 16 to 24 years of age. For this reason, we use as our primary measure whether the individual ever had a 1098-T form from 1999 to present and show robustness to other definitions at specific ages. Half of the sample has ever received a 1098-T form, with approximately one-third having one report of college attendance between ages 20 and 24. Unfortunately, we do not observe any data beyond attendance such as graduation.

Incarceration. Our incarceration measure captures only those who were incarcerated in federal or state

[4]Chetty et al. (2017) find that these forms cover the "vast majority" of college students. They write that in practice colleges often file these forms for all students, even those that do not pay tuition.

prison between 2011 and 2014.[5] Although this time-period is older than would be ideal (the sample youth are already in their early 30s), we still detect a sizable share of the sample in prison, with 10% of the males fitting this definition. For the 60 subjects that are incarcerated, we also have a measure of the sentence length. The median reported sentence length for this group is 3 years, with a mean of 10 years.

Marriage & Divorce. We observe marriage and divorce based on the subject's tax filing status and subsequent changes in it. Forty percent of subjects have been married, and of those married, one-half have become divorced.

Teen Birth. We have two ways to detect teen birth. The first is through tax returns, in which we code an individual as having a teen birth if they ever claim a dependent who was born before the tax filer's 20th birthday. The second method, which uses vital statistics, codes individuals as having a teen birth if they appear on any birth certificate before their own 20th birthday. This combined measure flags teen births for half of the women in our sample and one-quarter of the men.

Mortality. Death records are obtained from the Social Security Death Master File. Two percent of females and 4% of males in the sample have died. In general, we do not drop these people from the analysis so as not to bias estimates of treatment effects.

Economic Self-Sufficiency Outcomes

Wages & Log Wages. Our primary economic outcomes are collected from W-2 forms, which are submitted to the IRS by employers on behalf of employees. We observe individuals' wage records even if they do not file tax returns, and this earned income measure is not contaminated by the presence of a spouse. We average wages over ages 25-30 for precision, but also explore robustness to measurement at various ages. Of the 883 youth in our long-run analysis sample, only 91 have zero wage income averaged over these five years. They are dropped when we transform wage earnings to log. As an alternative transformation with a different treatment of zeros, we create an indicator for whether the subject is over the federal poverty line (FPL) for a single person of $11,170 in 2015 (US Dept of Health and Human Services 2015). Average annual wage earnings are $16,500 for men and $14,000 for women. About half of the subjects have income above this federal poverty line.

Non-Employment. As an additional measure of economic activity, we code as an outcome the share of years in which an individual received no W-2 and was thus not working between the ages of 25 and 30. The mean of this variable is 25%.

Unemployment, Social Security, and Earned Income Tax Credit. We code indicators for whether the individual ever received each of these government benefits. Unemployment Insurance and Social Security

[5] Federal and state prisons comprise approximately two-thirds of the US prison population.

benefits are reported to the IRS on special information returns pertaining to each individual. Because Supplemental Security Income is not taxable, we do not observe it in our data. However, we would observe Disability Insurance. Claiming of the Earned Income Tax Credit is at the household rather than individual level. Only 16% of our sample has received taxable Social Security benefits, while nearly half have received unemployment benefits. Rates of EITC take-up are extremely high in the sample, with 73% of males and 87% of females filing for this tax credit at some point.

Indices of Long-Run Outcomes

As one way to facilitate interpretation of many long-run outcomes in light of multiple hypothesis testing, we also combined all of these long-run outcomes into just two indices. This method, which requires the researcher to make judgement calls on what the appropriate signs of each outcome are, builds on the approach of Kling et al. (2007). The only difference between our approach and that of Kling et al. is that Kling et al. converted all outcome variables into common units of z-scores, whereas our variables are already all binary so we apply no such normalization to the variance. The signs we use are as follows. The social index is constructed as + College + Marriage – Divorce – Teen Birth – Prison – Mortality. The economic index is + Above FPL – Ever Non-Employed – UI Benefits – Soc. Sec. Benefits.[6]

4 Re-Analysis of Short-Run Outcomes

Before examining outcomes in the long run, we performed a new analysis of the outcomes collected 18 months after randomization. The primary flaw of the original analysis by Grossman and Tierney (1998) that we overcome is that many insignificant or weakly significant outcomes are reported with no attention given to testing the joint hypothesis that the treatment did not affect any outcomes. To test this hypothesis, we turn to a more detailed analysis of 21 successive questions that were asked to youth about activities in which they may have participated. We focus on these "hard outcomes" outcomes, such as skipping school or stealing, because movements in these tangible outcomes seem the most likely to predict changes in long-run socioeconomic outcomes.

The ITT estimate without controls for each of these 21 behavioral outcomes is shown in Figure 1. We have re-signed outcomes as appropriate so that the positive direction always indicates "better" behavior, and the outcome in all cases is binary. Perhaps the most salient result here is that only three of the 21 outcomes are significantly different from zero at the 5% level. Due to the large number of hypotheses tested, this overall lack of significance cautions against inferring much from any of these results. If these were 21 independent variables unaffected by treatment, we would expect 1.05 estimates to be significant at this level

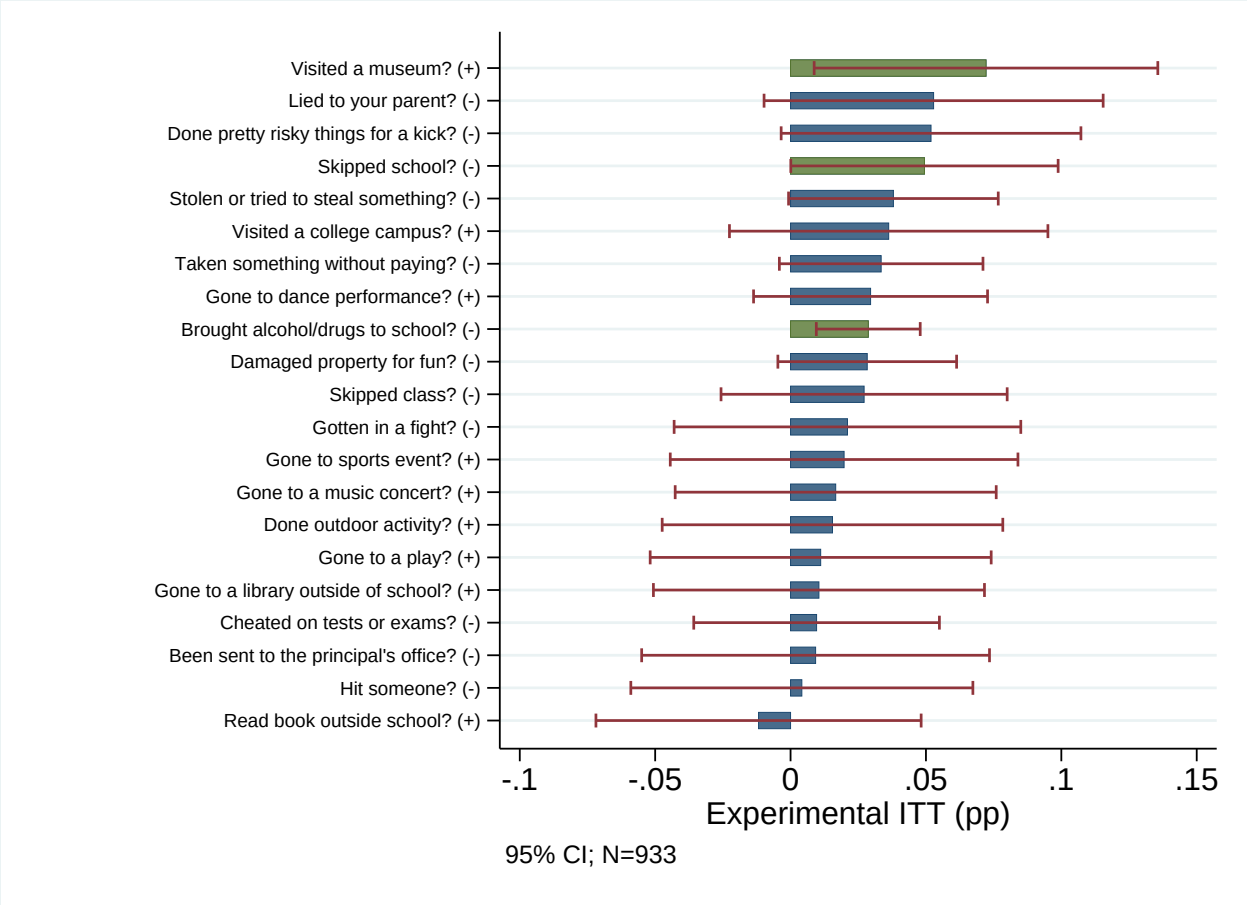
[6]The correct signs are less apparent for the economic index. However, our results are qualitatively robust to alternative definitions of this index.

due to chance alone (.05 times 21).

The fact that so few estimates are significant is consistent with the treatment not affecting subjects, but it is also consistent with the effects of the treatment being difficult to measure and noisy in this small sample. To discriminate between these hypotheses, we outline three strategies for multiple-hypothesis testing, which we apply consistently across short-run and long-run outcomes. The strategies are an index test, a directional permutation test, and a joint F test.

