The Perry Preschoolers at Late Midlife:

A Study in Design-Specific Inference*

James J. Heckman
University of Chicago

Ganesh Karapakula
University of Chicago

*James J. Heckman: Center for the Economics of Human Development, 1126 East 59th Street, Chicago, IL 60637; phone: 773-702-0634; fax: 773-702-8490; email: jjh@uchicago.edu. Ganesh Karapakula: Center for the Economics of Human Development, University of Chicago; email: vgk@uchicago.edu. We thank Kurtis Gilliat, John Eric Humphries, Meera Mody, Sidharth Moktan, Tanya Rajan, Azeem Shaikh, Joshua Shea, Winnie van Dijk, and Jin Zhou for providing helpful comments. We also thank Jorge Luis Garcia, Sylvi Kuperman, Juan Pantano, and Anna Ziff for help on related work. We thank Alison Baulos and Lynne Pettler-Heckman for their help in designing the sample survey. We thank Mary Delcamp, Iheoma Iruka, Cheryl Polk, and Lawrence Schweinhart of the HighScope Educational Research Foundation for their assistance in data acquisition, sharing historical documentation, and their longstanding partnership with the Center for the Economics of Human Development. We thank NORC at the University of Chicago for collecting the new data used in this paper. We thank Louise Derman-Sparks and Evelyn K. Moore for discussing and sharing documentation about how the intervention was delivered. This research was supported in part by: the Buffett Early Childhood Fund; NIH Grants R01AG042390, R01AG05334301, and R37HD065072; and the American Bar Foundation. The views expressed in this paper are solely those of the authors and do not necessarily represent those of the funders or the official views of the National Institutes of Health. The web appendix can be found at: `http://cehd.uchicago.edu/perry-design-specific-inference`.

Abstract

This paper presents the first analysis of the life course outcomes through late midlife (around age 55) for the participants of the iconic Perry Preschool Project, an experimental high-quality preschool program for disadvantaged African-American children in the 1960s. We discuss the design of the experiment, compromises in and adjustments to the randomization protocol, and the extent of knowledge about departures from the initial random assignment. We account for these factors in developing conservative small-sample hypothesis tests that use approximate worst-case (least favorable) randomization null distributions. We examine how our new methods compare with standard inferential methods, which ignore essential features of the experimental setup. Widely used procedures produce misleading inferences about treatment effects. Our design-specific inferential approach can be applied to analyze a variety of compromised social and economic experiments, including those using re-randomization designs. Despite the conservative nature of our statistical tests, we find long-term treatment effects on crime, employment, health, cognitive and non-cognitive skills, and other outcomes of the Perry participants. Treatment effects are especially strong for males. Improvements in childhood home environments and parental attachment appear to be an important source of the long-term benefits of the program.

Keywords: randomized controlled trial; early childhood interventions; life cycle treatment effects; randomization tests; re-randomization; worst-case inference; least favorable null distributions; partial identification; small-sample hypothesis testing

JEL codes: C1; C4; I21

James J. Heckman
Center for the Economics
of Human Development
University of Chicago
1126 East 59th Street
Chicago, IL 60637
Email: jjh@uchicago.edu

Ganesh Karapakula
Center for the Economics
of Human Development
University of Chicago
1126 East 59th Street
Chicago, IL 60637
Email: vgk@uchicago.edu

1 Introduction

The Perry Preschool Project was an experimental high-quality preschool program targeted toward disadvantaged African-American children in the 1960s. Previous studies through age forty report large treatment effects for numerous outcomes (see, e.g., Heckman et al., 2010a,b). These studies have greatly influenced discussions about the benefits of early childhood programs (Obama, 2013).

This paper analyzes participant outcomes through their mid-fifties using a new survey of the original participants. We examine previously analyzed outcome measures at later ages, such as self-reported employment outcomes and measures of criminal activity sourced from administrative records, and a host of new measures collected on cognition, personality, and biomarkers of health measured through epidemiological exams. We develop and apply new statistical methods for analyzing the data.

Critics of the Perry program question the validity of the conclusions drawn from the Perry data. They point to the small sample size of the experiment—just over a hundred observations. They also mention incomplete knowledge of and compromises in the randomization protocol used to form the control and treatment groups. Problems with attrition and non-response are also cited.

