The Long-Term Impacts of Mentors: Evidence from Experimental and Administrative Data*

Alex Bell & Neviana Petkova

October 22, 2023
Click here for current draft

Abstract

How do mentors shape kids' identities and later-life outcomes? To evaluate this question, we leverage program administrative records and microdata from a 1991 RCT that randomized disadvantaged children's eligibility for a popular mentoring program. Our re-analysis of the multitude of outcomes collected by the original short-run survey suggests that kids' behaviors improved during the time they were with mentors. A linkage to later-life administrative records shows that treated youth were 10 percentage points more likely to attend college and also showed positive (though less significant) effects on teen birth and marriage. RCT estimates of earnings effects are imprecise. However, using a larger dataset of program administrative records, we develop a supplementary research design comparing matched versus unmatched applicants that replicates key findings from the RCT, and also reveals significant long-term positive earnings gains from program participation on the order of 20%. Through the lens of a model in which adults of differing socioeconomic status influence kids' decision-making, we estimate that mentors may have the potential to mitigate on the order of $\frac{2}{3}$ of the disadvantage that ordinarily hampers low-income childrens' socioeconomic trajectories in adulthood. 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. JEL codes: J62, C12

*Alex Bell, University of California - Los Angeles; contact: alexbell@ucla.edu. Neviana Petkova, Office of Tax Analysis, US Treasury; contact: Neviana.Petkova@treasury.gov. We thank Raj Chetty, Nathaniel Hendren, and Lawrence Katz for advice. This paper also benefitted from the feedback of Alex Olssen as well as participants of the 2023 Spring NBER Children's Program Meeting, 2023 EEA meeting in Barcelona, and the Harvard Labor Lunch. We are particularly grateful to numerous current and former staff at the Boston office of Big Brothers Big Sisters for faciliating the research, in particular: Wendy Foster (whose editorial sparked this research), Nora Leary, Terry McCarron, Chris Masalsky, Mark O'Donnell, and Nicole White. We are also grateful to Big Brothers Big Sisters of America for facilitating research with the RCT data and in particular to David DuBois, Jean Baldwin Grossman, and Carla Herrera for help in locating, interpreting, and using the micro-data. This research has been supported by Harvard's Multidisciplinary Program in Inequality & Social Policy, the EO Foundation, and the Washington Center for Equitable Growth. 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

It is becoming well known that disparities in childhood environments—from homes to neighborhoods to schools—contribute to earnings and achievement gaps in adulthood. One channel may be that the people who surround us during our formative years shape our identities, which affects our choices and in turn economic outcomes. In this paper, we focus on a particular type of relationship: that of a mentor and and a mentee.

Evidence is growing that the types of adults that kids are exposed to shape their career trajectories. As data on interpersonal interactions mount in the digital world, research has begun to identify the role that social factors – namely relationships across social classes — might play in economic mobility.[1] An established 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). Consistent with 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.[2] Other recent work by Chetty et al. (2020) has suggested that mentors play a role in economic mobility due to the strong association between the presence of fathers and economic mobility of young black boys. Kraft et al. (2023) also offers evidence that natural mentors arising in school settings can impact student attainment.[3] But even if mentors matter, can short-term artificially created mentoring relationships confer the same lasting benefits that natural mentors seem to instill? This questions is critical for fostering economic opportunity, as youth from disadvantaged backgrounds do not naturally have equal access to mentors. Recent evidence from a randomized intervention in Germany suggests that assigned mentors may increase kids' prosocial behaviors and schooling decisions (Kosse et al., 2020; Falk et al., 2020), though whether these short-run improvements translate into long-run socioeconomic success remains to be seen.[4]

To better understand social factors in long-run economic mobility, this paper relies on micro-data from a randomized national control trial conducted nearly 30 years ago, as well as a larger dataset of program admin-

[1] See for instance the analysis of Facebook friendship data and economic mobility by Chetty et al. (2022a,b).

[2] Altmejd (2023) finds similar "inheritance" effects for field of study, which seem to be driven more by role modeling than knowledge transfers.

[3] There is a large body of work in psychology that also finds support for natural mentors, e.g. by DuBois and Silverthorn (2005a,b) in the same dataset though with a less rigorous research design (see also Rhodes et al., 1992; Hurd et al., 2016; Reynolds and Parrish, 2018).

[4] In theory, mentors could also do harm to youth by setting unrealistic expectations, increasing awareness of their own hardship, or diminishing trust following an unsuccessful relationship (see for instance Grossman and Rhodes, 2002). (For more discussion of youth interventions that harmed, see McCord and McCord, 1959; McCord, 1978; Dishion et al., 1999).

istrative records. All these records were linked with US administrative tax records to observe participants' long-run outcomes.

This paper contributes the first direct evidence on how childhood mentors can have lasting effects on behaviors and earnings into early adulthood. The results have strong implications for those in both academia and policy circles hoping to find concrete interventions to level the playing field that disadvantaged youth face in America. Methodologically, we demonstrate new frameworks for multiple-hypothesis testing in experiments and for combining multiple treatment effect estimates across datasets. To interpret the results, we also present a stylized model in which kids' behavior is motivated by social exposure to adults of differing socioeconomic status. Because our study isolates a purely social treatment, we argue this is the first direct evidence that identity – which is malleable through exposure to mentors – drives at least some decision-making leading to labor market inequality.

We find that mentors have substantial impact. Our re-analysis of the original self-reported behavioral outcomes collected during the RCT shows that after 18 months, kids who were randomized to be eligible to receive a mentor *said* they behaved better. Furthermore, administrative data corroborates that over the coming years, children in the treatment group were 10 percentage points more likely to attend college, and also show positive (though only weakly significant) impacts on marriage and teen birth. Comparing participants' income into their early 30s, we detect no significant treatment-control difference, but the results are so imprecise that we cannot rule out improvements as large as 18%. Standard estimates of the economic returns to education would be far too small to detect in this small experimental sample. To gain more resolution on the effects of the intervention, we develop an observational research design in a larger dataset of program administrative records on applicants. Using youth who apply but end up not getting matched to a mentor as controls, we replicate the RCT's estimated effect on college and also find a 19% increase in earnings between ages 20 and 25. Using the model as a lens through which to view our empirical results, we find that on the order of two-thirds of the socioeconomic disadvantage in adulthood of being born into a low-income family can be mitigated by a mentor, although some evidence points to this effect being smaller for earnings than for behaviors such as college attendance.

The program we analyze is the community-based mentoring model of Big Brothers Big Sisters. The treatment was a purely social one targeted to at-risk youth, with no financial transfers or tutoring. Relationships began 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. Still, these relationships however "weak" (in the sense of Granovetter (1973)) may confer benefits by virtue of how different the mentor is from the types of role models to which the youth would otherwise be exposed.

The remainder of this paper proceeds as follows. Section 2 provides further background on the mentoring program, the experimental design and potential non-compliance, and the observational research design. Section 3 describes the three datasets we rely on for our analysis: the RCT micro-data, program records from the Boston affiliate, and the US administrative tax records. Section 4 provides new insight into the original behavioral outcomes collected by the RCT with a more rigorous eye toward the multitude of hypotheses tested. Section 5 extends the analysis to long-run outcomes from administrative data, finding the strongest impact on behavioral outcomes, including college attendance. In Section 6, we uncover additional evidence from program administrative records that mentors increased future income. Section 7 presents a model of decision-making in which the socioeconomic status of mentors enters kids' utility functions through a process of identity formation and relates the results to intergenerational mobility. [5] Section 8 concludes with a discussion of costs and benefits.

2 Program Background and Estimation Strategies

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

[5] Some evidence, however, suggests this effect is lower for income than for behaviors such as college attendance.

or with other guardians.[6] The program targets youth with some form of adversity, but on the other hand 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.

[6] The following guidance was given regarding racial classification: "African American, Asian, Pacific Islander, Native American and youth of Hispanic origin are included in the category of minorities. White youth who are not of Hispanic origin are categorized as nonminorities. Biracial youth who express a preference for a particular racial category must be categorized accordingly; if no preference is expressed, consider them minorities."

The match support specialist evaluates the relationship and gives guidance, and can liaise with all parties.

2.1 Experimental Design

2.1.1 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.[7]

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 outcomes 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.

2.1.2 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.[8] 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-

[7]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.

[8]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.

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 assignment. 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.

2.1.3 Treatment & Control Dilution

Because the experiment randomized only eligibility for a mentor, treatment-control comparisons should be interpreted as ITT lower-bounds on the effect of actually being matched to a mentor. Concretely, we estimate the following ITT equation for individual i and a given outcome Y_i:

$$Y_i = \beta Treatment_i + \varepsilon_i$$

where β is the coefficient of interest and the variable $Treatment_i$ is an indicator equal to 1 if the youth was radomized to the treatment group and 0 if the youth was randomized to the control group (regardless of whether the youth was ultimately matched to a mentor).

While we do not have an accurate measure of long-run compliance with treatment and control randomization, several pieces of evidence suggest it was reasonably high.

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. (Of those who had been matched, the average length of the match at time of follow-up was just under one year.) Most of the unmatched youth were boys, due to the low supply of male volunteers relative to over-demand by boys for mentors. Because it is likely that some of the unmatched treatment group would have been matched after the study, treatment compliance is likely higher than 78% when studying long-run outcomes.

Just as some youth from the treatment group did not receive treatment (at least during the first 18 months), there is also potential for members of control to have received some form of treatment. 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. In the long run, however, it is important to note that all of the control group would have been eligible for a

BBBS mentor upon conclusion of the study. Unfortunately, the study did not track how many youth from 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.

2.2 Observational Research Design

To supplement the relatively small RCT data, we also incorporate a second research design in a larger administrative dataset from the Boston affiliate of the program. Whereas the RCT analysis should be interpreted as an ITT estimate (though with a scaling parameter thought to be close to 1), the observational research design seeks to estimate a TOT. Concretely, we estimate the equation

$$Y_i = \beta Matched_i + \gamma X_i + \varepsilon_i$$

where β the coefficient of interest and $Matched_i$ is an indicator for whether the applicant was matched to a mentor and X_i is a vector of youth-specific controls.

In essence, the secondary research design compares youth who applied to the program and were matched with a mentor with those who applied but were never matched to a mentor. In consultation with program staff, the geography of where applicants live was suggested to be among the most important determinants of whether the youth can ultimately be matched to a mentor.[9] To control for geographic selection into match status, we include fine-grained geographic controls, as well as a variety of other control variables collected by the agency at time of intake. However, we cannot control for all information that was available to the agency in creating the match: detailed free-text narratives are constructed about each youth at intake, but these narratives are subsequently destroyed due to their sensitive nature, and thus could not be made available for research.

Ex ante, it is unclear how selection might bias these matched-unmatched comparisons. To the extent that there is some scope for parents to advocate for their children to be matched, this could raise concern over positive selection. However, if caseworkers try harder to match youth in tougher circumstances, the matched-unmatched selection may well be negative. For concreteness, the 1984 service delivery manual for the Big Brothers Association of Greater Boston states: "Because of the large number of boys applying to our

[9]See also the editorial by Foster (2015) following a shooting in Boston describing how a shortage of mentors in specific neighborhoods had years earlier prevented the two teenage shooters and victim from being matched to mentors.

program, we are unable to accept those who are doing well." Program staff look for a balance in which the youth's emotional problems are not yet so severe that they cannot work with a mentor, but yet the youth is sufficiently at risk of adverse outcomes that they could benefit from a mentor. Empirically, Table A1 confirms that matched youth are substantially more likely to live near a subway. Also of note, they are less likely to have a father present, which is consistent with the stated goal of the program of matching youth that seem to be most in need.

3 Data and Descriptives

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 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.[10]

To provide additional context for the experimental group as well as a secondary observational research design, we also link administrative information on approximately 9,000 youth and 30,000 volunteer mentors who applied to Big Brothers Big Sisters of Massachusetts Bay between 1991 and 2010 to the tax records. We then subset the data to only those youth born between 1980 and 1994 so that all participants can be observed in the data until at least age 20. 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

[10] In their experimental evaluation of a German mentoring program, Kosse et al. (2020) incorporated participants' behavior in financially incentivized games to more convincingly conclude without relying solely on self-reports that their mentors improved prosocial behaviors.

Table 1: RCT Baseline Descriptives & Balance

	Baseline Variable	Overall Mean	Treatment Diff.	p
Tax Characteristics	Linked to Administrative Tax Records	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	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 administrative tax records, with the exception of the first row which has a sample of 959. Categories populated by fewer than 3 individuals are not shown.
+ p<0.1 * p<0.05 ** p<.01 *** p<.001

Table 2: RCT Sample Summary Statistics of Long-Run Outcomes

Variable	Sample		
	Women	Men	Full
Wages, age 25-30	$13,973.46	$16,526.25	$15,586.66
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
Behavioral Index	0.32	0.4139785	0.3793885
Behavioral 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 administrative tax records.