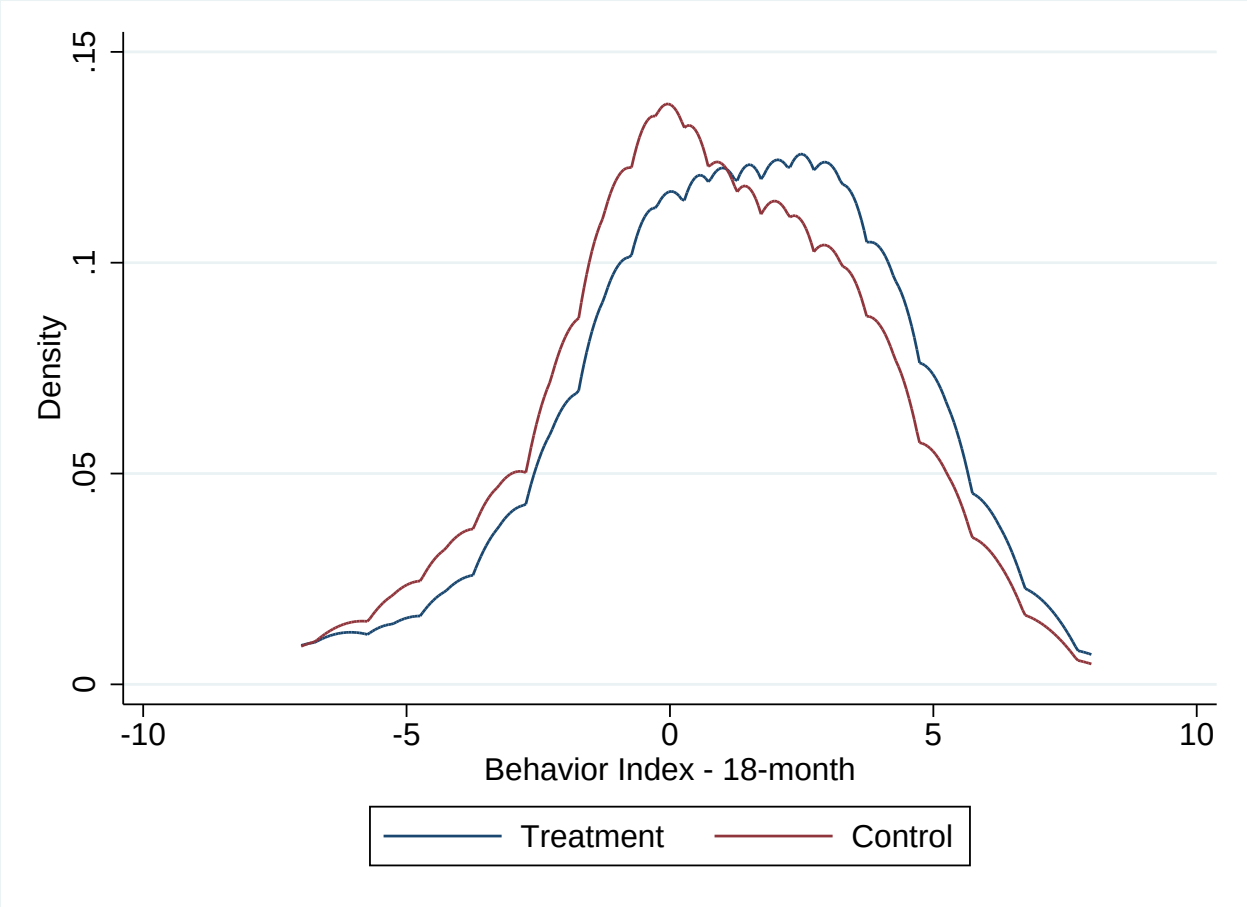
As we described for the long-run outcomes, we also build an index of short-run behavioral outcomes following the method of Kling et al. (2007). The signs of the outcomes in constructing the index are as reported in Figure 1. The range of the composite behavioral score is from -12 to 9, as questions were asked about 9 outcomes that seem desirable and 12 outcomes that constitute misbehavior, such as hitting someone. Because these 21 questions were asked both at baseline and at the conclusion of the study, we can build indexes of youth behavior pre- and post-treatment.

Figure 1: 18-month Follow Up: Have you ever in last year?



Notes: Coefficients correspond to separate regressions, with no controls. Sample size is 933 youth with non-missing behavioral outcomes. 95% CI shown.

Figure 2: 18-month Follow Up Behavior Index



Notes: Sample size is 914 youth with non-missing behavioral outcomes or baseline reports.

Table 3: 18-month Follow Up Behavior Index

	Outcome Behaviors Index			Baseline Behaviors Index
	(1)	(2)	(3)	(4)
Treatment	0.54**	0.60***	0.59***	-0.099
	(0.19)	(0.17)	(0.17)	(0.18)
Baseline Behaviors		X		
Baseline Behav. Index			0.52***	
			(0.032)	

Notes: Sample size is 914 youth not missing any behavior outcomes or baselines. Standard errors in parentheses. + p<.10 * p<.05 ** p<.01 *** p<.001

Figure 2 plots the distribution of this behavioral index by experimental group after treatment. The treatment group appears right-shifted, and Table 3 confirms with a high degree of statistical significance that treated youth are .54 points higher on this index. The difference is robust to controlling for participants' behaviors at baseline, either by means of indicator controls (Column 2) or by construction of the same index at baseline (Column 3). As a placebo, Column 4 regresses the behavioral index at baseline on treatment status, which yields a small and insignificant coefficient. The distributions of the behavioral index for treatment and control at baseline are also plotted in Figure A.6, with no discern-able difference prior to treatment.

Our second method for inference is a permutation test which formalizes the intuition that, although so few comparisons are statistically significant, nearly every comparison favors the treatment group. This fact alone cannot discriminate treatment effect from sampling variation; if these outcomes were very highly correlated, then this would just be a re-statement of the same insignificant finding 21 times. Alternatively, the diverse outcomes might be better thought of as independent measurements of one or more latent outcomes (e.g., "behavior") in which the error in measurement is uncorrelated. In such a model, it would be rare to recover so many correct-signed differences in outcomes if the underlying latent variable did not actually differ across treatment and control groups. To formalize this intuition and gauge the meaningfulness in our data of so many same-signed treatment effects, we randomly permute the treatment variable within the sample and count the share of simulations in which 18 or more treatment effect estimates go in the expected direction. The random permutation of treatment maintains the correlation of each observation's measured outcomes. Panel A of Figure 3 plots the distribution of this test statistic over 1,000 simulations. Such an extreme value as we observe in our data was observed in only 3 simulations, yielding a p-value of .0003 for the null hypothesis that the treatment did not cause an improvement in behavior. As a placebo, Figure A.4 plots the same analysis for behavioral measures at baseline: Nine of 21 outcomes are in the appropriate direction, and of our 1,000 permutations we observed 756 that contained at least such a good outcome ($p = .756$).

Whereas the index test and permutation test gained power due to the obvious directionality of the outcomes, the last test that we employ is an unsigned joint F-test. Commonly referred to a as a "balance test," we simply regress the treatment indicator on the outcomes to see whether the outcomes jointly predict treatment status. This regression yields an F-statistic, which does not correspond neatly to a null hypothesis, because the regression is not structural. In order to benchmark the magnitude of the test statistic, we compare it to the distribution of F-statistics when the treatment variable is permuted. The first column of Table 4 presents the results of this analysis. The F-statistic from the regression was 1.3, and we obtained larger F-statistics in 24% of permutations ($p = .24$). Thus, we are unable to reject the null hypothesis from this unsigned, and therefore lower-powered, test.[7]

A summary of our re-analysis of short-run outcomes is as follows. Many self-reported outcomes were collected, yielding few individually significant estimates of treatment effects. However, a disciplined analysis of the primary outcomes collected seems to in most cases provide evidence that the treatment improved participants' behaviors, at least according to their self-reports. With the short-run findings as motivation, we turn in the next section to examining the effects of treatment on long-run social and economic outcomes that can be measured in administrative datasets.

5 Long-Run Experimental Outcomes

As with the short-run analysis, the long-run analysis should not be interpreted as quantifying the effect of having a mentor versus not. Rather, the estimates are of intent-to-treat effects. As discussed in Section 2.2, upon the conclusion of the study in 1993, all participants were eligible to receive mentoring regardless of experimental status, though survey evidence suggests that much of the control group never got matched with a mentor. These long-run intent-to-treat estimates likely reflect a combination of two channels, neither of which we quantify well. First, the treatment group was more likely to ever during their lives be matched with a mentor. Second, the treatment group was able to have a mentor sooner; this may provide services at more formative ages or increase the total length of the match. If families choose to apply at the times that youth most need mentors (e.g., during crises), then the ITT may also contain the effect of having access to a mentor when he or she is most needed rather than a year or two later.

As was the case with the short-run behavioral outcomes, we again analyze a great many outcomes in a small sample. To gauge the significance of our results in light of the number of hypotheses tested, we use the same three strategies: the index test, the permutation test on signs, and the unsigned F test.

[7] As a robustness check, Column 2 reports this test within the sample that was matched to the tax data; the p-values are somewhat lower for among this sample.

5.1 Absolute Magnitudes

We first focus on the absolute magnitudes of our point estimates. Subsequently, we present relative magnitudes when each outcome is rescaled by its correlation with parental income.