This paper presents evidence of the efficacy of the Perry program that survives application of a statistically conservative inferential procedure that explicitly accounts for the limitations of the Perry data. We build a formal model of the randomization protocol that captures our imperfect knowledge of it. We partially identify the set of randomization protocols that are consistent with the available information about the randomization procedure. We construct worst-case randomization tests over this set using approximations of least favorable randomization null distributions.[1] These tests are conditional on the observed data and are theoretically conservative if our model of the randomization protocol is valid. We conduct Monte Carlo experiments to compare the false

[1] Our approach builds on that of Heckman et al. (2011) with respect to accounting for information about the true randomization protocol. However, we use more accurate baseline data and documentation to build our model of the randomization protocol. Our approach is more broadly applicable to other compromised social science experiments, including those using re-randomization designs. See Bruhn and McKenzie (2009), Morgan and Rubin (2012), Li et al. (2018), and Banerjee et al. (2017, 2016).

rejection rate of our procedure in practice with that of the standard inferential methods, such as asymptotic, bootstrap, and permutation tests, used in previous studies that ignore essential details of the experimental setup. In these exercises, the standard methods lead to more false positive results than desirable, justifying some of the doubts of the critics, whereas our worst-case randomization tests perform much better with much lower rejection rates, often below the nominal levels. Despite the conservative nature of our new procedures, we find many statistically significant treatment effects on economically important outcomes that survive our worst-case analyses.

We report significant effects on criminal activity, especially violent crime, of the participants. Reduced criminal activity is associated with higher lifetime employment and earnings, especially for males in their late twenties and thirties. We also find treatment effects on late-midlife health outcomes, executive functioning, and lifetime socioemotional skills. We also find that the participants had better childhood home environments and parental attachment during childhood, which are potential sources of the observed long-lasting treatment effects. A companion paper (Heckman and Karapakula, 2019) documents that the treated participants lead more stable married lives and provide their children better home environments compared to the untreated. The children of the treated participants fare better in various life domains compared to the children of the untreated. The companion paper also finds beneficial effects on the siblings of the original participants.

The structure of this paper is as follows. Section 2 presents and discusses the descriptions of the randomization procedure that appear in the published literature. There is some ambiguity in these descriptions, including how candidate treatment and control groups were formed and how reassignments of treatment status were made after the initial randomization. Section 3 presents a framework to formally characterize randomization rules consistent with the available descriptions. We use this framework to partially identify the class of randomization protocols compatible with the observed treatment and control groups. This analysis forms the basis for the conservative tests reported in this paper. Section 4 describes and motivates our estimators of program treatment effects. Section 5 discusses the conventional tests used in the literature and construction of our worst-case inferential procedures. Section 6 reports our empirical analyses. Section 7 concludes.

2 Perry Experimental Design and Background

The Perry Preschool Project was conducted in five waves between fall 1962 and fall 1965 near the Perry Elementary School, a public school in Ypsilanti, Michigan, a small city near Detroit. The intitial sample size was small: 128 children were allocated over five entry cohorts. Five of these children were dropped from the study due to extraneous reasons.[2] Data were collected at age 3, the entry age, and through annual surveys collected until age 15, with additional follow-ups conducted around ages 19, 27, 40, and 55. Program attrition remained low through age 55, with about 83% of the 123 original subjects interviewed. About 12% of the participants were deceased at the latest follow-up, and the rest of the attrited could not be interviewed.[3] Our estimates of treatment effects account for this attrition. Numerous measures were collected over this span, including economic, crime, and educational outcomes. Program intensity was low compared to that in many subsequent early childhood development programs.[4] Beginning at age 3, treatment in the following two years consisted of a 2.5-hour per day preschool on weekdays during the school year, supplemented by weekly home visits by teachers.[5] The Perry curriculum, developed over the course of the experiment, was based on the principle of active learning, guiding students through