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. Selected summary statistics for the Boston sample are presented in Table A1.

3.2 Long-Run Outcomes

We measure most of our long-run outcomes around 2014, when the children from the initial RCT are around 30.[11] Figure A2 plots the year of birth for children in the RCT sample; all would be at least 30 by the latest age of measurement.

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 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 behavioral outcomes, such as college attendance, could also be viewed as inputs to economic outcomes. Means of these outcomes for the RCT sample are presented in Table 2.

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

[11]These are the most recent data we are currently allowed to access for this project, and sensitivity analysis comparing older versus younger cohorts does not suggest any core findings in this paper would be meaningfully different if we observed older ages.

students.[12] These institutions 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 the RCT sample, the modal age in this year would have been 19, but their ages would have ranged from 16 to 24 years old. 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. For the Boston sample, the fact that cohorts are more recent works in our favor here, and we are able to reliably observe college attendance at age 20 as our main outcome; 33% of these youth were in college at age 20. 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 prison between 2011 and 2014.[13] Although this time-period is older than would be ideal (the RCT 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. We view this as a lower-bound estimate of incarceration rates because it includes only federal and state prisons (not local jails) and only in certain years. For the 60 RCT 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 RCT subjects have been married, and of those married, one-half have become divorced. Among the Boston youth, however, marriage rates are substantially lower at only 5%. This is at least partly a function of their younger age, though the different geography and predominantly male skew may also play factors.

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 RCT sample and one-quarter of the men. In Boston, the rate is much lower at about 4%.

Mortality. Death records are obtained from the Social Security Death Master File. Two percent of females and 4% of males in the RCT sample have died. In general, we do not drop these people from the

[12] 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.

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

analysis so as not to bias estimates of treatment effects.[14]

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 in the RCT sample 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.[15] 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.

Because the Boston sample is younger, our primary income measure takes the average over the five most recent years available, which on average corresponds to income between ages 20 and 25.

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% for the RCT sample. For Boston, we define this measurement between ages 20 and 25, and the mean is 36%.

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 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 the RCT 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: economic and behavioral. This method, which requires the researcher to make judgement calls on what the appropriate signs of each outcome are,

[14]For a year in which an individual was deceased and therefore not reporting income, their income would be coded as $0. Other outcomes, such as whether an individual ever attended college, would not necessarily revert to 0 after death.

[15]Recoding $0 to $1 or $1000 does not meaningfully change our results.

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 behavioral 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.[16]

4 Re-Analysis of Short-Run Experimental 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," such as skipping school or stealing, rather than more subjective constructs such as self-esteem.

The ITT estimate without controls for each of these 21 behavioral outcomes at 18 months is shown in Figure 1. (Figure A6, in the Appendix, shows raw means of these outcomes.) 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 due to chance alone (.05 times 21).

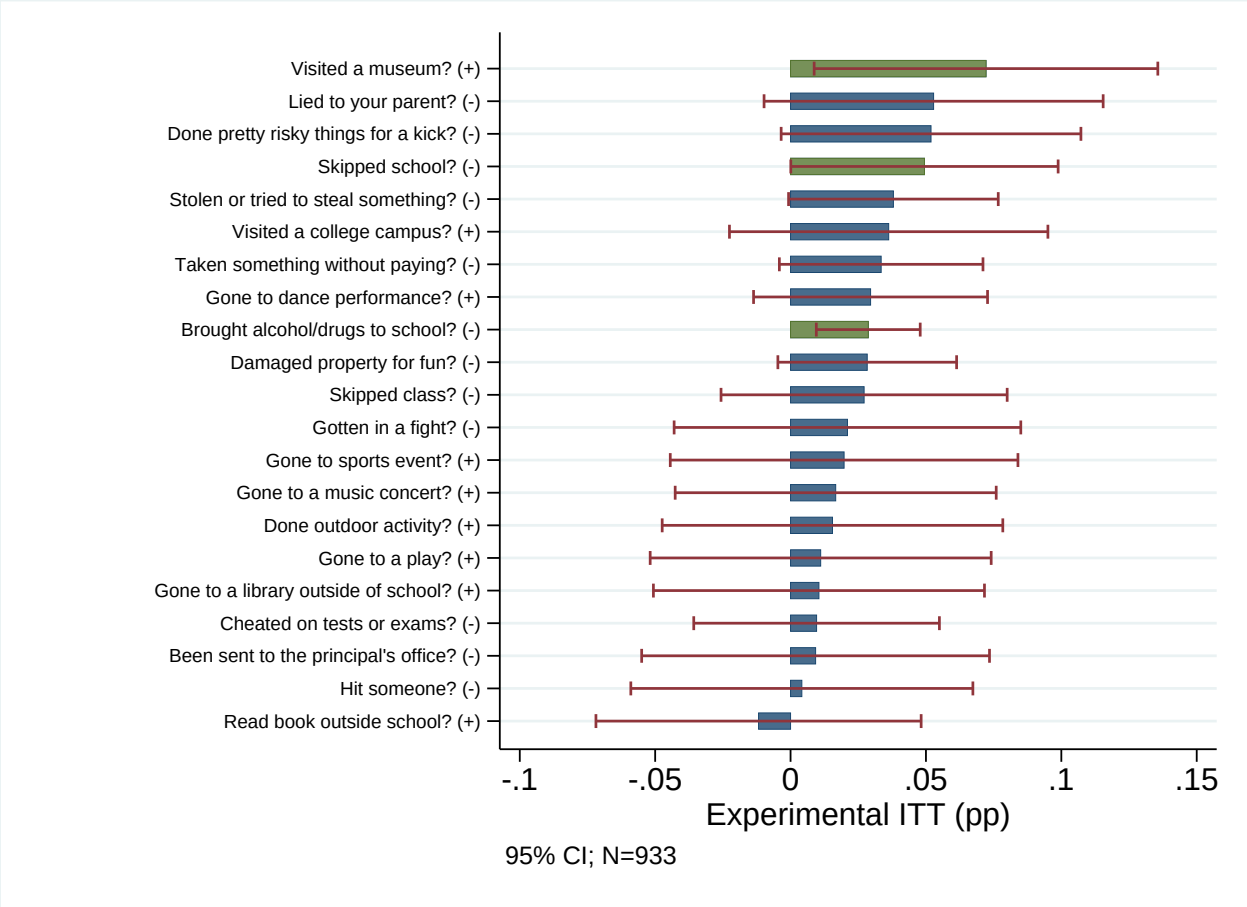
The fact that so few estimates are significant is consistent with the treatment not affecting subjects, but it is also consistent with the effects of the treatment being difficult to measure and noisy in this small sample. To discriminate between these hypotheses, we outline three strategies for multiple-hypothesis 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.

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

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?

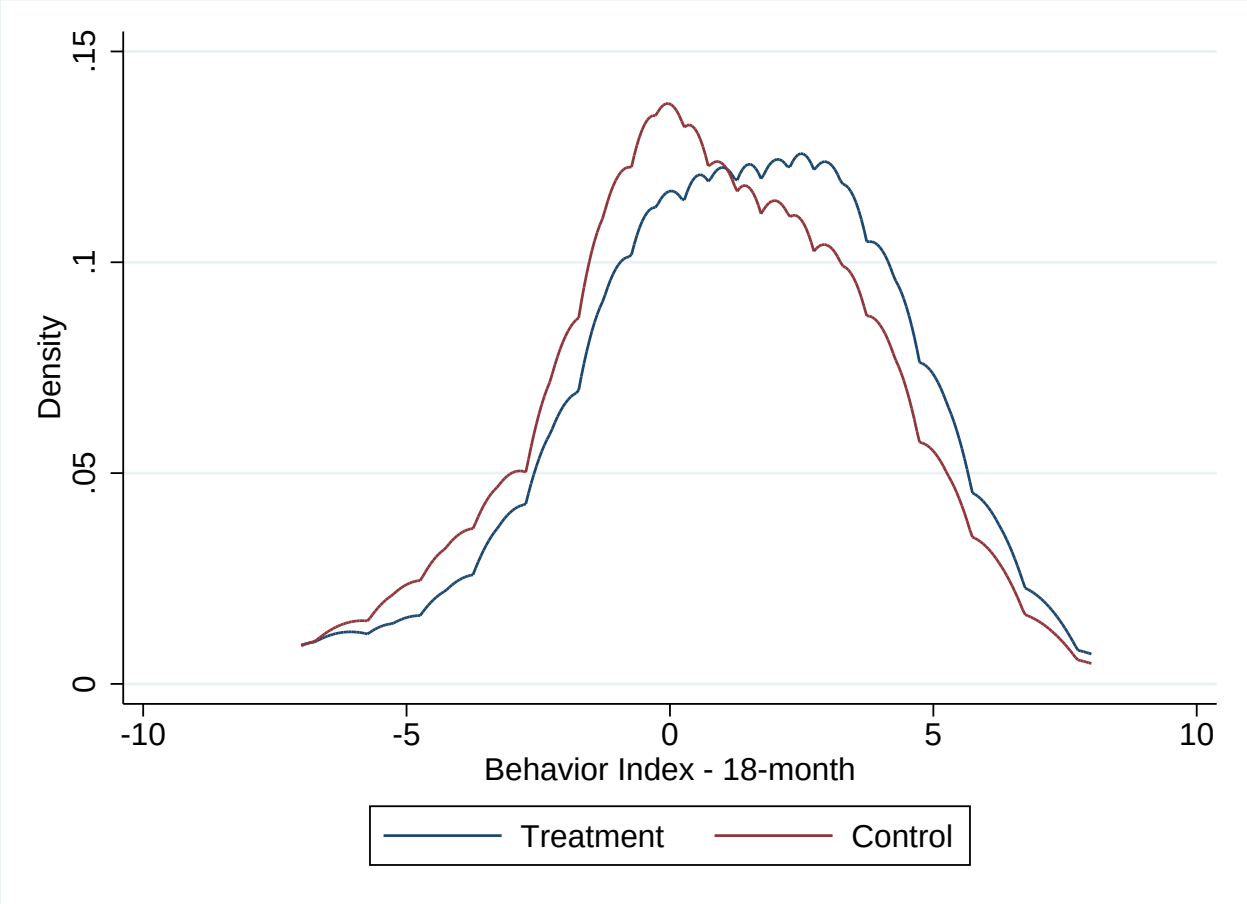


Notes: Coefficients correspond to separate regressions, with no controls. The outcomes have been sorted by the magnitude of the ITT point estimate and all variables have been signed so that the positive direction is the more socially desirable one. Sample size is 933 youth with non-missing behavioral outcomes. 95% CI shown.

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 A7, with no discernable 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

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

	(1)	(2)	(3)	(4)
	Outcome Behaviors Index			Baseline Behaviors Index
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

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 accross the various behavioral outcome measurements. 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 A5 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. 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.[17]

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 behavioral and economic outcomes that can be measured in administrative datasets.

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

5 Long-Run Experimental Intention-to-Treat Analysis

In this section, we leverage our dataset of administrative tax records merged with mentoring data to present the first evidence of the long-run effects of mentors on later-life 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 2.1, 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 numerous 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.

We first focus on the absolute magnitudes of our point estimates. Subsequently, we consider a model of mentors in intergenerational mobility that gives us a framework with which to relate the magnitudes of the outcomes to each other.

5.1 Long-Run Behavioral ITT Estimates

Table 5 shows strong evidence that the RCT-treated group exhibited better long-run behavioral outcomes. The table 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 behavioral outcomes, defined in Section 3.2. Although the magnitude does not lend itself well to an interpretation, there is strong evidence that the treatment has had some effect on this index of behavioral 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 (19.6%).[18] This estimate is individually significant at the $p = .01$ level.[19] The point

[18] Figure A4, in the appendix, shows the robustness of this result to measuring college attendance at only specific ages among applicable cohorts.

[19] Still, we caution against reading too much into the individual significance of this or any individual p-values in the face of

estimate is also very similar to the 9.4 percentage point effect of natural mentors in middle- or high-school on college attendance estimated by Kraft et al. (2023) using twin comparisons and two-way fixed effects in Add Health. This similarity suggests that artificially created mentoring relationships may indeed be as effective as natural ones. The large effects on college attendance also resonate with the large influence of mentors on students' academic tracking in Germany documented by Falk et al. (2020) in a randomized intervention. 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 no more than suggestive evidence of effects on the individual outcome in light of the number of hypotheses tested.