5.1.1 Long-Run Social & Behavioral Outcomes

Table 5 presents the results of separate regressions of eight separate outcomes on treatment with no controls. The first outcome presented is the summary index of all social outcomes, defined in Section 3.2. Although the magnitude is difficult to interpret, there is strong evidence that the treatment has had some effect on this index of social outcomes ($p < .001$).

To better understand which components of the index may be affected by the treatment, we turn to the remaining seven regressions. The most significant effect is on college attendance, which increases by 10 percentage points.[8] Although this estimate is individually significant at the $p = .01$ level, we do not want to over-state the individual significance of this or any individual p-values in the face of the number of hypotheses tested. Still, the point estimate represents a 19.6% increase over the control group's rate of ever attending college. The treatment group is 6 percentage points (16.2%) more likely to have been married at some point, with no discernible up-tick in divorces (if anything, a non-significant decline). The rate of teen births is 5 percentage points (16.1%) lower in the treatment sample. The indicator for mortality is not individually significant, but the point estimate is also favorable to the treatment group. The same is true of incarceration, although among the small sample of participants who have been incarcerated the average sentence length of the treatment group is significantly lower. Each of these regressions should be interpreted as suggestive evidence of effects on the individual outcome in light of the number of hypotheses tested.

Finally, the remaining multiple-hypothesis tests for social outcomes are presented in Figure 3 and Table 4. The p-value associated with all seven of the social outcomes pointing in the right direction is .013, and the p-value associated with these outcomes predicting treatment is .004. Much but not all of the outcomes' power in predicting treatment stems from college attendance, which is unsurprising given that treatment is such an individually significant predictor of college. Excluding college, the permutation p-value for the unsigned F-test using only the other the other outcomes is .029 (Column 4).

5.1.2 Long-Run Economic Self-Sufficiency Outcomes

Analogously to Table 5, Table 6 presents results on long-run economic outcomes. The overall significance level of the social outcomes is not mirrored in economic outcomes.

[8] Figure A.3, in the appendix, shows the robustness of this result to measuring college attendance at only specific ages among applicable cohorts.

Table 4: F-Tests for Balance of Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	18-month behavioral		**Social**		**Economic**	**All Long-Run**	
21 behavior outcomes	Full sample	Matched to tax data					
Wages, non-employment, UI, EITC, Soc. Sec.					x	x	x
College			x			x	
Marriage, Incarceration, Death			x	x		x	x
Divorce			x	x		x	x
Sentence Length			x	x		x	x
N	933	862	883	883	883	883	883
F	1.3	1.72	4.14	3.42	1.34	3.6	2.7
Asymptotic p	0.17	0.023	0.00017	0.0024	0.24	0.000028	0.002
Permutation p	0.24	0.061	0.004	0.029	0.29	0.0011	0.0212

Notes: The F-statistic of each column is from a regression of treatment on the outcomes indicated. The asymptotic p-values is from the regression output, and the permutation p-value is the share of F-statistics larger than the observed one when the treatment variable is permuted 10,000 times.

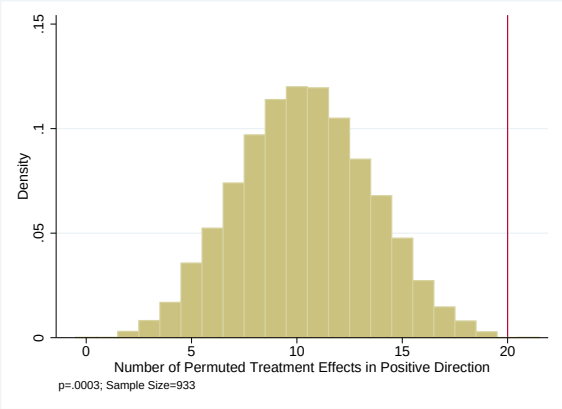
Table 5: Long-Run Results: Social Outcomes

	Social Index	College	Married	Divorced Given Married	Teen Birth	Deceased	State/Fed Incarcerated 2011-present	Sentence Length (Yrs)
Treatment Effect	0.22***	0.100**	0.060+	-0.077	-0.053+	-0.0018	-0.0015	-9.28+
	(0.061)	(0.033)	(0.033)	(0.053)	(0.03)	(0.012)	(0.017)	(5.27)
Constant	0.27	0.51	0.37	0.53	0.31	0.035	0.072	14.9
	(0.042)	(0.024)	(0.023)	(0.039)	(0.022)	(0.0089)	(0.012)	(5.16)
N	883	883	883	358	883	883	883	60
Asymptotic 2-tailed p	0.00037	0.0028	0.067	0.15	0.08	0.88	0.93	0.084
Exact p, 2-tailed	0.0004	0.0028	0.0685	0.1478	0.0819	0.8876	0.9365	0.0622

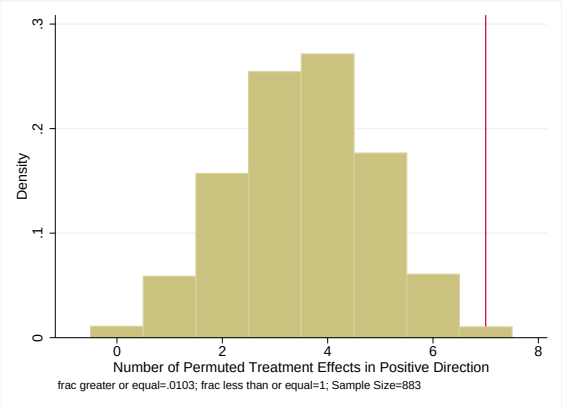
Notes: Standard errors in parentheses. No controls. + p<.10 * p<.05 ** p<.01 *** p<.001

Figure 3: Permutation Tests for Number of Correctly Signed Outcomes

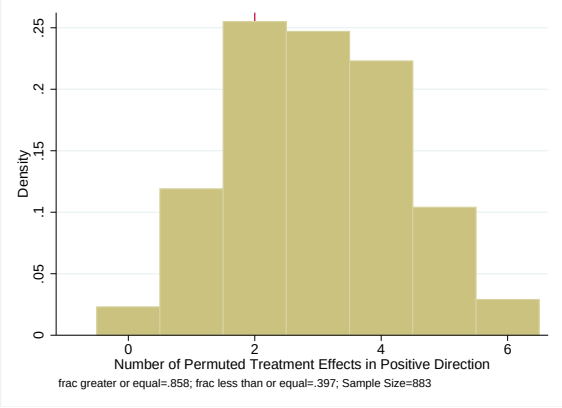
(a) 18-Month Behavioral Outcomes



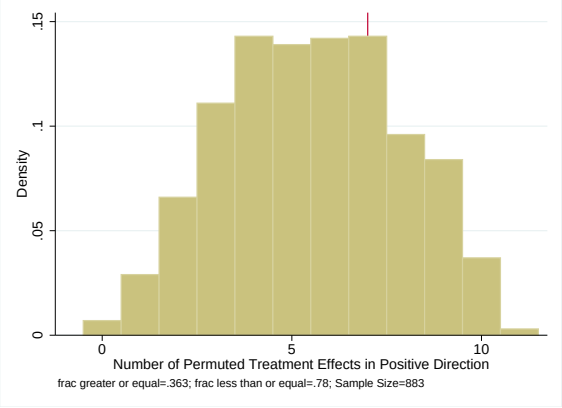
(b) All Social Outcomes



(c) Economic Outcomes



(d) All Social and Economic Outcomes



Notes: To construct each figure, 10,000 simulations were constructed in which treatment was permuted. For each simulation, we regressed each outcome of interest on permuted treatment and counted the number of treatment effects going in the desirable direction. The sample size for panel A is 933 youth with non-missing behavioral outcomes and for the remaining panels is 883 subjects matched to tax data. Social outcomes are college, marriage, divorce, teen birth, death, incarceration, and sentence length. Economic outcomes are wages, log wages, and indicators for non-employment, unemployment insurance, EITC, and Social Security.

Table 6: Long-Run Results: Economic

	Economic Index	Wages 25-30	Log(Wages 25-30)	Share non-employed years 25-30	UI Benefits	EITC	SS Inc
Treatment Effect	-0.025	-1632.6	-0.053	0.018	-0.071*	-0.016	0.019
	(0.067)	(1270.5)	(0.12)	(0.024)	(0.033)	(0.028)	(0.025)
Constant	-0.59	16424.2	9.04	0.26	0.49	0.79	0.15
	(0.048)	(1006.2)	(0.087)	(0.017)	(0.024)	(0.02)	(0.017)
N	883	883	792	883	883	883	883
Asymptotic 2-tailed p	0.71	0.2	0.65	0.28	0.034	0.58	0.44
Exact p, 2-tailed	0.7063	0.1989	0.6545	0.2852	0.0345	0.5769	0.4382

Notes: Standard errors in parentheses. No controls. + p<.10 * p<.05 ** p<.01 *** p<.001

The index of economic outcomes defined in Section 3.2 does not differ significantly by treatment status. As the signs by which the components should enter this index are arguably less clear than in the cases of the other indexes we use, the regressions of individual outcomes may be of more interest than the index. However, turning to its individual components, only one variable is individually significant at the 5% level: unemployment insurance benefits receipt is lower in the treatment. But in light of the number of hypotheses tested, we do not want to read much into this sole significant outcome.