[2]According to Schweinhart et al. (2005), "4 children did not complete the preschool program because they moved away and 1 child died shortly after the study began." We do not know the socioeconomic status (SES) and mother's working status of two siblings among these five children. We also do not know their IQs and the gender of the sibling in wave 1, although we know that the sibling in wave 0 is male. (We use the Perry convention that wave 0 is the first wave and wave 4 is the last one.) The baseline information on these two siblings is important in our formal model of the randomization protocol. We do not make assumptions regarding the gender of the sibling in wave 1 and the mother's working status of the siblings at baseline. In other words, we let these variables take values in $\{0, 1\}$. However, to keep our model computationally feasible, we impute the IQs and SES of these siblings based on a regression of these variables on gender, mother's working status, their interaction term, and an indicator for wave. However, if we had additional computational power, we could conduct our analysis with much weaker assumptions regarding the IQs and SES without resorting to imputations. Our imputations are among the limitations of our study. However, we do not impose any assumption on the gender of the sibling in wave 1 and the mother's working status of the siblings. Note that while we use the data on the five dropped children in our simulations of the randomization protocol for our approximate randomization tests, we treat the five participants as ignorable in our estimation of the treatment effects. Thus, our effective sample for estimation and inference is the core sample of 123 children.

[3]We find no difference in mortality between treatments and controls. Appendix Section 5 presents this evidence.

[4]The Abecedarian program is an example (see, e.g., Campbell et al., 2002). Cunha et al. (2006) and Elango et al. (2015) discuss a variety of these programs and compare their intensity.

[5]An exception is that the first entry cohort received only 1 year of treatment, beginning at age 4. In our estimation of treatment effects, we pool all five cohorts, even though the lower program intensity in the first cohort might in principle attenuate the magnitudes of the effects downward.

the formation of key developmental factors using intensive child-teacher interactions (Schweinhart et al., 1993; Weikart et al., 1978).

Eligibility Criteria The program enrolled five cohorts in the early 1960s, drawn from the catchment area surrounding the Perry Elementary School. Candidate families for the study were identified from a survey of the families of the students attending the Perry Elementary School, by neighborhood group referrals, and through door-to-door canvassing. The eligibility rules for participation were that the participants (i) be African-American; (ii) have low IQs at study entry;[6] and (iii) be disadvantaged as measured by an index of socioeconomic status based on parental employment level, parental education, and housing density (persons per room). The Perry families were more disadvantaged than most other African-American families in the United States, but were representative of a large segment of the disadvantaged African-American population. Heckman et al. (2010a) discuss the issue of the representativeness of the program compared to the general African-American population.

Heckman et al. (2010a) show that the Perry participants were particularly disadvantaged even when compared to those living in the disadvantaged community surrounding the school. We do not know whether, among eligible families in the Perry catchment, those who volunteered to participate in the program were more motivated than other families, and whether this greater motivation would have translated into better child outcomes. However, according to Weikart et al. (1978), "virtually all eligible children were enrolled in the project," so this potential concern appears to be unimportant.

2.1 Randomization Protocol

Understanding the randomization protocol is essential in constructing valid classical frequentist inference for any experiment. As noted by Bruhn and McKenzie (2009), many experimental stud-

[6]The initial eligibility criteria specified that the IQs, as measured by the Stanford–Binet IQ test (1960s norming), be between 70 and 85. (The average IQ in the general population is 100 by construction, so the upper limit is one standard deviation below the average. However, in practice, the IQ range was 61 to 88. Only about two-thirds of the participants had IQs in the range specified initially.

ies in economics do not report the complete set of rules (e.g., balancing criteria) used to form experimental samples and conduct hypothesis tests that ignore the exact randomization protocols. In analyzing the Perry data, this issue is salient. Reports vary about the procedure used and the exact rules followed in creating the experimental sample. We discuss the various descriptions of the randomization protocols. The available descriptions are vague and inconsistent across texts. We account for this ambiguity in designing and interpreting our hypothesis tests.

Before the Perry staffers began the randomization procedure in each of the last four Perry cohorts, any younger siblings of participants enrolled in previous waves were separated from children of freshly recruited families, whom we term singletons (Schweinhart, 2013; Schweinhart et al., 1985). As Schweinhart et al. (1985) explain,

> "*[A]ny siblings were assigned to the same group [either treatment or control] as their older siblings in order to maintain the independence of the groups.*"

This does not apply to the very first cohort by construction. The singletons from new families are then randomized into the two experimental groups as follows. Weikart et al. (1978) detail the first step of the randomization protocol:

> "*First, all [singletons] were rank-ordered according to Stanford–Binet [IQ] scores. Next, they were sorted (odd/ even) into two groups.*"

At the end of this step, the singletons are divided into two groups, one comprising those with even IQ ranks and another with odd IQ ranks. The latter group has one additional person if the singletons are odd in number; otherwise, the sizes of the two groups are equal.