Finally, the remaining multiple-hypothesis tests for behavioral outcomes are presented in Figure 3 and Table 4. The p-value associated with all seven of the behavioral outcomes pointing in the right direction is .013, and the p-value associated with these outcomes predicting treatment is .004.[20]

5.2 Long-Run Economic Self-Sufficiency ITT Estimates

Despite finding large and statistically significant effects on behavioral outcomes, our estimates for economic outcomes are imprecise.

Analogously to Table 5, Table 6 presents ITT estimates for long-run economic outcomes. The overall significance level of the behavioral outcomes is not mirrored in economic outcomes.

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.

the number of hypotheses tested.

[20] 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 outcomes is .029 (Column 4).

Table 4: F-Tests for Balance of Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	18-month behavioral		**L.R. Behavioral**		**Economic**	**All Long-Run**	
21 behavior outcomes	Full sample	Matched to tax records					
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.

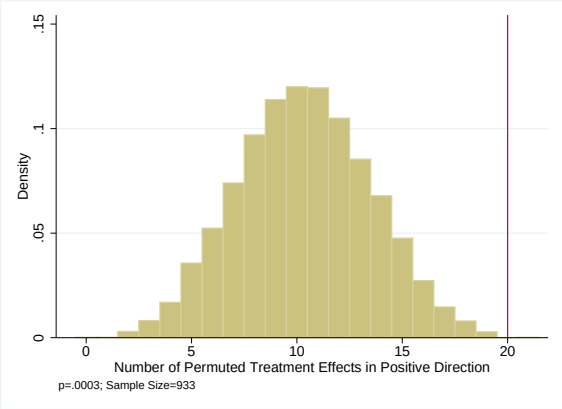
Table 5: Long-Run Behavioral Outcomes

	Behav. Index	College	Married	Divorced	Teen Birth	Deceased	Incarceration	Sentence (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

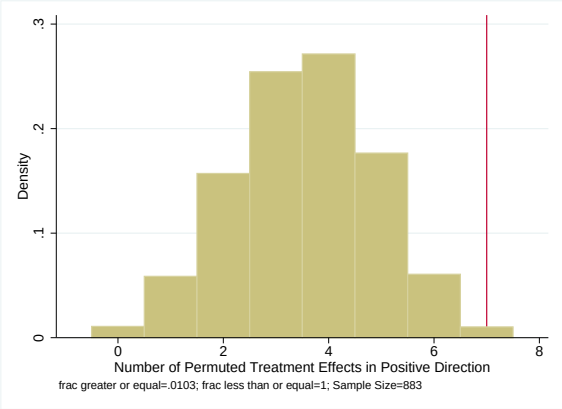
Notes: Standard errors in parentheses. No controls. + p<.10 * p<.05 ** p<.01 *** p<.001

Figure 3: Permutation Tests for Number of Correctly Signed Short-Run Outcomes

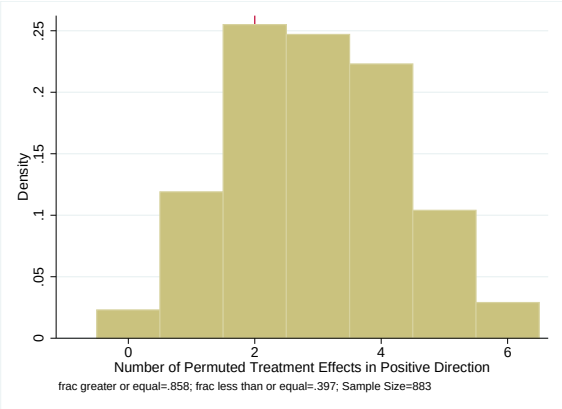
(a) 18-Month Behavioral Outcomes



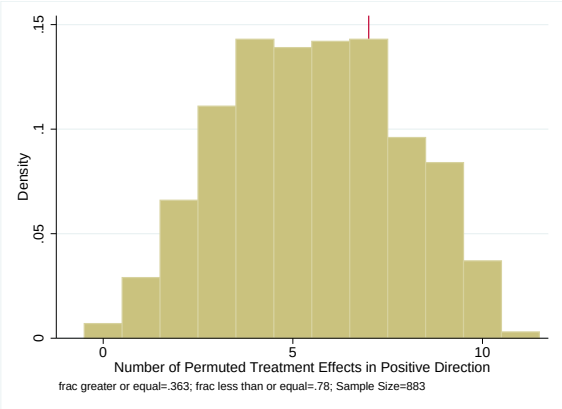
(b) Long-Run Behavioral Outcomes



(c) Economic Outcomes



(d) All Long-Run 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 administrative tax records. Behavioral 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 Economic Outcomes

	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

Moving to other point estimates, wages and log wages both decrease, and non-employment increases.[21] The only point estimates that might be said to go in the favorable direction are the decreases in unemployment insurance and EITC.[22]

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 behavioral outcomes, the economic outcomes may be more highly correlated as wages, log wages, and non-employment 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 A3 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

[21] Figure A4 shows that this insignificant result persists when defining income at different ages.

[22] 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.

threshold in the data.[23] In fact, even if one took the cross-sectional relationship between wages and long-run behavioral 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 behavioral 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.

6 Analysis of Boston Program Records

Given the RCT was not informative as to the effects of the treatment on earnings, we ask whether administrative program records can replicate RCT findings and yield more precise estimates of earnings impacts in the long run. Our strategy compares applicants who are matched with a mentor to those who are not. To address concerns about selection, we first show that we can replicate the results from the original RCT for college-going using this analysis framework. This indicates that selection does not appear to bias our results. Using this design, we also detect large and statistically significant income gains in the long run from being matched to a mentor, on the order of nearly 20%. If one were to view both the RCT and observational analyses as noisy but unbiased estimates of the same treatment effect, the optimal forecast for the earnings increase would be 15%.

Ex ante, it is not obvious that the two frameworks are estimating the same or similar treatment effect parameters. In particular, given the RCT estimates an ITT, there is reason to believe the RCT estimates are under-estimates of the TOT that we seek to estimate in program data. The sample available from Boston is also nearly entirely boys, and they are more recent cohorts than those in the RCT, though it is not obvious in which direction this could push results.

Before moving to a new data source, we first ask whether we can use the matched-versus-unmatched research design to replicate the findings of the RCT within the RCT data. A limitation is that we cannot replicate the full research design with fine-grained geographic controls because the youth selected for the study were intentionally scattered far across the country, leaving little ability to compare outcomes of youth who lived close to each other. Still, Table A4, in the Appendix, shows that matched-unmatched comparisons do not seem out of line with treatment-control comparisons. Among the full sample, the observational research design shows the most significant positive impacts on the behavioral index and college, with an 11 percentage point impact on college closely mirroring the RCT estimate. At the cost of precision, we can also restrict the sample only to the 453 youth assigned to the treatment group, roughly three-quarters of whom were matched during the study. In this sample, we find only a weakly significant ($p < .10$) effect on college attendance, though the point estimate of 9.6 percentage points is still very in line with RCT estimates.

[23] 10% of the sample would have outcomes of .2 log points higher.

Table 7: Matched-Unmatched Comparisons for College Attendance

	(1)	(2)	(3)	(4)	(5)	(6)
Matched	**0.0930*** **	**0.0810*** **	**0.0705*** **	**0.0676** **	**0.0700**	**0.0639**
	(0.0156)	(0.0162)	(0.0185)	(0.0228)	(0.0418)	(0.0489)
Controls:		Parent Income, YOB, Application Year	Parent Income, YOB, Application Year, Tract	Parent Income, YOB, Application Year, Block Group	Parent Income, YOB, Application Year, Block	Parent Income, YOB, Application Year, Block, Agency Controls
N	4067	4067	4067	4067	4067	4067
R^2	0.00812	0.0480	0.215	0.408	0.692	0.745

Notes: The outcome of all regressions is whether the youth attended college at age 20. Standard errors in parentheses. + p<.10 * p<.05 ** p<.01 *** p<.001

6.1 Comparisons Among Boston Matched and Unmatched Youth

Turning to the Boston program data, our first question is whether an observational design in this dataset can replicate the strong positive results for college attendance shown by the RCT. Column 1 of Table 7 shows that, with no controls, matched youth are 9.3 percentage points more likely to attend college than unmatched youth. Adding controls in Column 2 for parent income, year of birth, and application year attenuates the point estimate slightly (the coefficients remain within each other's confidence intervals). Columns 3, 4, and 5 add progressively more stringent geographic controls, from Census Tract to Block Group to Block. Across these three specifications, the coefficient remains close to 7 percentage points; when controlling for Blocks (the most granular level), the estimate is insignificant. While the addition of controls typically not only reduces bias but increases precision, in this case there is a cost to precision by reducing the effective sample size because so few applicants live on exactly the same Block. Column 6 additionally adds in agency controls about the youth, though these also do not increase precision. Table A2, in the Appendix, shows bounds for the point estimates in Columns 2-6 using various seleciton-on-observables correction methods. Using the correction suggested by Oster (2019), the lower bound of any of the specifications is 0.0551. We take the specification in Column 3, which absorbs Tract fixed effects, as our baseline specficiation because it strikes a balance between precision and reduction in bias from geographic selection.

Table 8 shows our main observational analysis for all outcomes.[24] While the economic index lacks significance (though points in the right direction), the behavioral index is positive and signficant. The remaining columns show estimates for individual outcomes (for college attendance, Column 3 replicates Column 3 of Table A2). Columns 4 and 5 show that matched youth earn 19.8% more than unmatched

[24]Table A3 shows selection-on-observables bounds for these estimates as well.

Table 8: Matched-Unmatched Comparisons for All Outcomes

	(1) Economic Index	(2) Behav. Index	(3) College-20	(4) Log Indv. Inc.	(5) Log Fam. Inc.	(6) Prison	(7) Teen Birth	(8) Living Wage	(9) Non-employed 20-25	(10) Youth (14-16) Employment
DV Mean	-.428	.470	.333	8.523	8.523	.0388	.0390	.311	.361	.483
Mentored	**0.0424**	**0.131*** **	**0.0705*** **	**0.198** **	**0.174** **	**-0.0139**	**-0.0129**	**0.0282**	**-0.053** **	**0.0678*** **
	(0.0336)	(0.0267)	(0.0185)	(0.0620)	(0.0607)	(0.00824)	(0.00883)	(0.0175)	(0.0192)	(0.0201)
N	4067	4067	4067	3793	3811	4067	4067	4067	4067	4040
R^2	0.190	0.205	0.215	0.262	0.266	0.165	0.163	0.272	0.211	0.214

Notes: All specifications include controls for parent income, year of birth, application year, and Census Tract indicators. Standard errors in parentheses.
+ p<.10 * p<.05 ** p<.01 *** p<.001

youth, and have 17.4% more total income. These estimates are close to the 18% upper bound of the confidence interval for the earnings ITT effect implied by the RCT.[25] While large in percentage terms, it should be noted that mean income at this age in this sample is low at only about $10,000. Comparisons are insignificant for incarceration, teen birth, and earning a living wage, although matched youth on average were 6.8% more likely to have held a job during their teenage years, and were employed .05 more years between ages 20-25. While we have no estimate from the RCT for youth employment rates due to data availability for their cohort, the estimate for employment during ages 20-25 is within the upper bound ITT from the RCT of .065 years.

Because the Boston sample is particularly young at age of outcome measurement, Figure A8, in the Appendix, shows how selected results differ among applicants who were as young as 20 or as old as 30 by age of measurement. In general, the percentage differences in income appear relatively constant over the life cycle (and the level differences grow). The one outcome that appears to diverge much more substantially at higher ages is the likelihood of earning a living wage. Based on this result Table A5 repeats the research design for this outcome restricted to older cohorts. Among a smaller sample of youth that is more solidly in their post-college years, we are able to detect a moderately significant ($p < .05$) increase in the rate of earning a living wage on the order of 8.9 percentage points, off a base of 49% for this sample.

6.2 Combined Estimators

While the RCT ITT estimate is not perfectly comparable to the Boston-based TOT estimate, a simpified framework would allow us to combine the two estimates. To facilitate a more precise combined estimate, consider a model in which the two research designs are both noisy but unbiased estimates of the same

[25]The fact that the earnings estimate here slightly exceeds the confidence interval from the RCT may be accounted for by differences in sample composition. Relatively to the RCT sample, the Boston data are younger and much more heavily male.