Moving to other point estimates, wages and log wages both decrease, and non-employment increases.[9] The only point estimates that might be said to go in the favorable direction are the decreases in unemployment insurance and EITC.[10]

Although the point estimates for economic outcomes are generally in the unfavorable direction, we are unable to lend support to the hypothesis that treatment worsened economic outcomes (or moved them at all). The unsigned permutation p-value we obtain from using these economic outcomes to predict treatment is .29. The permutation test on the number of correct-signed outcomes showed that 85.8% of permutations would have yielded more favorable results, while 39.7% yielded less favorable. In contrast to the social outcomes, the economic outcomes may be more highly correlated as wages, log wages, and non-employment

[9] Figure A.3 shows that this insignificant result persists when defining income at different ages.

[10] Even still, it is not abundantly clear that receipt of unemployment benefits should be regarded as a negative outcome as it signifies previous formal employment. Similarly, EITC also requires working and often does not amount to much money without dependents.

are all just re-parameterizations of each other.

In terms of the absolute magnitudes, this experiment is much too small to say much about economic outcomes. Reasonable effects on wages between -28% and +17% are all within our confidence intervals. To put the imprecision into perspective, suppose that treatment had caused 10% of treated youth to attend college for exactly 2 extra years (Figure A.2 in the Appendix shows that two-year colleges are common for this sample). Even if the return per each year of schooling is a generous 10% to wages, the average effect of the treatment on the full sample would only be a 2% increase in wages, which is well below the detectable threshold in the data.[11] In fact, even if one took the cross-sectional relationship between wages and social outcomes as causal (which seems likely to be an over-statement), the movements on the social variables would only be expected to generate a 9% increase in wages. Thus, given the positive results on social outcomes it may be reasonable to suspect the treatment increased wages by as much as 9%, although this experiment would be too small to detect a result of that magnitude.

5.2 Relative Outcome Magnitudes

Given the results of varying precision on absolute magnitudes, our next question is whether mentoring improves social outcomes more than it does economic outcomes. Alternatively, economic outcomes may merely be measured with less precision.

To disentangle these two hypotheses, we normalize the scales of all the outcomes according to their relationships with parent income to ask how much treated youth have "moved up" the parent income ladder on either dimension. For context, Figure 4 plots the income distributions of the youths' parents as well as that of mentors for whom we have data from the Boston affiliate. When mentors' income is measured during their middle-aged years, the average mentor would rank approximately 30 percentiles higher in the national distribution of parent income than the parents of youth in the study.

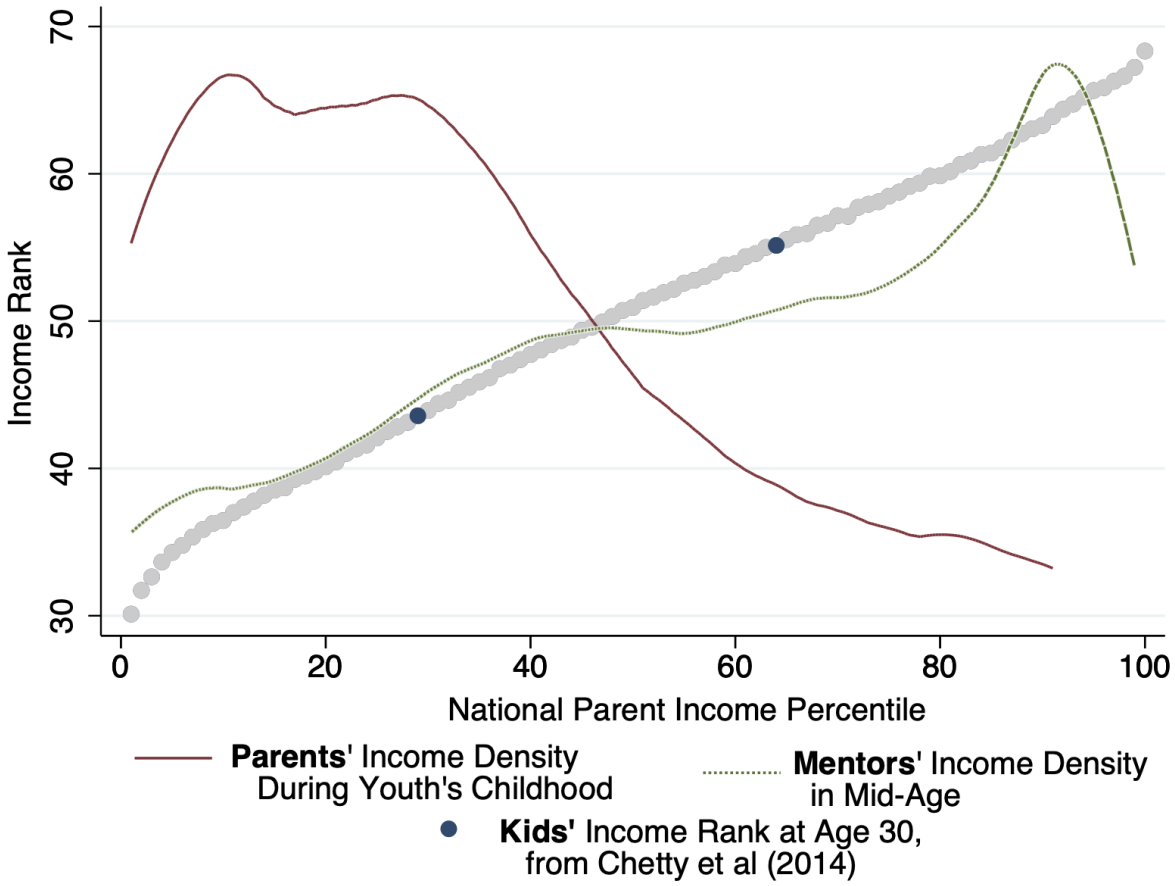
Figure 5 plots the magnitudes of these indexes and their components when all outcomes are re-scaled by their within-sample correlations with parent income (in percentile bins, following Chetty et al. (2014)).[12] Mechanically, a 1 unit increase in any outcome now corresponds to the expected outcome of kids belonging to parents who were 1 percentile higher. On the social index, the outcomes of treated youth are comparable to youth from 21.5 percentiles higher, about two-thirds of the way to where we would expect the children born to the mentors to fall.[13] The insignificant point estimate on economic self-sufficiency from the previous section is relatively small in magnitude when converted to parent percentiles, at -4.2 percentiles.

[11] 10% of the sample would have outcomes of .2 log points higher.

[12] The re-scaling was performed by dividing each outcome by the coefficient on parent rank when regressing the outcome on parent rank.

[13] Because this estimate is an ITT (22% of treatment youth were not assigned a mentor by the time of the follow-up survey), it is possible that the ATE of having a mentor is higher than 21.5 percentiles.

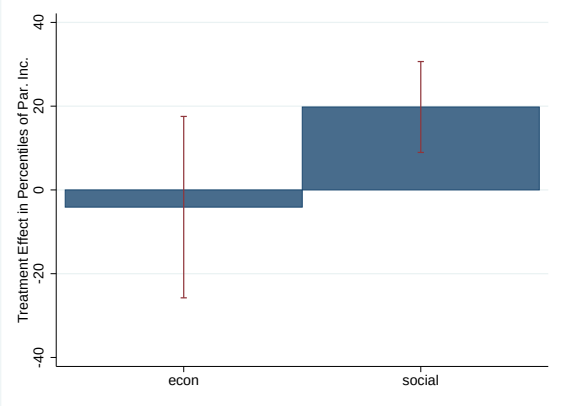
Figure 4: Boston Volunteer & Parent Income Ranks



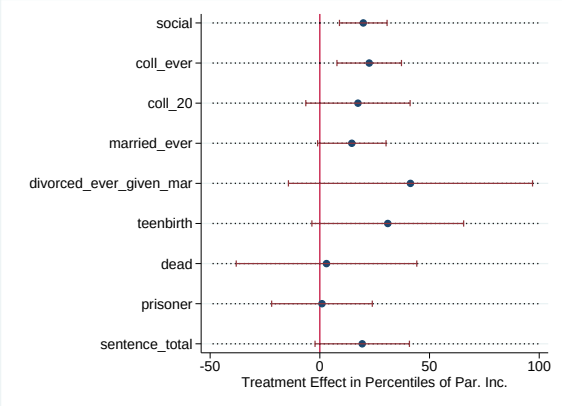
Notes: Sample size is 883 youth linked to tax data.

Figure 5: Treatment Effects, in Parent Income Percentiles

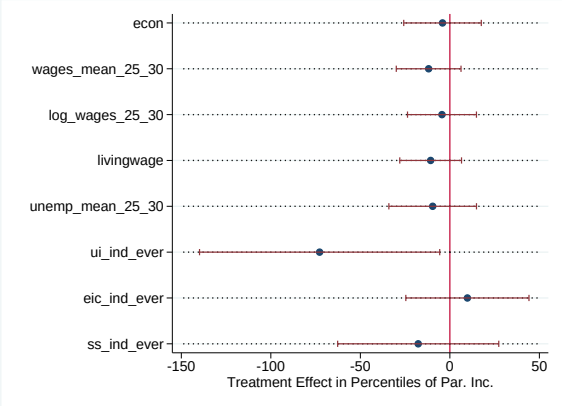
(a) Indexes



(b) All Social Outcomes



(c) All Economic Outcomes



Notes: Sample is 883 youth matched to tax data. 95% confidence intervals shown.