In the second step, children are exchanged between the two groups to balance the mean of an index of socioeconomic status (SES), the proportions of boys and girls, and the proportion of children with working mothers, in addition to mean IQ (Schweinhart et al., 1993). The number of exchanges is not specified, and the exchanges are not necessarily restricted to those between

consecutively ranked pairs.[7] After the first two steps, there are two undesignated groups that differ in number by at most one, and the two groups are balanced with respect to mean IQ, mean SES, percentage of boys, and the proportion of children with working mothers, in a manner acceptable to the staffers, using balancing rules that are undocumented.

All sources agree that a toss of a fair coin decides assignment of the two groups to treatment and control conditions in the third step. The fourth and final step concerns children with working mothers who are placed in the treatment group after the third step. In the fourth step, some of these children are transferred to the control group (Schweinhart et al., 1993, 1985; Schweinhart and Weikart, 1980; Zigler and Weikart, 1993). Although there is no consistent account of the number of transfers, the sources describe the fourth step as involving one-way transfers of some children of working mothers from the treatment group to the control group.[8] Weikart et al. (1978) provide reasons for the transfers: "*no funds were available for transportation or full day care, and special arrangements could not always be made*." We interpret this statement as implying that special arrangements could be made for at least some working mothers to enable their children to attend preschool and participate in home visits if placed in the treatment group. We assume that the staffers are impartial as to which working mothers get special arrangements.

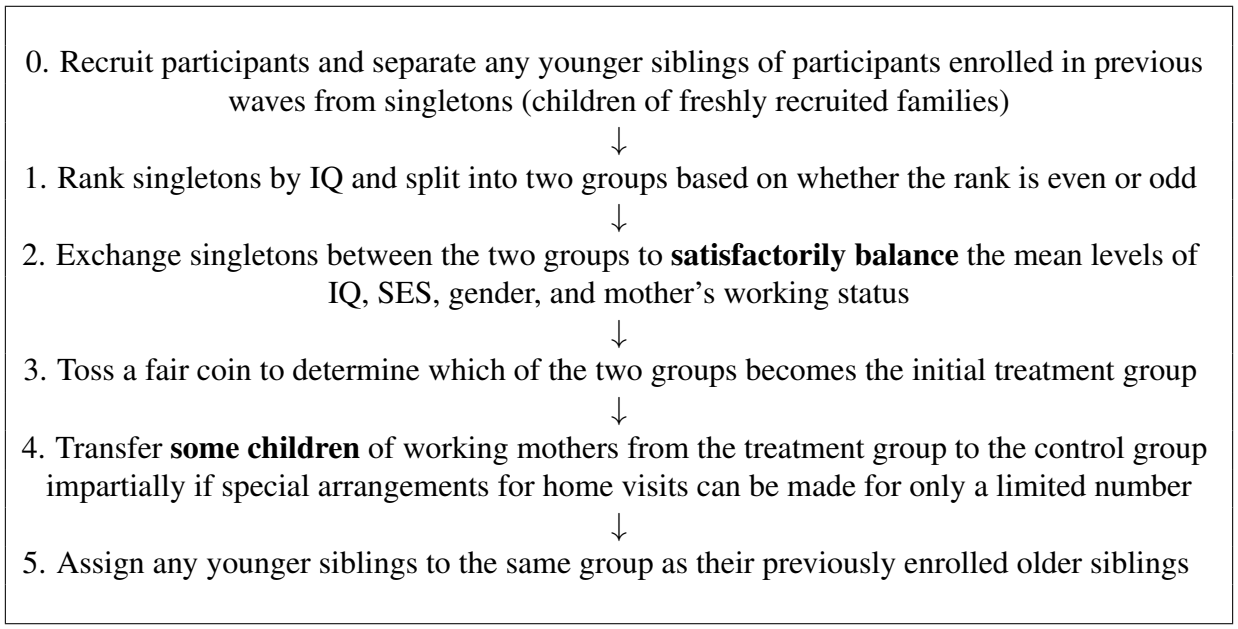
Figure 1 summarizes the randomization protocol. It highlights the sources of ambiguity in boldface: (a) the undocumented criteria and rules used to satisfactorily balance the two undes-