Table 9: Combined Treatment Effect Estimates

	College		
	Experimental	Observational	Combined
Estimate	0.10	0.071	0.078
SE	0.033	0.019	0.016
Variance	0.0011	0.00034	0.00026
Weight	0.24	0.76	
Lower Bound	0.035	0.034	0.046
Upper Bound	0.16	0.11	0.11
	Log Wages		
	Experimental	Observational	Combined
Estimate	-0.053	0.20	0.15
SE	0.12	0.062	0.055
Variance	0.014	0.0038	0.0030
Weight	0.21	0.79	
Lower Bound	-0.29	0.076	0.037
Upper Bound	0.18	0.32	0.25

Notes: Experimental and observational estimates and standard errors correspond to those reported in Tables 5, 6, and 8.

parameter. The combination of estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ that minimizes the mean-squared error from the true parameter β takes the form $a\hat{\beta}_1+(1-a)\hat{\beta}_2$, where it can be shown by taking the derivative of the mean-squared error that $a = \frac{var(\hat{\beta_2})}{var(\hat{\beta_1})+var(\hat{\beta_2})}$. The variance of the resulting combined estimate is given by $\frac{1}{\frac{1}{var(\hat{\beta_1})}+\frac{1}{var(\hat{\beta_2})}}$, which is smaller than the variance of either individual estimate.

Table 9 shows combined estimates for college and log wages. In either case, the optimal combination puts roughly three-quarters of the weight on the more precise observational estimate, and the remaining quarter on the experimental estimate. The combined estimate for the effect of a mentor on college is a .08 percentage point increase (with a confidence interval from .05 to .11). Using log wages, the combined estimator predicts a 15% increase in wages (with bounds of 4% to 25%). For the Boston youth that earn approximately $10,000 per year on average, this amounts to an increase of $1,500 yearly earnings, with a range of $400-$2,500.

While the above procedure assumes the observational analysis is unbiased, alternative procedures exist that would allow for bias in the observational estimates. Following Athey et al. (2020), one could estimate the bias between the estimates using the fairly precise outcome of college attendance, and use that result to adjust the combined estimate for income. Because the observational estimate for college is actually *lower* than the experimental one, such a correction would suggest if anything, the true effect on income could be even larger than what is estimated from the observational comparison.

6.3 Comparisons by Match Length

While youth and mentors are asked to stay matched for at least one year, administrative records from Boston showed many matches lasting longer than this, with a mean length of 2.4 years. Figure A1, in the

Appendix, shows the full distribution of match lengths for both the Boston and the experimental data. While the distributions line up quite closely during the first 18 months (during which experimental matches were tracked), the program administrative data from Boston contain information on a long right tail of match lengths, with approximately 10% of matches remaining intact for six years or longer.

Consistent with the results of treatment-control and matched-unmatched comparisions, we also find variation in outcomes correlated with treatment intensity. Table 10 shows the relationship between several long-run outcomes and match length. Focusing first on only the 2,781 Boston youth that were matched, we find that youth that remained matched longer scored higher on the behavioral index, were more likely to go to college, had higher income, lower rates of teen birth, spent fewer years with no wage income, and were more likely to hold a job as a young teenager. These relationships are more precise if we include unmatched youth as well (whose match lengths were 0). In this specification, we see significant correlations for both the economic and behavioral indices. Each year of being matched with a mentor is associated with a 1.6 percentage point increase in college attendance, 2.8% increase in income (approximately $300 in their early 20s), and a weakly positive increase in earning above the federal poverty line. It is also associated with a 2 percentage point increase in holding a job as a young teenager (ages 14-16), and a 1.3 percentage point decrease in the likelihood of not working for a year between ages 20 and 25. Because the Boston sample is almost entirely male, the rate of observing a teen birth is relatively rare (just under 4%), but we still observe a significant decline with each year of the match of .36 percentage points (a roughly 10% decline per year).

While we would not ex ante expect these correlations with match length to be perfeclty unbiased estimates of the causal effect of each year of mentoring, their signs and magnitudes turn out to be quite reasonable. In each case, the effect per match year turns out to be smaller than the total effect of mentoring, estimated either via the RCT or the observational design. Depending on the outcome, the effect per year is roughly a third to a tenth of the overall effect. Given the mean match length is 2.4 years, most per-year estimates are lower than one would expect if treatment effects were completely linear over time, though imprecision in the exact measurement of match length could also attenuate these estimates.

Whether these correlations reflect the causal effect per year of mentoring, unfortunately, will need to remain the topic of future research. The optimal length of a mentoring relationship is a question of great importance to program administrators, though finding sufficient exogenous variation in match length has proven elusive (Grossman and Rhodes, 2002; Grossman et al., 2012). In the absence of randomized length of match, factors that could confound estimation include how stable the youth's living situation is or the severity of the youth's behavioral problems. Figure A9, in the Appendix, shows the result of several research designs that a priori may hold some promise to isolate the causal effect of match length on college-going in the dataset of Boston matched yourth. We explored strategies including regression with controls, family

Table 10: Comparisons by Match Length

		Matched Youth Only		All Applicants	
	Mean	Effect per Year	(SE)	Effect per Year	(SE)
Economic Index	-0.428	0.010	(0.0072)	0.013*	(0.0061)
Behavioral Index	0.47	0.012*	(0.0053)	0.024***	(0.0046)
College Attendance	0.333	0.0088*	(0.0041)	0.016***	(0.0035)
Individual Income	9,685	235.8*	(98.9)	276.4**	(86.8)
Log Individual Income	8.523	0.014	(0.012)	0.028**	(0.011)
Incarceration	0.039	0.00062	(0.0018)	-0.00045	(0.0015)
Teen Birth	0.039	-0.0026+	(0.0014)	-0.0036**	(0.0013)
Living Wage	0.311	0.0037	(0.0038)	0.0054+	(0.0032)
Non-Employed between 20-25	0.361	-0.0095*	(0.0040)	-0.013***	(0.0034)
Worked During ages 14-16	0.483	0.014***	(0.0042)	0.020***	(0.0036)

Notes: All regressions include controls for year of birth. Dependent variable means correspond to the full sample of 3,984 youth applicants. The number of matched youth is 2,781. (The total sample size is is 4,067 applicants and 2,864 matched, but 83 youth were dropped from the analysis because their match had not yet terminated within the data.)

fixed effects (comparing siblings who were matched for different lengths of time), and exploiting volunteer moves (as indicated by tax filings). While estimates of the effect of match length on college attenandance remain positive throughout the specifications, the results of these designs that leverage more credible sources of variation in match length are not statistically significant.

7 Modeling Mentors in Economic Mobility

To put the magnitudes of our treatment effect estimates into a broader perspective, we develop a model in which parents and mentors both infleunce kids' economic outcomes in adulthood through a process of social identity formation. Our random variation in access to mentors allows us to empirically quantify the importance of this social channel in economic mobility across generations.

Thus far, our discussion has focused on absolute magnitudes. But if the goal of the program is to level the playing field that youth from disadvantaged families face, we can at least view these outcomes on a uniform scale. To relate these magnitudes to each other and to the nature of intergenerational economic mobility, we need a model in which outcomes depend in part on a socioeconomic identity. We modify a simple model of social identity formation based on the framework put forward by Akerlof and Kranton (2000). The identity-in-utility framework provides a useful lens through which to interpret the results, though it is not without its limitations.

Suppose individual j's utility depends on j's own actions a_j, the vector of others' actions a_{-j}, and own identity I_j:

$$U_j = U_j(a_j, a_{-j}, I_j)$$

The identity function itself can be further decomposed:

$$I_j = I_j(a_j, a_{-j}; s_j, \varepsilon_j, P)$$

To translate actions into utils, the identity function itself takes three additional parameters. Key to the identity function in our formulation is the socioeconomic status with which the youth identifies themsevles and others, which we denote s_j.[26] For our application, we view s_j as a choice variable. The ability of a youth to choose their socioeconomic identity is, in a way, also present in the two-audience signaling model and equilibrium characterized by Austen-Smith and Fryer (2005) and a motivating theme of ethnographic work by Wilson (1987, 1996). Identity also incorporates the youth's own characteristics, denoted ε_j, and the commonly accepted prescriptions for different socioeconomic classes, P. In the context of the model, the treatment is a shock to a_{-j}, which can influence utility by changing one's evaluation of their own identity.

Concretely, a young girl may have started with some knowledge of herself (ε_j) that lead her to identify as low-SES (low s_j). She likely takes as given a social prescription (P) that low-SES kids like her don't go to college (for any number of reasons), and in this way her identity would lead her to shun college. So, learning that a mentor the youth is close with went to college (a_{-j}) may surprise the girl. An update may ensue, in which the child re-considers whether low s_j is really the right way for her to see her identity, given her characteristics ε_j and those of the mentor. In this way, the mentor challenges the youth to see herself as someone else – someone she wouldn't have known she could be until she saw it. Through a shift in s_j, the youth sees college as a more promising choice for her as well.

In the context of the literature on interventions for at-risk youth, what stands out about this model is what is missing. In modeling the treatment as impacting the identity component of the utility function, we take off the table competing models of lifting a financial budget constraint or increasing human capital.[27] These mechanisms can be ruled out largely on the basis of the program design. Accounts and existing

[26] While identity is a heterogeneous concept—for instance varying by gender and race—we argue that to interpret our application, one should consider identity as indexed only by social status. By exposing youth to same-gendered mentors, socioeconomic status seems likely to be the primary mechanism of action.

[27] For outcomes such as college attenandance, it may be reasonable to suspect that a pure information mechanism also contributes, such as if a mentor shows the youth's family how to fill out a financial aid application (following Bettinger et al., 2012). Our data do not allow us to identify such a channel specifically; however, we note that the point estimate for college is similar to that for other behaviors, and the pure information channel seems ill-suited at explaining other short- and long-run behavioral outcomes. In addition, surveys of the youth and parents involved in the RCT suggest that information sharing was not a primary activity that matches engaged in (e.g., the topic that youth most commonly reported talking to mentors about was "Fun things you'd like to do together").

evidence do not indicate that the mentors provided substantial financial or tutoring inputs. In surveys from the RCT, matches reported spending the most time on social acitivities such as eating at a restaurant or going to movies or the mall, and doing homework was among the least common activities. Similarly, parents most agreed that the program helped youth the most in terms of increasing opportunities to go places or do new things, or improving self-esteem, social skills, trust, and relationships. Improving school performance or attendance were among the least likely benefits that parents reported.

7.1 Relative Outcome Magnitudes

To operationalize the model of identity formation according to social status, we normalize the scales of all the outcomes according to their relationships with parent income. Intuitively, the resulting answer the quesiton of by how much treated youth have "moved up" the parent income ladder on the given 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.

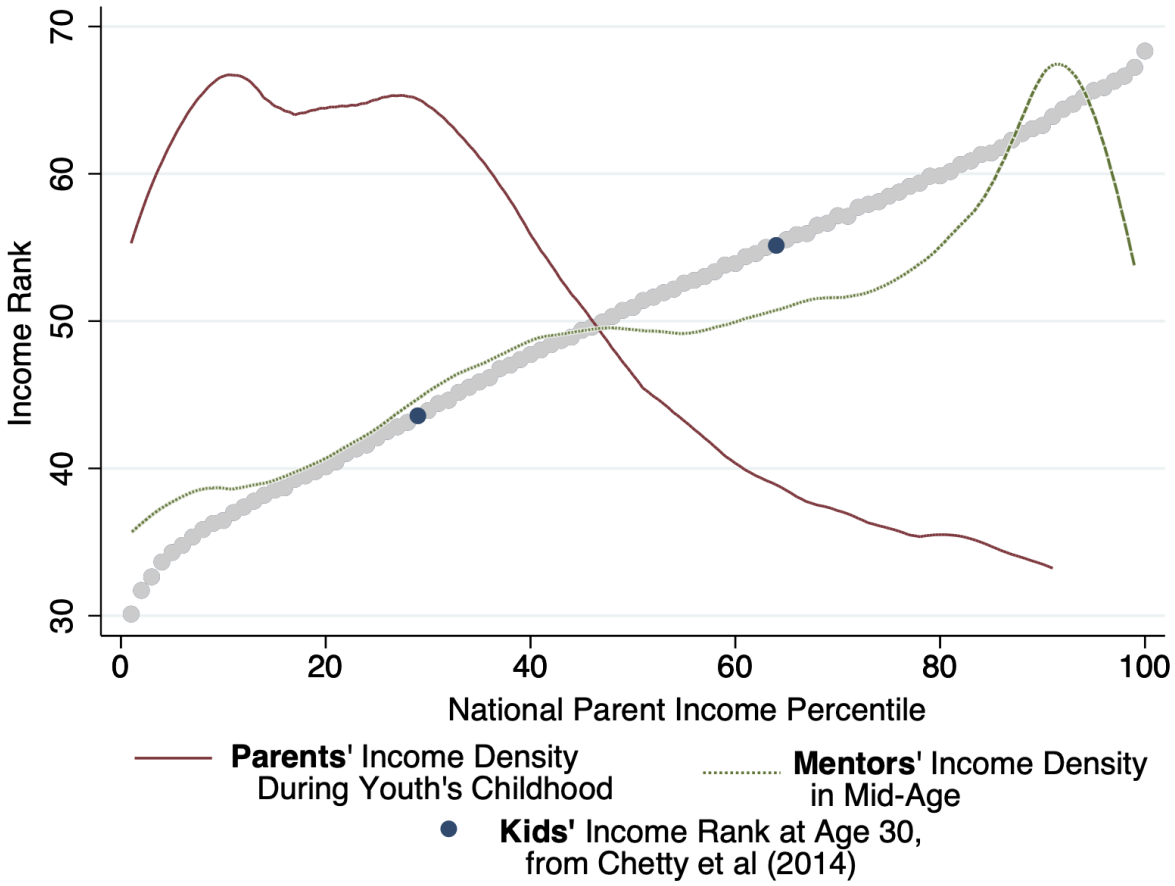
Figure 5 plots the magnitudes of these indexes and their components when all outcomes are re-scaled by their within-sample correlations with parent income (in percentile bins, following Chetty et al. (2014)).[28] Mechanically, a 1 unit increase in any outcome now corresponds to the expected outcome of kids belonging to parents who were 1 percentile higher. On the behavioral 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.[29] 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. While the RCT analysis lends some weak evidence to the hypothesis that the economic effects are smaller than the behavioral effects, that pattern does not emerge from matched versus unmatched comparisons.[30]

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

[29] 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.