The null hypothesis that the treatment effects on the two indices are the same can be tested using either a permutation test or an asymptotic test. In the permutation test, we conducted 10,000 simulations in which the treatment variable was randomly re-assigned. In only 2.43% of the random permutations was a difference between the coefficients of at least the actual magnitude observed, corresponding to a two-tailed p-value of .0243. For the asymptotic test, we applied a seemingly unrelated regression model, which allows for asymptotic inference across equations when the error terms of the regressions may be correlated (Zellner and Huang, 1962; Zellner, 1963). The χ^2 statistic from the test of coefficient equality is 5.17, corresponding to a p-value of .0229. From either test, we can conclude at the $\alpha = .05$ level that mentoring enables kids to climb higher up the rungs of the parent income ladder in terms of their social outcomes than it does for economic self-sufficiency.

Is it surprising to find that the economic benefits of mentors are smaller than the social benefits? On the one hand, many of the social outcomes that we measure, such as college attendance, are also likely to be causal factors of kids' income. Thus, it is difficult to imagine that the treatment group's increased education does not at least somewhat increase their earnings potential.[14] But on the other hand, there are likely a great many external factors that also contribute to kids' earnings potential, such as housing instability or under-resourced public schools, that are not being varied by the mentoring intervention. Even if the mentors could completely remove all disparities between kids from richer and poorer families in terms of their social choices like criminality or whether to attend college, this would probably not be enough to equalize earnings opportunities without first resolving the other structural inequalities of society. Still, given that resources are finite and this mentoring intervention does have some proven benefits, we conclude with a discussion of costs and alternative interventions.

6 Discussion and Conclusions

Exposure to successful adults is an important input to success to which not all kids have equal access. Structured community-based mentoring programs such as Big Brothers Big Sisters create artificial mentoring relationships for disadvantaged youth. Our re-analysis of a thirty-year-old RCT that randomized kids' eligibility for mentoring finds that these mentors had and continue to have substantial effects on participants' behavioral and social outcomes. The effect size for our index of long-run social success is two-thirds as high as one would expect if the participants were actually raised in families as economically successful as the mentors were. Although the effect of mentoring on our index of economic success was significantly smaller and not detectable, we also cannot rule out fairly large effects.

[14] Of course, we cannot rule out models that the intervention decreased earnings, for instance by changing occupation preferences.

The mechanism of these effects was not that academic support nor financial investments in the child's education. In fact, mentors were instructed not to engage in these activities. Instead, mentors served as role models for real-world decision making, which may have improved children's perceptions of the rewards to foregoing impulses. Mentors may also have exposed youth to new careers and life trajectories, such as college, and may made it easier for kids to have positive and trusting relationships with adults, such as their parents and teachers. In theory, there could have also been room for mentors to harm youth by setting unrealistic expectations, increasing awareness of their own hardship, or diminishing trust following an unsuccessful relationship.[15] However, if such wrinkles did occur, the positive effects at the mean suggest that benefits of participation tended to dwarf such potential harm.

Based only on the effects on measures of kids' social success, this program seems reasonably attractive from a cost-benefit analysis. The cost per match is on the order of $2,000-$3,000 per year. Locations with more matches tend to have lower costs per match, consistent with some fixed costs of program administration; Figure 6 plots the cost per match across all locations in the study.[16] As the mentors receive no financial compensation, the primary costs should be thought of as the professional caseworkers who arrange and support these matches.

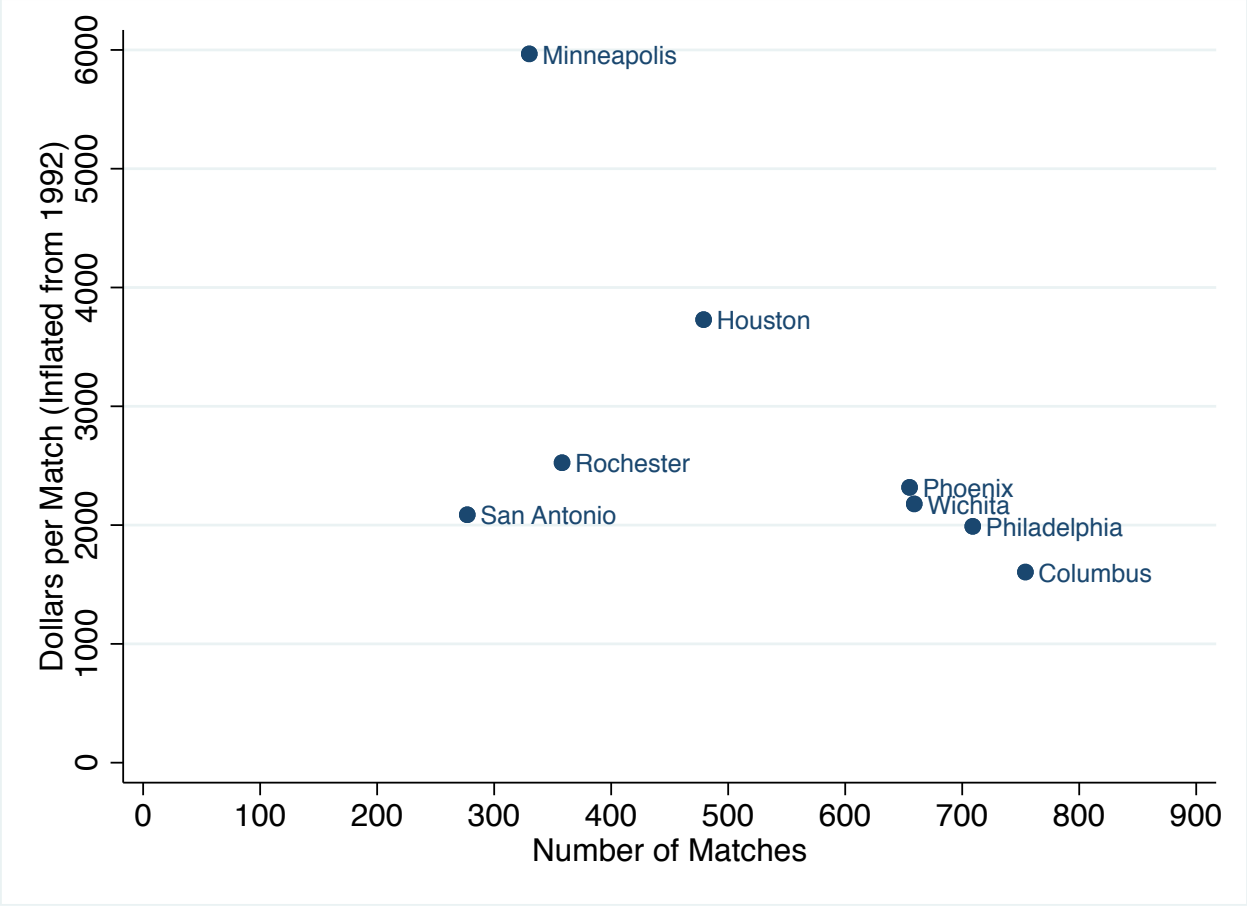
For the closest research-evaluated program of comparison, the increase of 10 percentage points to college attendance we observe is slightly larger than the magnitude of the effect of another program that included mentoring evaluated by Rodriguez-Planas (2012). The Quantum Opportunity Program, which provided mentoring, conditional cash transfers, and tutoring to low-performing high school students, raised the likelihood that participants attended post-secondary training by 7.4 percentage points. However, that effect came at a much higher price tag of $25,000 per enrollee. By this rough yardstick, the professionally supported volunteer mentoring that we have evaluated seems much more cost-effective.

Aside from mentoring, other comparisons can be drawn to a growing literature evaluating other programs that aim to increase opportunities for disadvantaged youth. The Becoming a Man program evaluated by Heller et al. (2015) delivers cognitive behavioral therapy in group settings to at-risk youth. Although the volunteer mentors in our study typically have no training in therapy, the professional caseworkers that supervise the relationship through regular conversations may effectively enable the mentors to play a similar role in challenging youths' unproductive ways of thinking that surface during outings. The authors estimated the benefits of BAM to be between $1,100 and $1,850 per participant per year, but a direct comparison of benefits is difficult because we do not access to the type of short-run arrest records used by the Heller et al. (2015)

[15] Salient examples of such social interventions gone awry include McCord (1978) and Dishion et al. (1999).

[16] Costs seem to be similar today. For instance, the Boston branch's budget in fiscal year 2016 was $6,783,022 according to their publicly available IRS F-990. That same year, they had 2,441 active matches, which implies a cost of $2,778 per active match.

Figure 6: Program Cost



Notes: Number of total active matches and agency budgets are as reported in Tierney et al. (1995). Budgets are converted to 2018 dollars.

evaluation. Although the intervention was of a very different nature, the financial aid experiment evaluated by Bettinger et al. (2012) produced increases in college attendance similar to what we have observed, though at a lower cost of about $700 per student. The recent Moving to Opportunity evaluation by Chetty et al. (2016) reported increases in college attendance among young participants of only 2.5 percentage points, at a cost of $2,660 per family.[17]

In conclusion, our findings suggest that structured mentoring programs hold a great deal of promise to bring socioeconomic opportunities to disadvantaged youth. Although our results lead us to conclude that mentors should not be touted as a panacea for all of economic inequality, structured mentoring relationships can lead kids to make better decisions and contribute tangibly to their social development. Such effects are in turn likely to contribute to their longer-term economic success.