[7]According to Schweinhart et al. (1993), "*They exchanged several similarly ranked pair members so the two groups would be matched on [the baseline variables]*." Even though the phrase "similarly ranked pair members" might suggest consecutively ranked members, this is not necessarily the case. We use Perry data from wave 4 as an example to conclude that the exchanges were not necessarily between consecutively ranked pairs. In wave 4, there were 19 participants, excluding any younger siblings in the program. The IQs of these 19 people were: 61, 71, 75, 76, 76, 76, 78, 78, 79, 79, 80, 80, 81, 82, 83, 83, 83, 85, 88, involving many ties. Regardless of which method was used to break the ties, from a pure ranking procedure they would have obtained two initial groups: one with IQs {61, 75, 76, 78, 79, 80, 81, 83, 83, 88} and another group with IQs {71, 76, 76, 78, 79, 80, 82, 83, 85}. The final observed treatment group has IQs in the set: {61, 75, 76, 78, 80, 81, 83, 83, 83, 88}. Note that the person with IQ 79 is replaced with a person with IQ 83. The final observed control group has IQs in the set: {71, 76, 76, 78, 79, 79, 80, 82, 85}. Note that the person with IQ 83 is replaced with a person with IQ 79. From this, we can conclude that an exchange happened between participants with IQs 79 and 83, who do not comprise a consecutively ranked pair. Thus, after the IQ rank ordering, the exchanges between the two initial groups were not always within consecutively ranked IQ pairs. Thus, in the second step the Perry staffers did not strictly implement a matched pair design.

[8]This is also manifested in the observed data. For example, as explained later in Section 3.2, the number of singletons in wave 2 is 22, with 12 in the control group and 10 in the treatment group. If there were exchanges between the initial experimental groups instead of one-way transfers to the control group, there would have been 11 singletons in both the control and treatment groups instead of 12 and 10, respectively.

ignated groups with respect to the mean levels of baseline variables in the second step; and (b) the nature of constraints on the provision of special home visitation arrangements for children of working mothers.

Figure 1: Schematic of the Actual Randomization Protocol



3 Modeling the Randomization Protocol and Bounding the Unknown Parameters

This section develops a formal model of the randomization protocol consistent with the imprecise verbal accounts reported in Section 2. We model the Perry staff's satisficing behavior with respect to balancing the baseline covariate means of the experimental groups in the first two steps. We recognize and bound the capacity constraints on home visits for children of working mothers. Using our model, we can bound the level of covariate balance the staffers deem acceptable at the end of the first two stages of the protocol. We can also bound the possible number of transfers at the fourth stage of the assignment procedure. Our model and the identified bounds are used in

Section 5 to conduct worst-case randomization tests using least-favorable null distributions (for treatment effects) constructed using our knowledge of the randomization protocol.

3.1 Formalizing the Randomization Protocol

Let $\mathcal{S}_c$ be the set of unique identifiers of participants in cohort[9] $c \in \{0,1,2,3,4\}$ with no elder siblings already enrolled in the Perry Preschool Project. The cardinality of the set of singletons is $|\mathcal{S}_c|$.[10] The participants in the set $\mathcal{S}_c$ are ranked according to their IQs by the Perry staffers, using an undocumented method to break any ties. The participants with odd and even ranks are then split into two undesignated groups, with $\lceil |\mathcal{S}_c|/2 \rceil$ and $\lfloor |\mathcal{S}_c|/2 \rfloor$ members, respectively.[11] Staffers exchange participants between the two groups until the mean levels of four variables (Stanford–Binet IQ, index of socioeconomic status, gender, and mother's working status) are balanced to their satisfaction.[12] The exact metric the staffers used to determine satisfactory covariate balance is not documented.

In this paper, we assume that they use Hotelling's multivariate two-sample t-squared statistic τ_c^2, which is closely related to the Mahalanobis distance metric widely used for matching.[13] The Hotelling statistic for the observed experimental groups in each cohort does not correspond to the

[9]Each of the cohorts corresponds to one of the five waves (labeled 0 through 4) of study participants recruited from the fall seasons of 1962 through 1965. Waves 0 and 1 were randomized in the fall of 1962, while the waves 2, 3, and 4 were randomized in the fall of 1963, 1964, and 1965, respectively. We follow the labeling convention for the cohorts by the Perry analysts who designate the first cohort as "0."

[10]Note that the other participants in cohort c who are not singletons have older siblings already enrolled in the Perry experiment in a previous wave. The non-singletons are not randomized but rather assigned to the same treatment status as their elder siblings already enrolled in the study.