[30] 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 both tests, we find moderate evidence that the treatment had a larger effect on the behavioral index than the economic index, when both are rescaled to parent income bins.

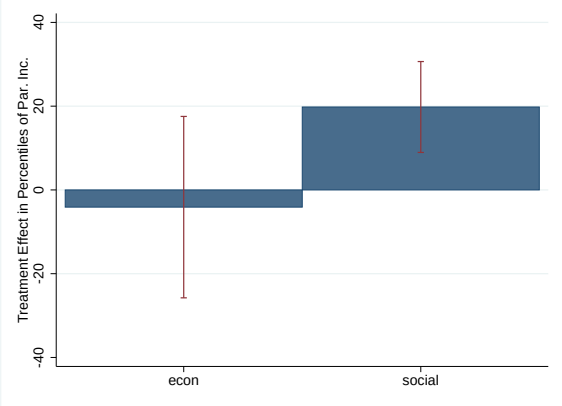
Figure 4: Boston Volunteer & Parent Income Ranks



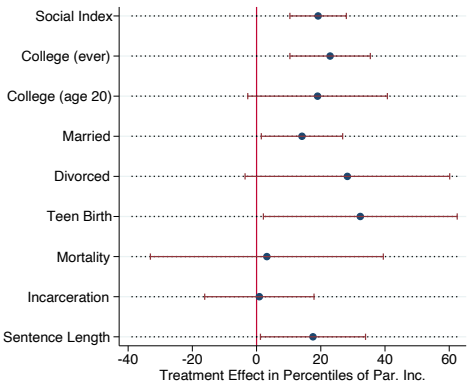
Notes: Sample size is 883 youth linked to administrative tax records.

Figure 5: RCT Treatment Effects, in Parent Income Percentiles

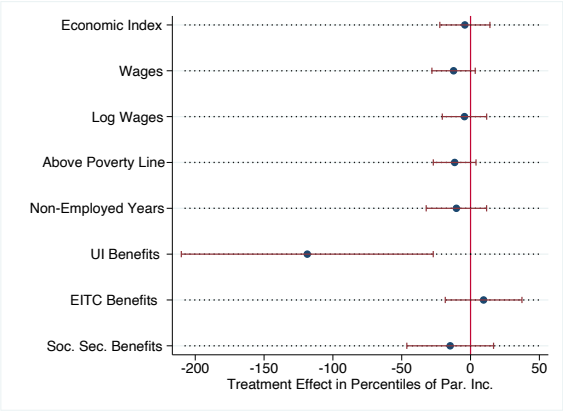
(a) Indexes



(b) All Behavioral Outcomes



(c) All Economic Outcomes



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

Table 11: Matched-Unmatched Comparisons, Rescaled to Parent Income Bins

	(1) Economic Index	(2) Behav. Index	(3) College-20	(4) Log Indv. Inc.	(5) Log Fam. Inc.	(6) Prison	(7) Teen Birth	(8) Living Wage	(9) Non-employed 20-25
Matched	12.02	26.78***	19.15***	25.48**	22.93**	35.31	28.20	15.48	30.91**
	(9.511)	(5.448)	(5.013)	(7.993)	(8.015)	(20.96)	(19.24)	(9.594)	(11.21)
N	4067	4067	4067	3793	3811	4067	4067	4067	4067
R^2	0.190	0.205	0.215	0.262	0.266	0.165	0.163	0.272	0.211

Notes: All specifications include controls for parent income, year of birth, application year, and Census Tract indicators. Standard errors in parentheses.
+ p<.10 * p<.05 ** p<.01 *** p<.001

The analogue to Table 8, Table 11 presents matched versus unmatched comparisons for Boston applicants rescaled to parent income bins. Of note, this analysis suggests the youths' income in early adulthood moved up 23-25 parent bins, of a similar order of magnitude to the improvement in college attendance.

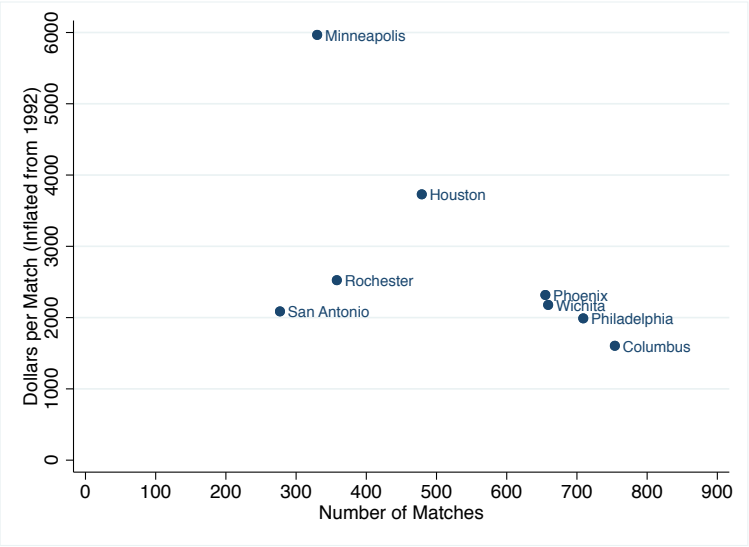
This magnitude of improvement in economic mobility seems consistent with the analysis of economic connectedness by Chetty et al. (2022a). For instance, they find that the strong negative correlation of poverty rates with upward mobility at the zip-code level is reduced by on the order of two-thirds when controlling for economic connectedness (falling from -0.543 to -0.195).[31] Our results thus support their conclusion that social interactions with economically successful individuals are a major factor in geographic differences in intergenerational economic mobility.

8 Discussion and Conclusions

Exposure to successful adults is an important input to success to which not all kids have equal access. We provide the first evidence that mentoring by successful adults may have positive long-run effects on mentees' social and economic outcomes. 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. Although we detect no statistically significant effect of mentoring in the RCT sample, comparisons between matched and unmatched youth in a larger dataset of applicants suggests the program had substantial impacts on income as well. Furthermore, youth who spent more years with their mentor had better long-run outcomes. Putting these effects into a model of social exposure and intergenerational mobility, we find that on many long-run dimensions, kids' outcomes move up roughly two-thirds of the way to that of families as economically successful as the mentors

[31] Using other measures of income and poverty rates at different geographic levels, Chetty et al. (2022a, Table 2) report coefficient reductions ranging from 50% to virtually 100% after controlling for economic connectedness.

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.

were.

The mechanism of these effects was not academic support nor financial investments in the child's education. In fact, mentors were instructed not to engage in these activities. Instead, we model the mechanism as a more holistic improvement of the youth's self-concept and identity. This in turn caused youth to make better decisions, for instance about future schooling.

How do the costs of the program compare to its benefits? 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.[32] As the mentors receive no financial compensation, the primary costs should be thought of as the professional caseworkers who arrange and support these matches.

If one takes as given that the program increases earnings on the order of 20%, it is not hard to imagine the program would pay for itself through increased tax revenue. Assuming a tax rate of 18%, the government would recoup enough taxes to have paid for a year of mentoring in about 7 years. If the 20% boost remains constant as incomes grow throughout life, that cost would be recouped even quicker. Since it seems likely that participants also have a positive willingness to pay for the program, the marginal value of public funds described by Hendren and Sprung-Keyser (2020) for the program would be infinite.

In fact, based only on the RCT estimates on measures of kids' behavioral outcomes, this program seems

[32]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.

reasonably attractive from a cost-benefit analysis. For a close 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. The lower-intensity German mentoring program evaluated by Falk et al. (2020) had a cost estimated to be on a more similar order of magnitude to Big Brothers Big Sisters–1,000 euros—although the outcomes they studied do not map as neatly to the ones we have studied.

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 have access to the type of short-run arrest records used by the Heller et al. (2015) 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.[33]

If structured mentoring programs hold great promise to bring socioeconomic opportunities to disadvantaged youth at reasonable cost, where then is the bottleneck? The key barrier, at present, lies in having enough reliable volunteers to serve as mentors, which leads to lost opportunities for youth who are seeking mentors but instead linger for years on wait lists. To some extent, there may be scope for policy to invest in the recruitment of volunteers and to raise awareness of its benefits. As quantitative research mounts on the long-run benefits of mentors, this evidence may encourage more people to volunteer their time toward such an effort.

[33]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).

References

Akerlof, George A. and Rachel E. Kranton, "Economics and Identity," *The Quarterly Journal of Economics*, 2000, *115* (3), 715–753. Publisher: Oxford University Press.

Altmejd, Adam, "Inheritance of fields of study," Technical Report 2023:11, IFAU - Institute for Evaluation of Labour Market and Education Policy April 2023.

Athey, Susan, Raj Chetty, and Guido Imbens, "Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes," June 2020. arXiv:2006.09676 [econ, stat].

Austen-Smith, David and Roland G. Fryer, "An Economic Analysis of "Acting White"," *The Quarterly Journal of Economics*, 2005, *120* (2), 551–583. Publisher: Oxford University Press.

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.

_ , Matthew O. Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B. Fluegge, Sara Gong, Federico Gonzalez, Armelle Grondin, Matthew Jacob, Drew Johnston, Martin Koenen, Eduardo Laguna-Muggenburg, Florian Mudekereza, Tom Rutter, Nicolaj Thor, Wilbur Townsend, Ruby Zhang, Mike Bailey, Pablo BarberÃ¡, Monica Bhole, and Nils Wernerfelt, "Social capital I: measurement and associations with economic mobility," *Nature*, August 2022, *608* (7921), 108–121. Number: 7921 Publisher: Nature Publishing Group.

_ , **and** _ , “Social capital II: determinants of economic connectedness,” *Nature*, August 2022, *608* (7921), 122–134. Number: 7921 Publisher: Nature Publishing Group.

_ **, 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).

_ , _ **, 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. and Naida Silverthorn, “Characteristics of Natural Mentoring Relationships and Adolescent Adjustment: Evidence from a National Study,” *The Journal of Primary Prevention*, 2005, *26* (2), 69–92. Place: Germany Publisher: Springer.

_ **and** _ , “Natural Mentoring Relationships and Adolescent Health: Evidence From a National Study,” *American Journal of Public Health*, March 2005, *95* (3), 518–524.

_ **, 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.

Falk, Armin, Fabian Kosse, and Pia Pinger, “Mentoring and Schooling Decisions: Causal Evidence,” 2020.

Foster, Wendy, “Haunted By The Lost Chance To Save A Life,” June 2015.

Granovetter, Mark S., "The Strength of Weak Ties," *American Journal of Sociology*, 1973, *78* (6), 1360–1380.

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.

Grossman, Jean B. and Jean E. Rhodes, "The test of time: predictors and effects of duration in youth mentoring relationships," *American Journal of Community Psychology*, April 2002, *30* (2), 199–219.

_ , Christian S. Chan, Sarah E. O. Schwartz, and Jean E. Rhodes, "The test of time in school-based mentoring: the role of relationship duration and re-matching on academic outcomes," *American Journal of Community Psychology*, March 2012, *49* (1-2), 43–54.

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.

Hendren, Nathaniel and Ben Sprung-Keyser, "A Unified Welfare Analysis of Government Policies*," *The Quarterly Journal of Economics*, August 2020, *135* (3), 1209–1318.

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.

Hurd, Noelle M., Joseph S. Tan, and Emily L. Loeb, "Natural Mentoring Relationships and the Adjustment to College among Underrepresented Students," *American Journal of Community Psychology*, June 2016, *57* (3-4), 330–341.