References

BBBSA, "About Us," September 2016.

Bell, Alex, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen, "Who Becomes an Inventor in America? The Importance of Exposure to Innovation," *The Quarterly Journal of Economics*, 2019.

Bergman, Peter, Raj Chetty, Stefanie DeLuca, Nathaniel Hendren, Lawrence F Katz, and Christopher Palmer, "Creating Moves to Opportunity: Experimental Evidence on Barriers to Neighborhood Choice," Working Paper 26164, National Bureau of Economic Research August 2019.

Bettinger, Eric P., Bridget Terry Long, Philip Oreopoulos, and Lisa Sanbonmatsu, "The Role of Application Assistance and Information in College Decisions: Results from the H&R Block Fafsa Experiment*," *The Quarterly Journal of Economics*, August 2012, *127* (3), 1205–1242.

Bureau, US Census, "Historical Living Arrangements of Children," November 2017.

Chetty, Raj, John N Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan, "Mobility Report Cards: The Role of Colleges in Intergenerational Mobility," Working Paper 23618, National Bureau of Economic Research July 2017.

_ , Nathaniel Hendren, and Lawrence Katz, "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Project," *American Economic Review*, 2016, *106* (4).

[17] To report the cost figure in present-day terms, we have reported the cost of a similar program recently implemented by Bergman et al. (2019).

_ , _ , **Maggie R. Jones, and Sonya R. Porter**, "Race and Economic Opportunity in the United States: an Intergenerational Perspective," *The Quarterly Journal of Economics*, May 2020, *135* (2), 711–783.

_ , _ , **Patrick Kline, and Emmanuel Saez**, "Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States," *The Quarterly Journal of Economics*, 2014, *129* (4), 1553–1623.

Cunha, Flavio, "Eliciting Maternal Beliefs about the Technology of Skill Formation," 2013 Meeting Paper 1051, Society for Economic Dynamics 2013.

Dishion, T. J., J. McCord, and F. Poulin, "When interventions harm. Peer groups and problem behavior," *The American Psychologist*, September 1999, *54* (9), 755–764.

DuBois, D. L., N. Portillo, J. E. Rhodes, N. Silverthorn, and J. C. Valentine, "How Effective Are Mentoring Programs for Youth? A Systematic Assessment of the Evidence," *Psychological Science in the Public Interest*, August 2011, *12* (2), 57–91.

DuBois, David L., Carla Herrera, and Julius Rivera, "Investigation of Long-Term Effects of the Big Brothers Big Sisters Community-Based Mentoring Program: Final Technical Report for OJJDP," February 2018.

Grossman, J. B. and J. P. Tierney, "Does Mentoring Work?: An Impact Study of the Big Brothers Big Sisters Program," *Evaluation Review*, June 1998, *22* (3), 403–426.

Heckman, James J., "Skill Formation and the Economics of Investing in Disadvantaged Children," *Science*, June 2006, *312* (5782), 1900–1902.

_ **and Stefano Mosso**, "The Economics of Human Development and Social Mobility," *Annual Review of Economics*, 2014, *6* (1), 689–733.

Heller, Sara B., Anuj K. Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A. Pollack, "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago," Working Paper 21178, National Bureau of Economic Research May 2015.

Herrera, Carla, Jean Baldwin Grossman, Tina J. Kauh, and Jennifer McMaken, "Mentoring in Schools: An Impact Study of Big Brothers Big Sisters School-Based Mentoring," *Child Development*, January 2011, *82* (1), 346–361.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz, "Experimental Analysis of Neighborhood Effects," *Econometrica*, 2007, *75* (1), 83–119.

Klinger, Laura, "Big Brothers Big Sisters of America Announces Local Agencies and Boards of the Year," July 2018.

McCord, Joan, "A thirty-year follow-up of treatment effects.," *American Psychologist*, 1978, *33* (3), 284–289.

of Health and Human Services, US Dept, "Prior HHS Poverty Guidelines and Federal Register References," November 2015.

of Justice, The U.S. Dept, "Mentoring Children Affected by Incarceration: An Evaluation of the Amachi Texas Program," Technical Report 2011.

Rodriguez-Planas, Nuria, "Longer-Term Impacts of Mentoring, Educational Services, and Learning Incentives: Evidence from a Randomized Trial in the United States," *American Economic Journal: Applied Economics*, October 2012, *4* (4), 121–139.

Sacerdote, Bruce, "How Large are the Effects from Changes in Family Environment? A Study of Korean American Adoptees," *The Quarterly Journal of Economics*, February 2007, *122* (1), 119–157.

Spencer, Renee, "It's Not What I Expected: A Qualitative Study of Youth Mentoring Relationship Failures," *Journal of Adolescent Research*, July 2007, *22* (4), 331–354.

Tierney, Joseph P., Jean Baldwin Grossman, and Nancy L. Resch, "Making a Difference: An Impact Study of Big Brothers/Big Sisters (Re-issue of 1995 Study)," Technical Report 1995.

Wit, David J. De, Ellen Lipman, Maria Manzano-Munguia, Jeffrey Bisanz, Kathryn Graham, David R. Offord, Elizabeth O'Neill, Deborah Pepler, and Karen Shaver, "Feasibility of a randomized controlled trial for evaluating the effectiveness of the Big Brothers Big Sisters community match program at the national level," *Children and Youth Services Review*, March 2007, *29* (3), 383–404.

Zellner, Arnold, "Estimators for Seemingly Unrelated Regression Equations: Some Exact Finite Sample Results," *Journal of the American Statistical Association*, December 1963, *58* (304), 977–992.

_ **and David S. Huang**, "Further Properties of Efficient Estimators for Seemingly Unrelated Regression Equations," *International Economic Review*, 1962, *3* (3), 300–313.

Appendix

A All Outcomes Reported in Grossman and Tierney (1998)

Although no formal pre-analysis plan was made public, documents drafted at baseline indicate that five broad hypotheses were to be tested. These five hypotheses also formed the structure of the paper. Unfortunately, the specific variables used to test each hypothesis were not generally specified. Control variables were also not pre-specified. The broad hypotheses were as follows:

- Social and Cultural Enrichment. Examples given pre-analysis are attending a play, musical performance, or sporting event.
- Attitudes toward oneself. Includes self-concept, sense of competence, and self-esteem.
- Relationships with family and friends.
- Schooling. Includes school attendance, performance, attitudes toward school, and school behavior.
- Antisocial activities. Includes disciplinary problems in school, alcohol/drug use, and involvement in the criminal justice system.

The table below summarizes the findings reported in tables by Grossman and Tierney (1998). All regressions include several baseline controls. All significant findings indicate beneficial effects, and we have annotated significant results as follows: + p<.10 * p<.05 ** p<.01 *** p<.001.

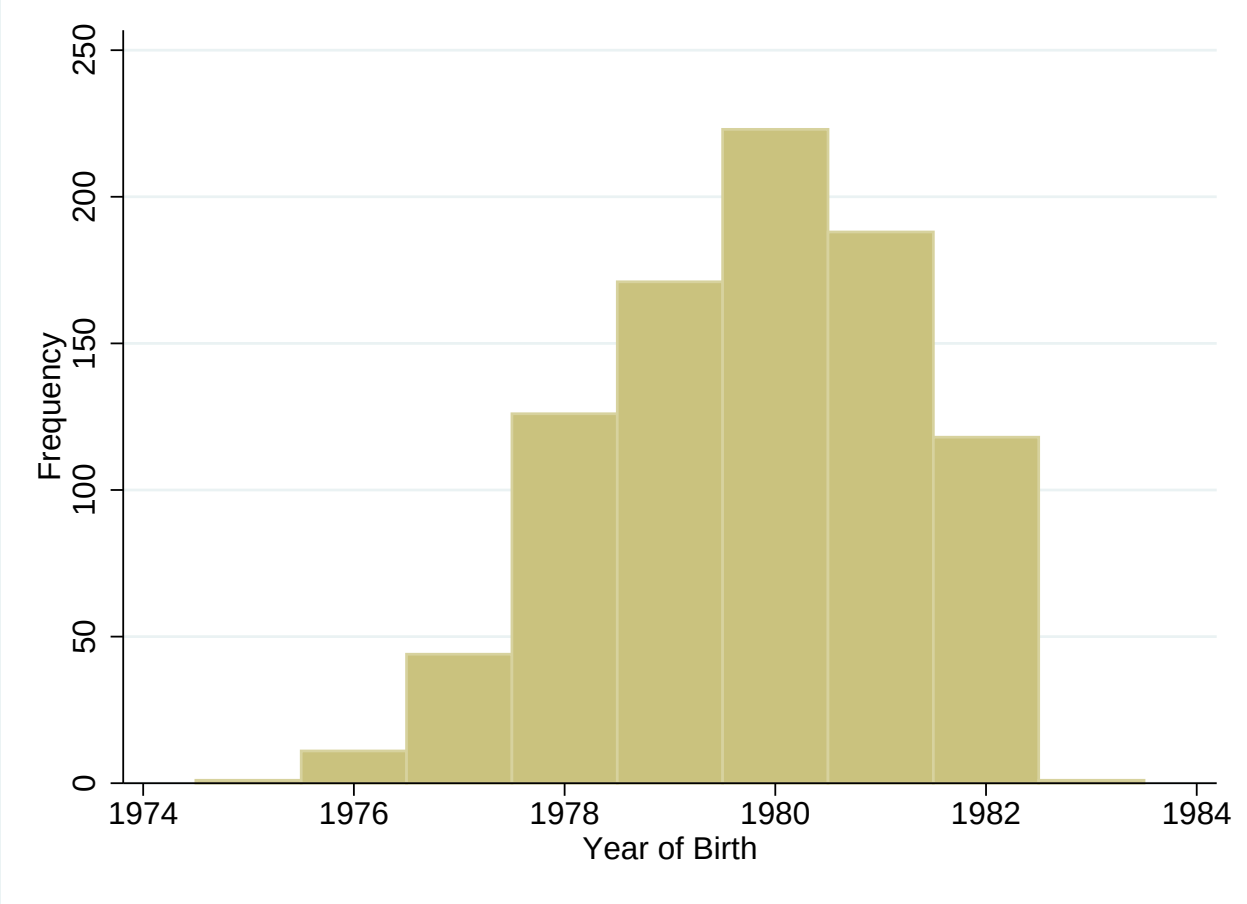
Broad Hypothesis	Measure Reported	Level of Sig.	Direction
Social & Cultural Enrichment			
	Hours spent on social/cultural activities		beneficial
	Number of social/cultural events attended		harmful
Self Concept			
	Global Self-Worth		beneficial
	Social Acceptance		beneficial
	Self-Confidence		beneficial
Family & Peer Relationships			
	Summary Parental Relationship Measure	*	beneficial
	Parental Trust	*	beneficial