[11]Note that $\lceil \cdot \rceil \equiv \text{ceil}(\cdot)$ is the ceiling function and that $\lfloor \cdot \rfloor \equiv \text{floor}(\cdot)$ is the floor function that assigns least upper integer bounds and greatest lower integer bounds to the argument in the function.

[12]Note that an exchange means a swap between two participants belonging to different undesignated groups. Since the Perry experiment did not use a matched pair design, an exchange or swap is not restricted to occur between participants with consecutive IQ ranks. Exchanges between participants with non-consecutive IQ ranks are also allowed.

[13]The Hotelling's multivariate two-sample t-squared statistic τ_c^2 maps a partition $(\mathcal{A},\mathcal{B})$ of $\mathcal{S}_c$ (such that $|\mathcal{A}| = \lceil |\mathcal{S}_c|/2 \rceil$ and $\mathcal{B} = \mathcal{S}_c \setminus \mathcal{A}$) to $\mathbb{R}_{\geq 0}$ and is given by $\tau_c^2(\mathcal{A},\mathcal{B}) = (\bar{Z}_\mathcal{A} - \bar{Z}_\mathcal{B})' \left(|\mathcal{A}|^{-1}\hat{\Sigma}_\mathcal{A} + |\mathcal{B}|^{-1}\hat{\Sigma}_\mathcal{B}\right)^{-1} (\bar{Z}_\mathcal{A} - \bar{Z}_\mathcal{B})$, where $\bar{Z}_\mathcal{A} = |\mathcal{A}|^{-1} \sum_{i\in\mathcal{A}} Z_i$, with Z_i as the vector containing the i-th participant's IQ, index of socioeconomic status, gender, and mother's working status, $\bar{Z}_\mathcal{B} = |\mathcal{B}|^{-1} \sum_{i\in\mathcal{B}} Z_i$, and $\hat{\Sigma}_\mathcal{A} = (|\mathcal{A}|-1)^{-1} \sum_{i\in\mathcal{A}} (Z_i - \bar{Z}_\mathcal{A})(Z_i - \bar{Z}_\mathcal{A})'$, while $\hat{\Sigma}_\mathcal{B} = (|\mathcal{B}|-1)^{-1} \sum_{i\in\mathcal{B}} (Z_i - \bar{Z}_\mathcal{B})(Z_i - \bar{Z}_\mathcal{B})'$. We use this metric for computational feasibility. If adequate computational power was available, we could also incorporate into our model the raw mean differences in the four variables, studentized versions of such mean differences, or other measures of mean differences between two groups. Of course, it is possible that the Perry staffers were just eyeballing mean differences and did not use any formal metric.

possible minimum value of the Hotelling statistic and is sometimes far away from it.[14] Thus, it appears that program officials were satisficing rather than optimizing (minimizing covariate imbalance) in constructing the two groups.

This process results in a partition $(\mathcal{A}_c^*, \mathcal{B}_c^*)$ of the set $\mathcal{S}_c$ chosen uniformly from

$$\mathbb{U}_c(\delta_c) = \{(\mathcal{A}, \mathcal{B}) : \mathcal{A} \subset \mathcal{S}_c,\ \mathcal{B} = \mathcal{S}_c \setminus \mathcal{A},\ |\mathcal{A}| = \lceil |\mathcal{S}_c|/2 \rceil,\ \tau_c^2(\mathcal{A}, \mathcal{B}) \leq \delta_c\}, \tag{1}$$

where δ_c is a satisficing threshold that captures how stringent or lax the Perry staffers were in trying to balance the mean levels of the two groups.[15] Define $D_{i,c}^{(0)}$ as an indicator of whether participant $i \in \mathcal{S}_c$ belongs to $\mathcal{A}_c^*$: $D_{i,c}^{(0)} = \mathbb{I}\{i \in \mathcal{A}_c^*\}$.

In the next stage, the Perry staffers flip a fair coin to determine whether $\mathcal{A}_c^*$ or $\mathcal{B}_c^*$ becomes the preliminary treatment group. Let Q_c be an indicator of whether the coin flip results in a head. If $Q_c = 1$, then $\mathcal{B}_c^*$ becomes the treatment group. If $Q_c = 0$, then $\mathcal{A}_c^*$ becomes the treatment group. Let $D_{i,c}^{(1)}$ denote membership in the preliminary treatment group. Thus

$$D_{i,c}^{(1)} = Q_c\,(1 - D_{i,c}^{(0)}) + (1 - Q_c)\,D_{i,c}^{(0)}. \tag{2}$$

In the next step, some children of working mothers initially placed in the treatment group

[14] For cohort 0, the proportion of possible group formations with a lower Hotelling statistic is at least 29.24%. The corresponding numbers for cohorts 1, 2, 3, and 4 are 64.51%, 14.79%, 15.52%, 75.56%, respectively.