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.

Kosse, Fabian, Thomas Deckers, Pia Pinger, Hannah Schildberg-HÃ¶risch, and Armin Falk, "The Formation of Prosociality: Causal Evidence on the Role of Social Environment," *Journal of Political Economy*, February 2020, *128* (2), 434–467. Publisher: The University of Chicago Press.

Kraft, Matthew A., Alexander J. Bolves, and Noelle M. Hurd, "How Informal Mentoring by Teachers, Counselors, and Coaches Supports Students Long-Run Academic Success," May 2023.

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

_ **and William McCord**, "A Follow-Up Report on the Cambridge-Somerville Youth Study," *The Annals of the American Academy of Political and Social Science*, 1959, *322*, 89–96.

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.

Oster, Emily, "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, April 2019, *37* (2), 187–204.

Reynolds, John R. and Michael Parrish, "Natural Mentors, Social Class, and College Success," *American Journal of Community Psychology*, March 2018, *61* (1-2), 179–190.

Rhodes, Jean E., Lori Ebert, and Karla Fischer, "Natural mentors: An overlooked resource in the social networks of young, African American mothers," *American Journal of Community Psychology*, August 1992, *20* (4), 445–461.

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.

Wilson, William J., *The truly disadvantaged: the inner city, the underclass, and public policy*, paperback ed., [nachdr.] ed., Chicago: Univ. of Chicago Press, 1987.

Wilson, William Julius, "When Work Disappears," *Political Science Quarterly*, 1996, *111* (4), 567–595. Publisher: [Academy of Political Science, Wiley].

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 and Tables

Table A1: Boston Sample Descriptive Statistics and Differences by Match Rates

Variable	Sample Mean	Matched minus Unmatched
Subway within 1 mile	0.51	0.073***
	(0.0079)	(0.017)
Parent Income	30992.6	772.1
	(0.019)	(1303.7)
No father present	0.65	0.070***
	(0.0075)	(0.017)
Male	0.92	0.0069
	(0.0043)	(0.0096)
Age at Application	11.2	-1.29***
	(0.035)	(0.075)
Year of Birth	1990.5	-0.21*
	(0.044)	(0.092)
Racial minority	0.68	0.010
	(0.0073)	(0.016)
Youth in Counseling	0.27	0.0022
	(0.0070)	(0.015)
Match Length (years)		2.34***
		(0.042)

Notes: Sample size is 4,067 applicants. Standard errors in parentheses.
+ p<.10 * p<.05 ** p<.01 *** p<.001

Table A2: Matched-Unmatched Comparisons for College Attendance (with bounds)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Matched	**0.0930*** **	**0.0810*** **	**0.0705*** **	**0.0676** **	**0.0700**	**0.0639**	**0.0654** **
	(0.0156)	(0.0162)	(0.0185)	(0.0228)	(0.0418)	(0.0489)	(0.0203)
Controls:		Parent Income, YOB, Application Year	Parent Income, YOB, Application Year, Tract	Parent Income, YOB, Application Year, Block Group	Parent Income, YOB, Application Year, Block	Parent Income, YOB, Application Year, Block, Agency Controls	Parent Income, YOB, Application Year, Tract, Agency Controls
N	4067	4067	4067	4067	4067	4067	4067
R^2	0.00812	0.0480	0.215	0.408	0.692	0.745	0.303
R_{max}= 1.3*$\tilde{R}$ Bound		0.0767	0.0635	0.0598	0.063	0.0551	0.0569
R_{max}=.75 Bound		-0.1302	0.0123	0.0459	0.068	0.0637	0.0236
R_{max}=1 Bound		-0.2055	-0.0149	0.03	0.0596	0.0538	0.0002

Notes: The outcome of all regressions is whether the youth attended college at age 20. Standard errors in parentheses.
+ p<.10 * p<.05 ** p<.01 *** p<.001

Table A3: Matched-Unmatched Comparisons for All Outcomes (with bounds)

	(1) Economic Index	(2) Behav. Index	(3) College-20	(4) Log Indv. Inc.	(5) Log Fam. Inc.	(6) Prison	(7) Teen Birth	(8) Living Wage	(9) Non-employed 20-25	(10) Teen Employ-ment
Outcome Mean	-.428	.470	.333	8.523	8.523	.0388	.0390	.311	.361	.483
Matched	**0.0424**	**0.131*** **	**0.0705*** **	**0.198** **	**0.174** **	**-0.0139**	**-0.0129**	**0.0282**	**-0.053** **	**0.0678*** **
	(0.0336)	(0.0267)	(0.0185)	(0.0620)	(0.0607)	(0.00824)	(0.00883)	(0.0175)	(0.0192)	(0.0201)
N	4067	4067	4067	3793	3811	4067	4067	4067	4067	4040
R^2	0.190	0.205	0.215	0.262	0.266	0.165	0.163	0.272	0.211	0.214
$R_{max}=$ 1.3*$\tilde{R}$ Bound	0.0367	0.1253	0.0635	0.1946	0.1725	-0.0162	-0.0107	0.0275	-0.0516	0.0571
R_{max}=.75 Bound	-0.0133	0.0804	0.0123	0.1772	0.1648	-0.0413	0.013	0.0243	-0.0413	-0.0215
R_{max}=1 Bound	-0.0382	0.0572	-0.0149	0.1665	0.16	-0.0531	0.0241	0.0223	-0.0359	-0.0632

Notes: All specifications include controls for parent income, year of birth, application year, and Census Tract indicators. Standard errors in parentheses.
+ p<.10 * p<.05 ** p<.01 *** p<.001

Table A4: Implementation of Observational Research Design in RCT Sample

	Econ. Index	Soc. Index	College	Mean Wages	Log Wages	Incarc.	Teen Birth	Marriage
Full Sample	0.032	0.18**	0.11***	-1040.0	-0.027	0.00081	-0.019	0.047
	(0.069)	(0.063)	(0.034)	(1234.4)	(0.12)	(0.018)	(0.031)	(0.034)
Only Treat.	0.14	0.018	0.096+	888.9	0.045	0.0060	0.066	0.0012
	(0.10)	(0.11)	(0.055)	(1898.4)	(0.20)	(0.028)	(0.046)	(0.055)

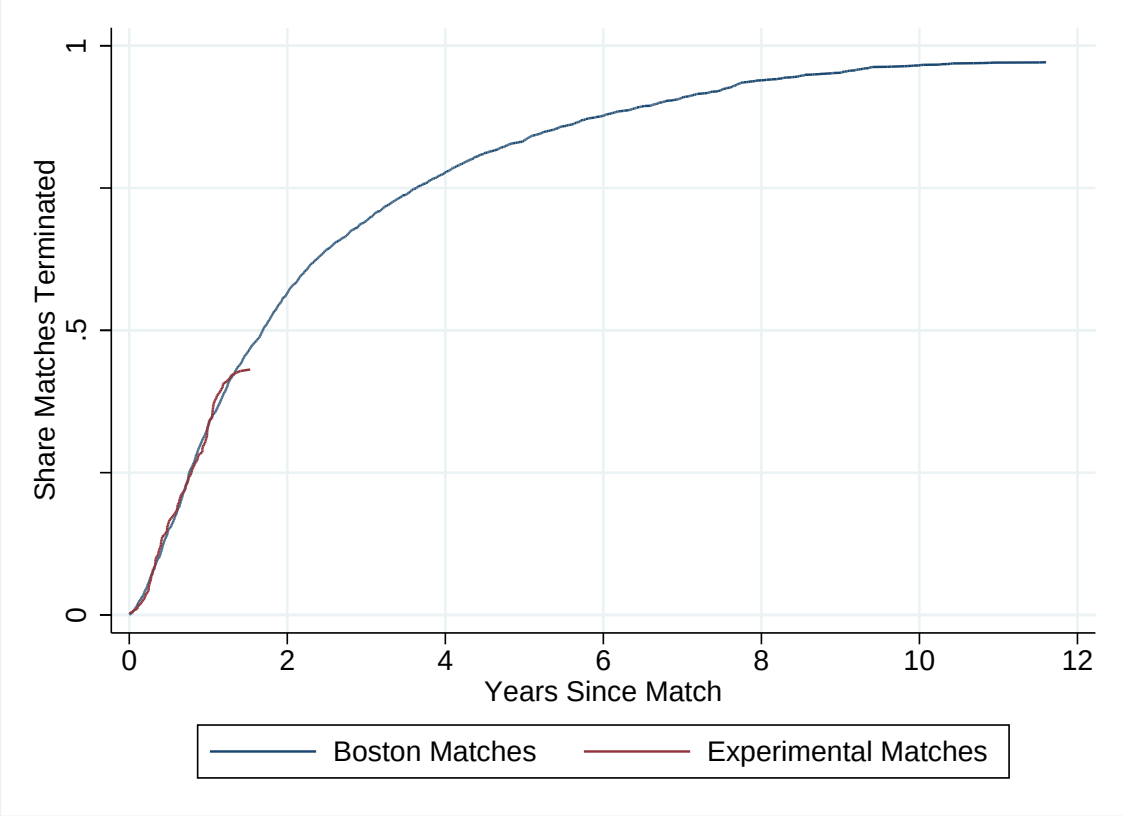
Notes: Each coefficient corresponds to a separate regression of the outcome on an indicator for whether the youth was matched to a mentor during the 18-month study. The first row applies the design to the full sample of 883 youth in the RCT data that were matched to administrative tax records, and the second row restricts the sample to only the 453 of those youth that were randomly assigned to the treatment group (such that all were eligible to be matched).
+ p<.10 * p<.05 ** p<.01 *** p<.001

Table A5: Living Wage Regressions Among Older Boston Youth

	(1)	(2)	(3)	(4)
Matched	0.0304+	0.0282	0.0684*	0.0888*
	(0.0157)	(0.0175)	(0.0308)	(0.0436)
Controls?	N	Y	N	Y
Sample	All ages		Older cohorts	
Outcome mean	.311		.487	
N	4,067		1,318	

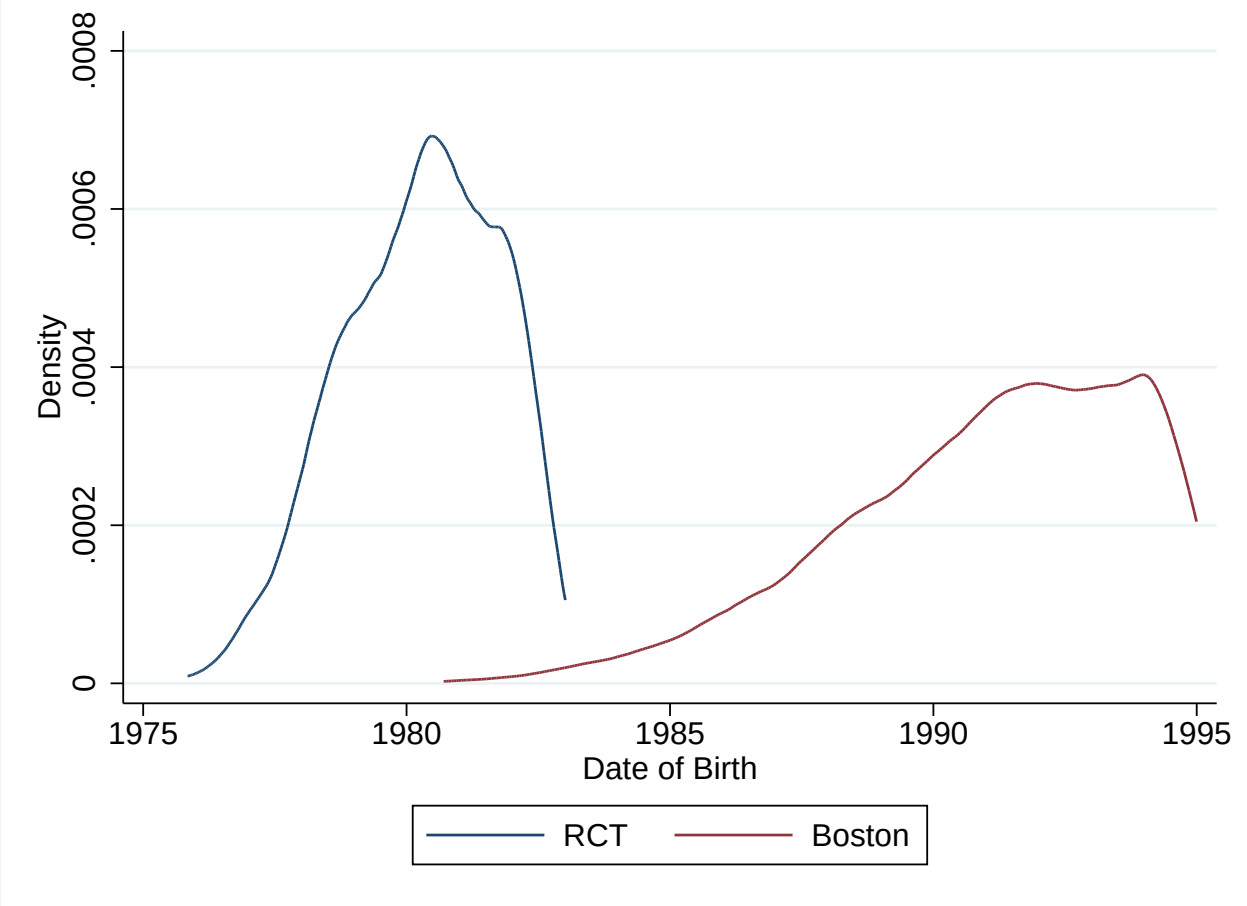
Notes: This table presents matched versus unmatched differences in the share of youth earning a living wage. The outcome is defined as an indicator for whether the average of yearly individual income over 2010-2014 exceeded the federal poverty line of $11,170. The first two columns use the full sample of Boston youth, whereas the second two columns restrict to cohorts born before 1990, who would have been no younger than 21 during the full period of outcome measurement.
youth aged at least 24 in 2014.
+ p<.10 * p<.05 ** p<.01 *** p<.001