	Parental Communication		beneficial
	Parental Anger & Alienation		harmful
	Intimacy in Communication		beneficial
	Instrumental Support		harmful
	Emotional Support	+	beneficial
	Conflict		beneficial
Academic Outcomes			
	GPA	+	beneficial
	Skipping School	**	beneficial
Antisocial Behaviors			
	Drug Use	*	beneficial
	Alcohol Use	+	beneficial
	Hitting	*	beneficial
	Stealing		beneficial
	Damaging Property		beneficial

Of the 20 outcomes reported by Grossman and Tierney (1998), 1 is significant at the 1% level, 5 are significant at the 5% level and 8 are significant at the 10% level. The corresponding numbers we would expect by chance alone would be slightly lower at .5, 1, and 2, assuming that the authors did not choose which outcomes to report based on which outcomes were found to be significant. Furthermore, 17 of the 20 outcomes that are reported point in the direction of treatment being beneficial, while only 3 of the reported outcomes are in the direction of the treatment being harmful.

B Appendix Figures

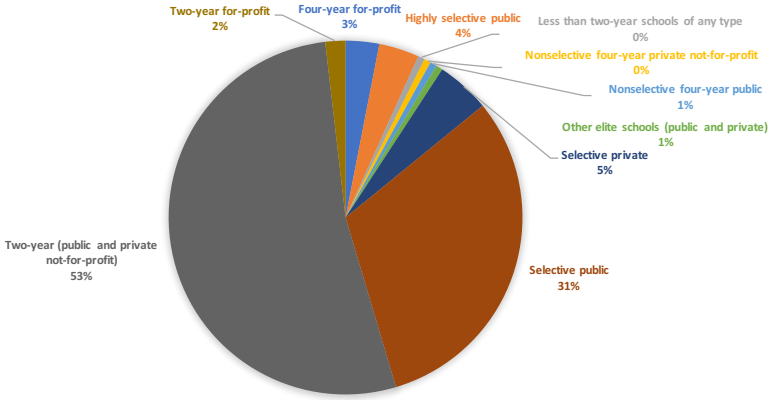
Figure A.1: Baseline Behavior Index



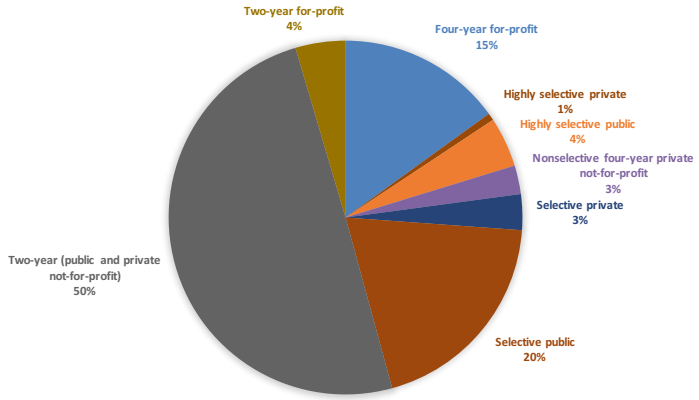
Notes: Year of birth is the year of birth assigned to the taxpayer.

Figure A.2: College Types

(a) Age 20

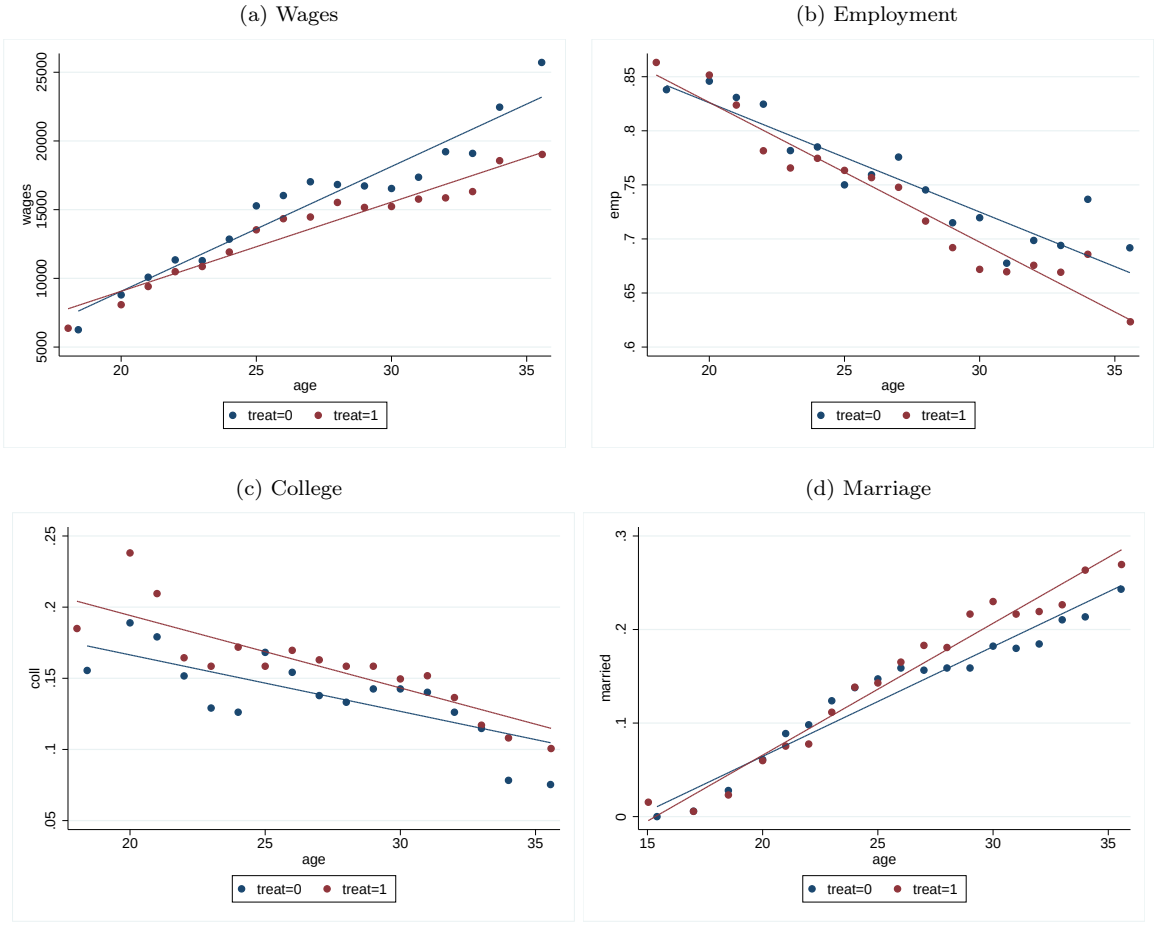


(b) Age 25



Notes: College type is based on IPEDS classification. Sample size is 883 youth linked to tax data.