[15] The satisficing threshold δ_c is the maximum level of covariate imbalance that satisficed Perry staffers. The threshold δ_c is unknown to the analyst but can be partially identified, as explained later. We assume a uniform probability over $\mathbb{U}_c$ for the choice of the partition $(\mathcal{A}_c^*, \mathcal{B}_c^*)$ for the purpose of keeping the model simple and computationally feasible. In general, we might suspect the following: given two partitions of $\mathcal{S}_c$ with the same level of Hotelling's statistic, there might have been a higher probability mass on the partition closer to the initial grouping based on odd and even IQ ranks. In addition, the staffers might have also preferred not to make additional exchanges if they expected relatively insignificant reductions in covariate imbalance. In other words, the probability that the Perry staffers chose a particular partition $(\mathcal{A}_c^*, \mathcal{B}_c^*)$ could have depended on their preferences over substitution between two things: similarity of $(\mathcal{A}_c^*, \mathcal{B}_c^*)$ to the initial IQ rank-based grouping; and the level of covariate imbalance (as measured by Hotelling's statistic) resulting from the partition $(\mathcal{A}_c^*, \mathcal{B}_c^*)$. However, there is no unique way to formalize this notion. Such a general model may not even be computationally feasible.

are transferred to the control group.[16] To model this process, we introduce additional notation. Define $M_i = 1$ as an indicator of whether participant i's mother was working at baseline, for all $c \in \{2,3,4\}$. For groups 2 and higher, let m_c be the number of children of working mothers initially placed in the treatment group: $m_c = \sum_{i \in \mathcal{S}_c} M_i D_{i,c}^{(1)}$. For the entry cohorts, let $m_{0,1}$ be the number of children of working mothers initially placed in the treatment group for cohorts 0 and 1, that is, $m_{0,1} = \sum_{i \in \mathcal{S}_0 \cup \mathcal{S}_1} M_i D_{i,c}^{(1)}$.[17]

Define η_c as a parameter indicating the maximum number of children of working mothers in cohort $c \in \{2,3,4\}$ for whom special arrangements could be made to enable home visits.[18] We define $\eta_{0,1}$ to be the parameter indicating the maximum number of children of working mothers in the pooled cohorts 0 and 1 for whom special home visitation arrangements could be made, averting their transfer to the control group if placed in the initial treatment group.[19]

Special arrangements are made for $\min(\eta_{0,1}, m_{0,1})$ children of working mothers in the entry cohorts and for $\min(\eta_c, m_c)$ such children in each cohort $c \in \{2,3,4\}$ to enable special home visits (as opposed to weekday home visits for children of non-working mothers). If there are any remaining children with working mothers in the initial treatment group, they are transferred to the control group. We assume that the Perry staffers impartially chose (with equal probability) the children for whom the special accommodations are made.[20] To formalize this assumption, let $V_{i,c}$

[16]The Perry teachers conducted special home visits for working mothers at times other than weekday afternoons, when they visited the homes of non-working mothers. Because of logistical and financial constraints, the teachers were able to visit the homes of only a limited number of working mothers at times other than weekday afternoons. Thus, the children of working mothers in the preliminary treatment group for whom these special arrangements could not be made were transferred to the control group.

[17]The reason for this slightly different notation for the entry cohorts is given later.

[18]Thus, η_c can be thought of as slots available for special visits to the homes of working mothers. Equivalently, it is the number of children of working mothers who would remain in the final treatment group if all of them were placed in the preliminary treatment group.

[19]Note that cohorts 0 and 1 were both randomized in the fall of 1962, while each of the subsequent cohorts were randomized in separate years from 1963 through 1965. Since cohorts 0 and 1 had a common set of teachers, they share the number of slots available for the special home visits. Thus, we pool these two cohorts while defining $m_{0,1}$ and $\eta_{0,1}$. However, cohorts 2 through 5 have separate parameters for the slots available for special home visits.