Figure A1: Distributions of Match Lengths



Notes: This figure plots the length of match for all matches in the Boston and RCT samples. In both cases, the CDF does not reach 1 because some matches were not closed during the timeframe of data collection. In Boston, a small number of matches were still open at the time the data were extracted; in the RCT sample, most matches were still open, and we are unable to observe closure dates for matches that lasted beyond 18 months. In Boston, the median match length was 1.6 years and the mean was 2.4 years. One quarter of matches lasted less than 9 months, and another quarter lasted longer than 3 years. Approximately 10% of matches reported lasting 6 years or more.

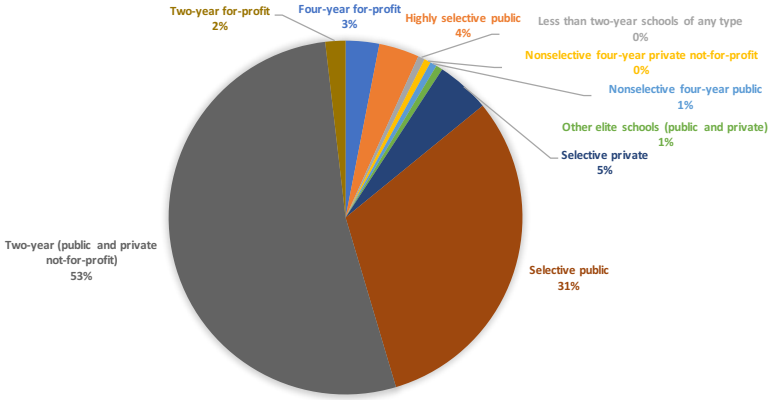
Figure A2: Birth Year of RCT Participants and Boston Applicants



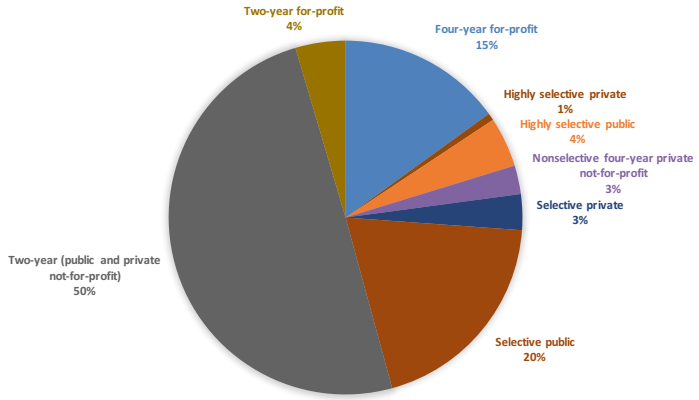
Notes: The RCT sample contains 956 observations and the Boston sample contains 4,067.

Figure A3: College Types

(a) Age 20



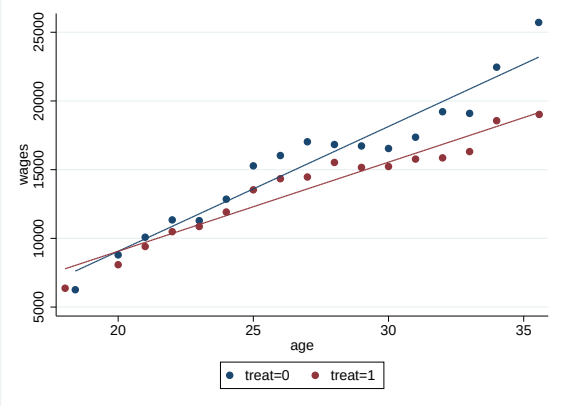
(b) Age 25



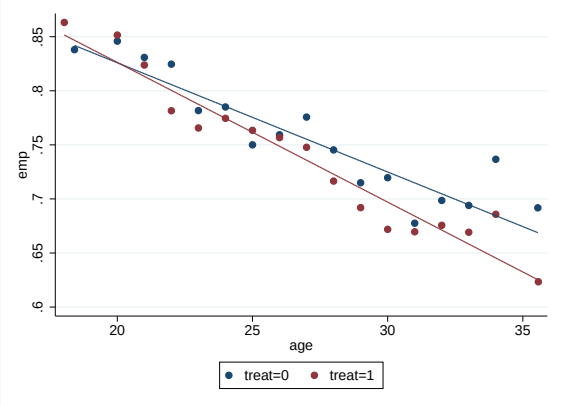
Notes: College type is based on IPEDS classification. Sample size is 883 youth linked to administrative tax records.

Figure A4: Dynamic Considerations, RCT Sample

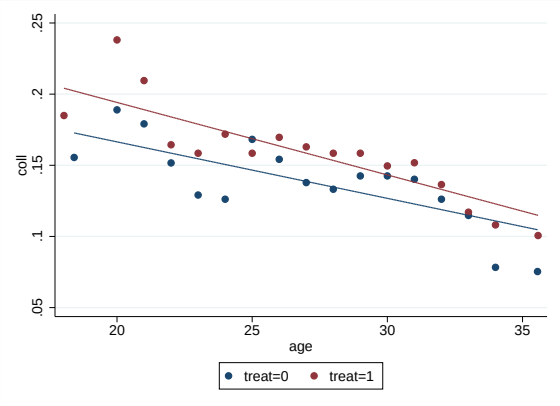
(a) Wages



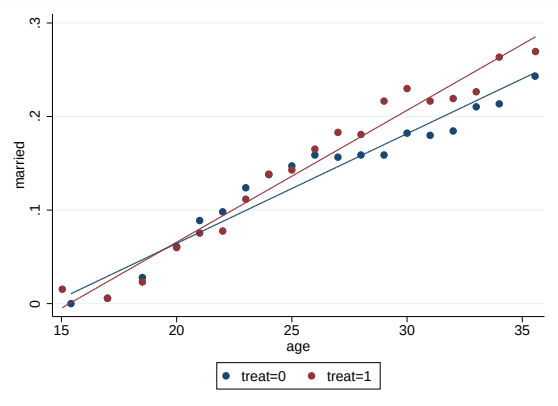
(b) Employment



(c) College



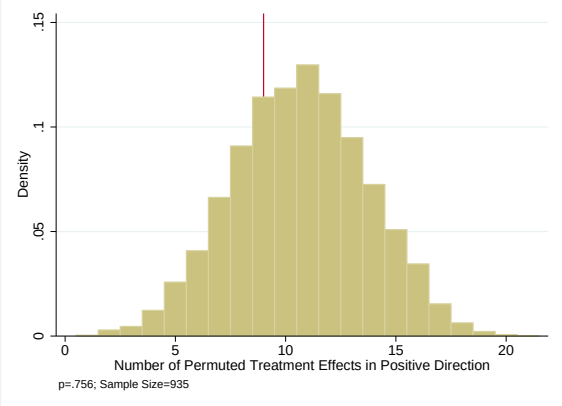
(d) Marriage



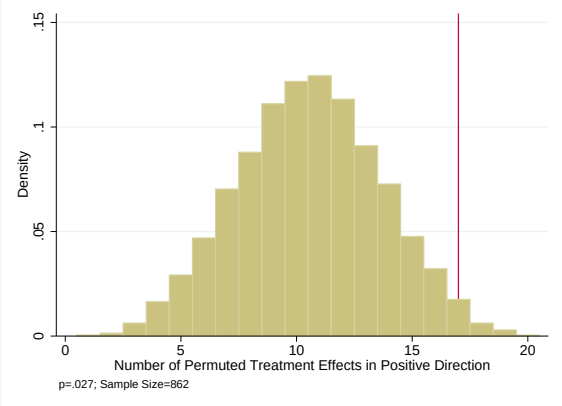
Notes: Sample is 883 youth matched to administrative tax records.