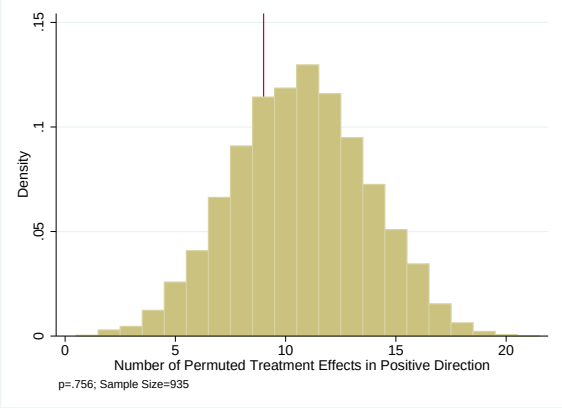
Figure A.3: Dynamic Considerations



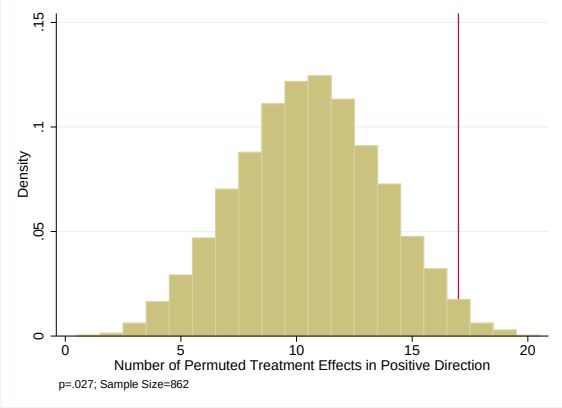
Notes: Sample is 883 youth matched to tax data.

Figure A.4: Additional Permutation Tests for Number of Correctly Signed Outcomes

(a) Behavioral Measures at Baseline

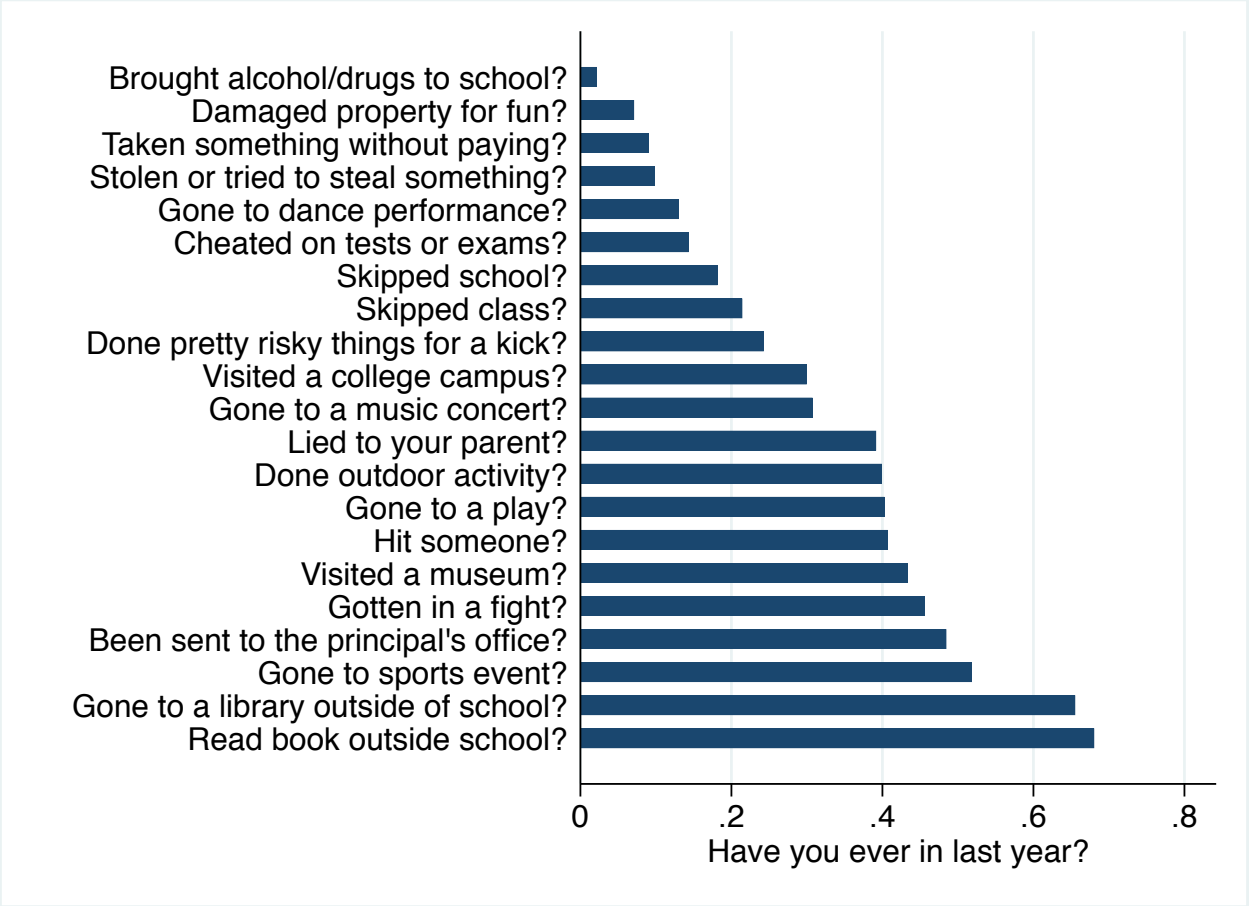


(b) Behavioral Outcomes Among Youth Linked to Taxpayers



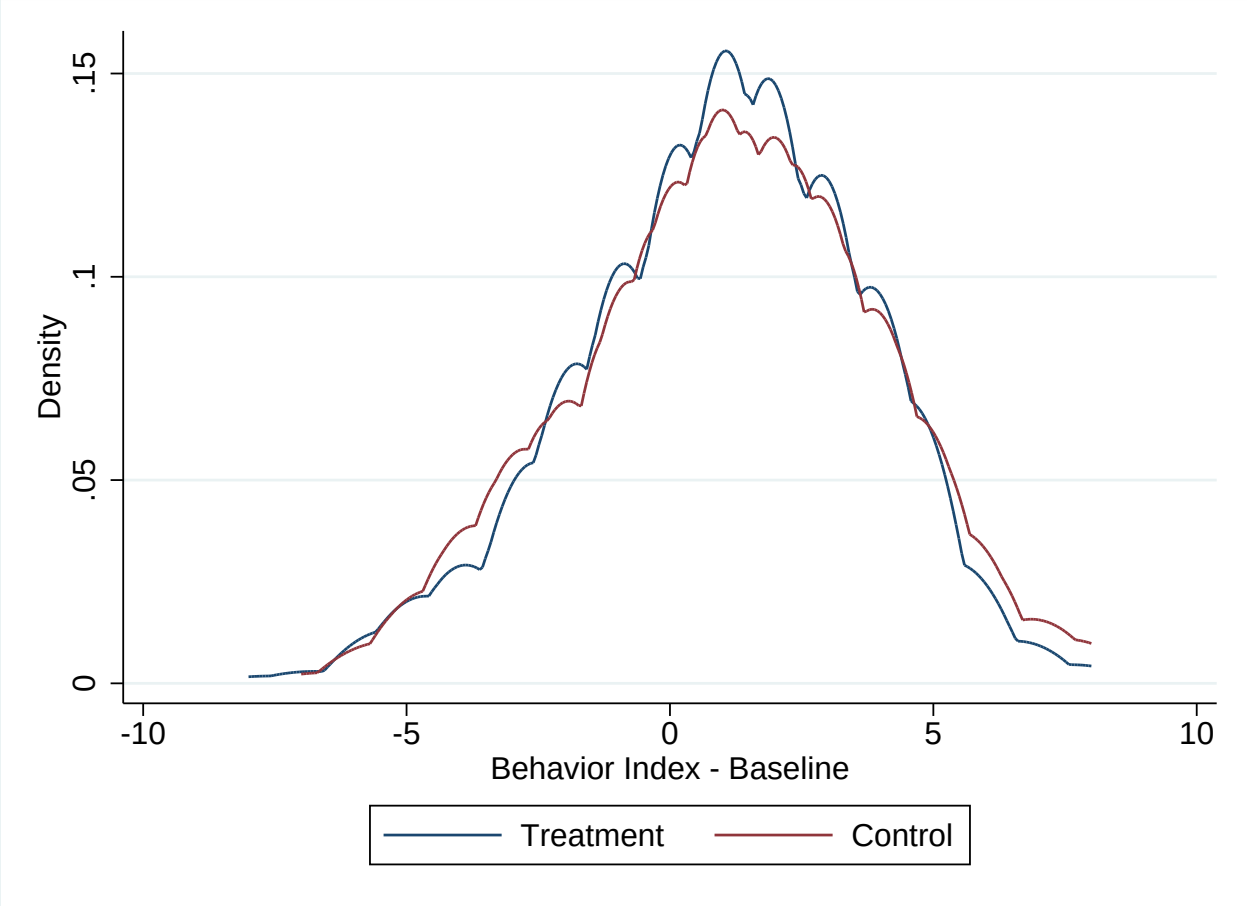
Notes: See notes for Figure 3 pertaining to the simulated permutation test. Panel a shows no significant difference between the reported behaviors of treatment and control subjects at baseline. The sample size is 935 youth with non-missing baseline behavioral reports. Panel B restricts the sample to 862 subjects with non-missing behavioral outcomes and that are matched to tax data.

Figure A.5: Means of 18-Month Behavioral Outcomes3



Notes: Sample size is 914 youth with non-missing behavioral outcomes or baseline reports.

Figure A.6: Baseline Behavior Index



Notes: Sample size is 914 youth with non-missing behavioral outcomes or baseline reports.