Figure A5: Additional Permutation Tests for Number of Correctly Signed Outcomes

(a) Behavioral Measures at Baseline

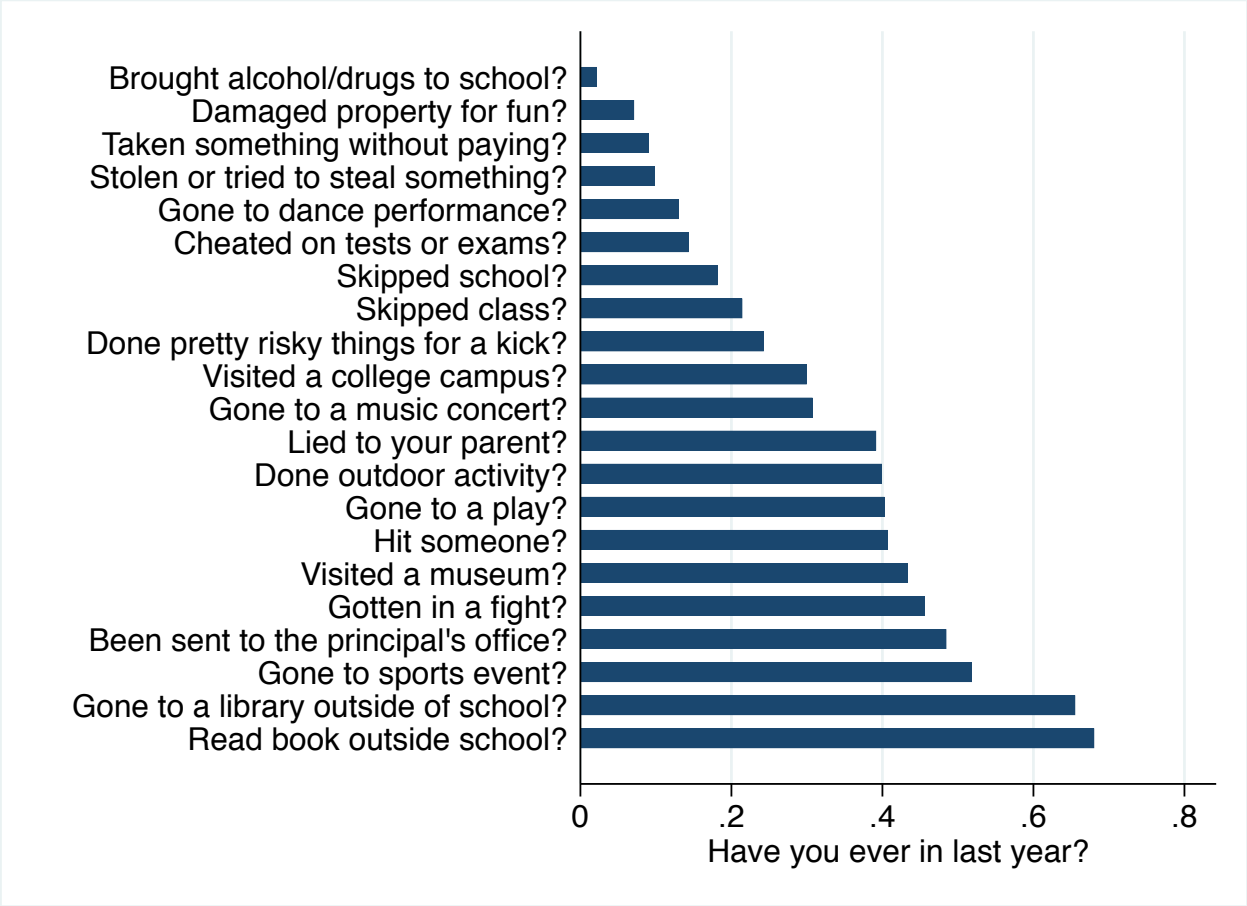


(b) Behavioral Outcomes Among Youth Linked to Taxpayers



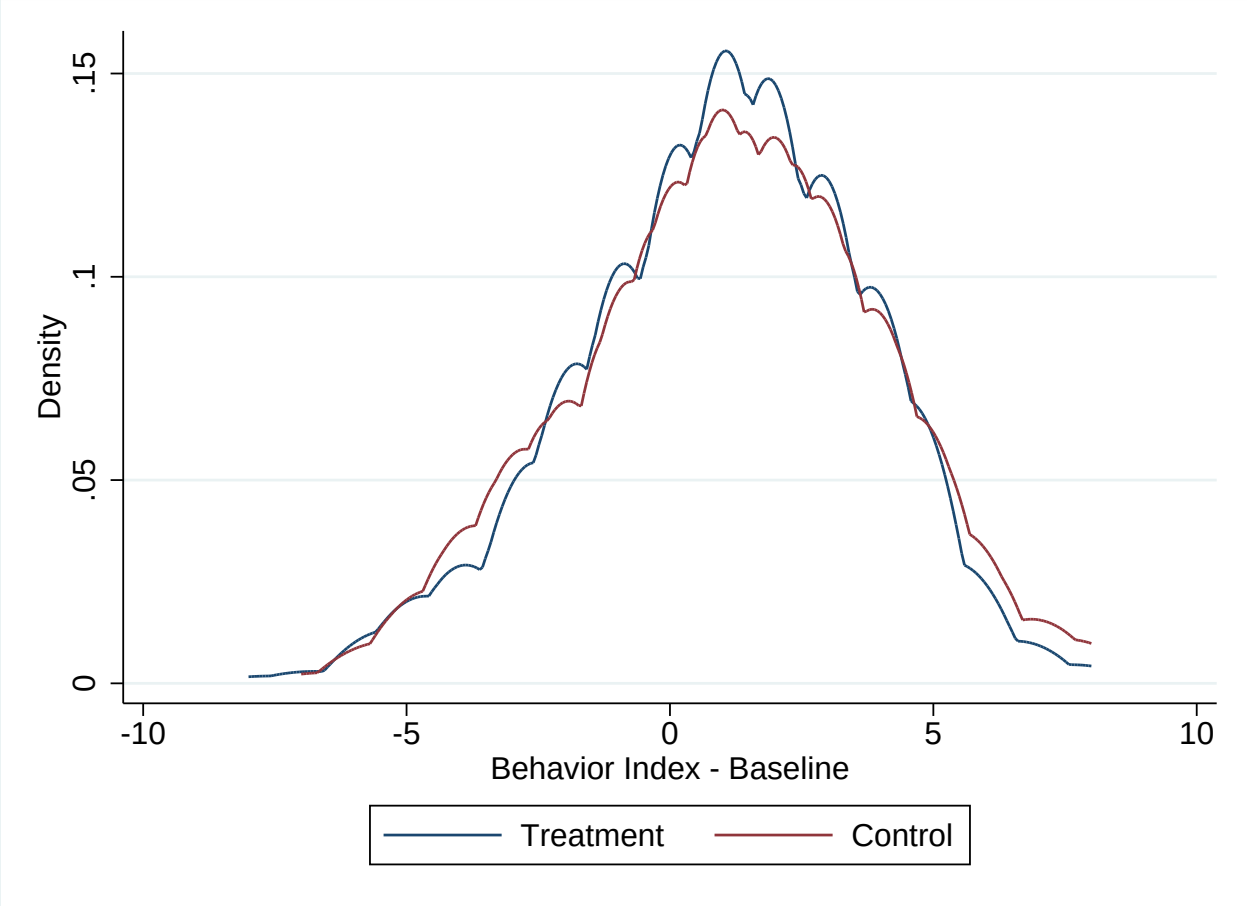
Notes: See notes for Figure 3 pertaining to the simulated permutation test. Panel a shows no significant difference between the reported behaviors of treatment and control subjects at baseline. The sample size is 935 youth with non-missing baseline behavioral reports. Panel B restricts the sample to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

Figure A6: Means of 18-Month Behavioral Outcomes



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

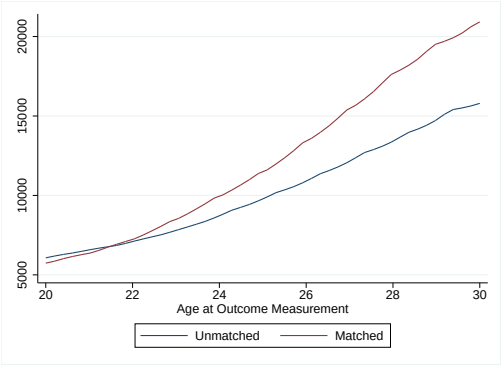
Figure A7: Baseline Behavior Index



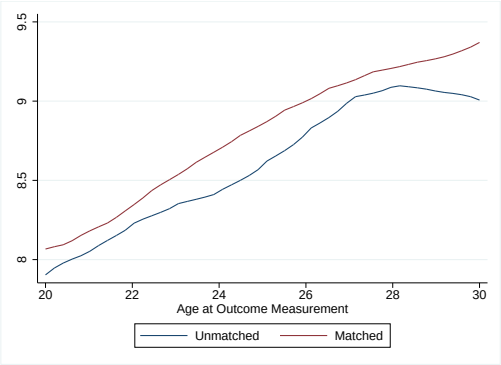
Notes: Sample size is 914 youth with non-missing behavioral outcomes or baseline reports.

Figure A8: Dynamic Considerations, Boston Sample

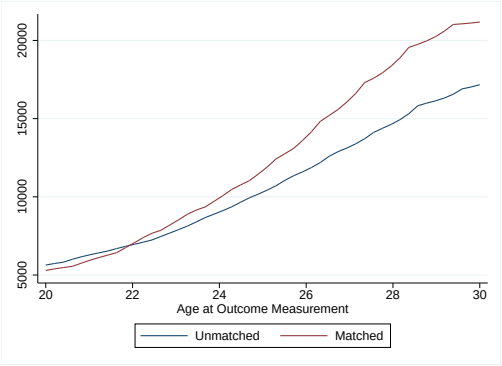
(a) Individual Income



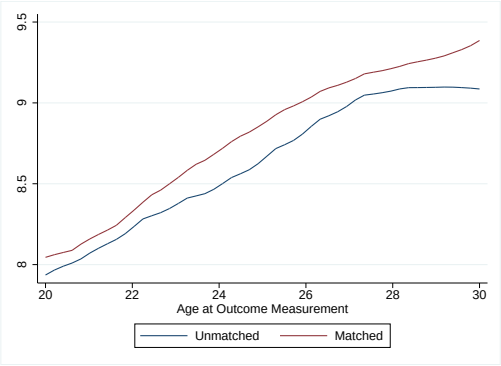
(b) Log Individual Income



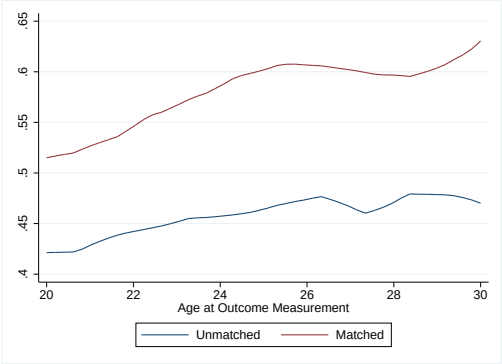
(c) Family Income



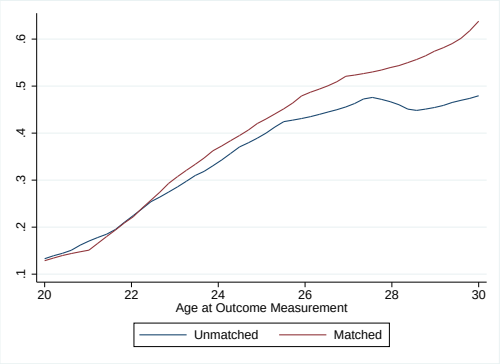
(d) Log Family Income



(e) College (ever)

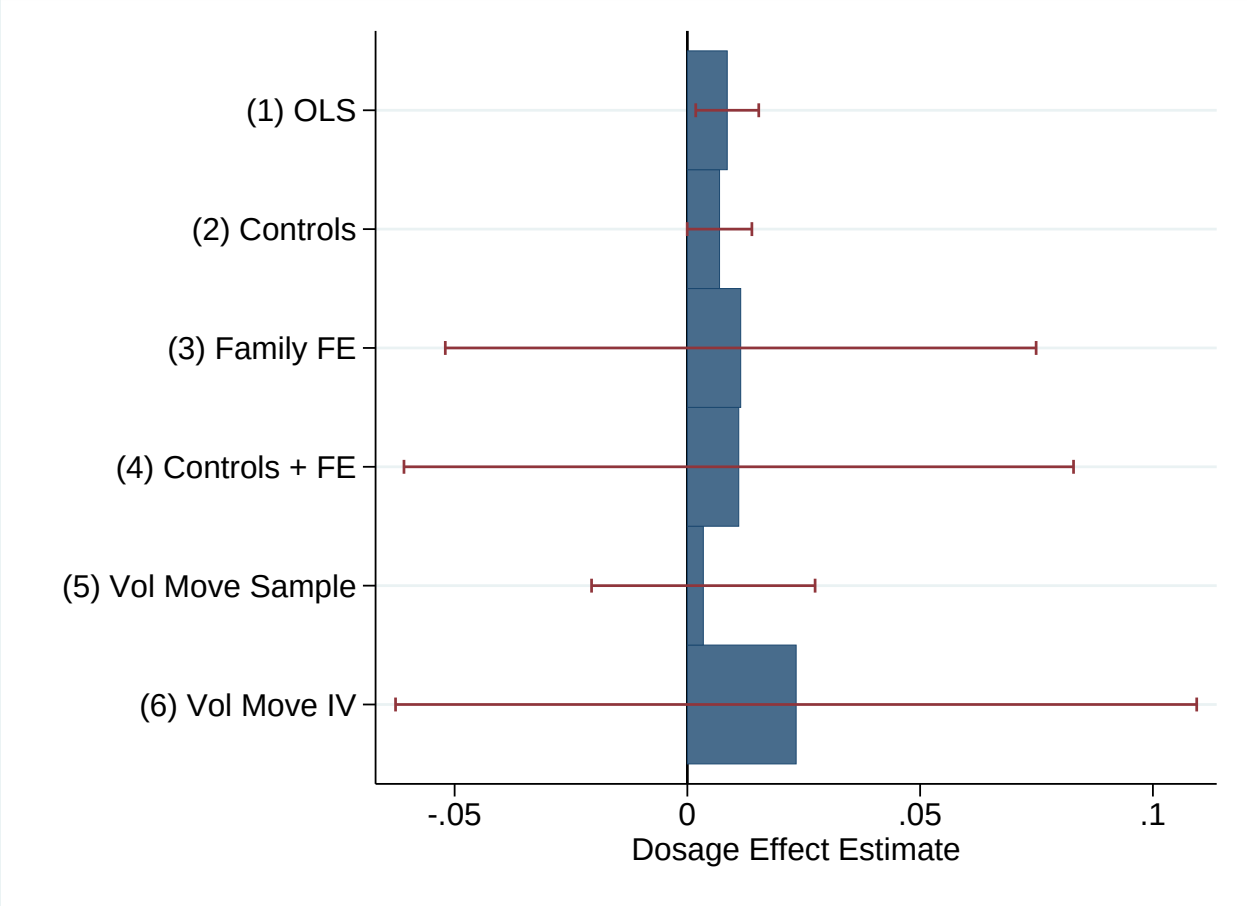


(f) Living Wage



Notes: Each figure plots data for the 4,067 applicants in the Boston dataset using local polynomial smoothing. All outcomes are as of 2014 (or averages over years ending in 2014) and the x-axis corresponds to the age of the applicant in 2014.

Figure A9: Various Research Designs for Estimating Per-Year Effects of Mentors



Notes: This figure depicts the results of several research designs for measuring the effect per year of mentoring on college attendance in the Boston data. Column 1 regresses the outcome on length of match. Column 2 includes various controls for parent income and youth pre-existing behavioral problems. Column 3 is Column 1 plus parent fixed effects, whereas Column 4 includes both the controls and parent fixed effects. Column 5 restricts Column 1 to the subset of matches that reported to BBBS closing due to a volunteer's move. Column 6 instruments for match length with the length of time until we first observe the volunteer move in the administrative tax records.