[20]We are implicitly assuming that (i) all working mothers would be able to send their children to preschool and participate in weekly home visits if special arrangements could be made for them and that (ii) all working mothers have a similar kind of availability for alternative home visiting arrangements. The higher the extent to which the reality deviated from these two assumptions, the more unrealistic and limited the assumption that the staffers chose working mothers with equal probability for special arrangements. A model allowing for heterogeneity in availability of working mothers (for special home visiting arrangements) does not appeear to be computationally feasible.

be a binary indicator of whether the participant $i \in \mathcal{S}_c$ was in the initial treatment group but was transferred to the control group for a lack of special accommodations for home visits. The vector $(V_{i,c} : i \in \mathcal{S}_c,\, M_i D_{i,c}^{(1)} = 1)$ is assumed to be drawn uniformly from the set $\{v \in \{0,1\}^{m_c} : ||v||_1 = \min(\eta_c, m_c)\}$ for all $c \in \{2,3,4\}$. Since the two entry cohorts face a common capacity constraint with respect to special home visitation accommodations, the vector $(V_{i,c} : i \in \mathcal{S}_0 \cup \mathcal{S}_1,\, M_i D_{i,c}^{(1)} = 1)$ is assumed to be drawn uniformly from the set $\{v \in \{0,1\}^{m_{0,1}} : ||v||_1 = \min(\eta_{0,1}, m_{0,1})\}$. In addition, note that $V_{i,c} = 0$ (by construction) for participants $i \in \mathcal{S}_c$ such that $M_i D_{i,c}^{(1)} = 0$ for all $c \in \{0,1,2,3,4\}$.[21] In this notation, the participant's final treatment status $D_{i,c}^{(2)}$ is given by

$$D_{i,c}^{(2)} = M_i D_{i,c}^{(1)} V_{i,c} + (1 - M_i D_{i,c}^{(1)}) D_{i,c}^{(1)}. \tag{3}$$

Any Perry subjects with identifiers not in $\bigcup_{c=0}^{4} \mathcal{S}_c$ receive the same treatment status as their elder siblings already enrolled in the Perry study. Thus, the final treatment status D_i of the i-th subject is given by $D_i = D_{i,c}^{(2)}$ if $i \in \bigcup_c \mathcal{S}_c$. Otherwise, if participant i is not from a freshly recruited family, the assignment is given by $D_i = D_h$, where the h-th subject is the i-th subject's eldest sibling enrolled in the Perry study, if $i \in \mathcal{I} \setminus \bigcup_c \mathcal{S}_c$, where $\mathcal{I}$ is the set of identifiers of all Perry subjects.

3.2 Partially Identifying Satisficing Thresholds and Capacity Constraints

We now demonstrate how we can partially identify the satisficing thresholds δ_c and the special home visitation capacity constraints η_c using cohorts 2 and 3 as examples. We then present a general framework for partially identifying these parameters.

[21] In other words, $V_{i,c} = 0$ for the participants who were either initially placed in the control group or placed in the initial treatment group but have non-working mothers.

Example 1: Wave 2 (A Case with One Transfer in the Last Stage)

Wave 2	$D_i = 0$	$D_i = 1$	Total
$M_i = 0$	9	7	16
$M_i = 1$	3	3	6
Total	12	10	22

This example discusses the steps for bounding the parameters δ_2 and η_2 for wave 2. Shown above is the contingency table of mother's working status M_i and final treatment status D_i for participants $i \in \mathcal{S}_2$ in cohort 2 with no elder siblings already enrolled in the Perry study. There are 22 such participants in total. Since there are an even number of participants, each of the initial two undesignated groups (as well as the initial treatment and control groups in the next stage) would have been $\lceil |\mathcal{S}_2|/2 \rceil = \lfloor |\mathcal{S}_2|/2 \rfloor = 11$ in size. However, we observe only 10 members in the final treatment group but 12 members in the final control group. This implies that there must have been one transfer from the initial treatment group to the control group. Thus, one of the 3 children of working mothers in the final control group was in the initial treatment group. However, we do not know exactly which one of these children was transferred, so there are 3 possibilities for the initial treatment group. Let $\tau^2_{2,1}, \tau^2_{2,2}, \tau^2_{2,3}$ be the Hotelling two-sample statistics for these three possibilities. One of these Hotelling statistics was the actual level of covariate imbalance between the initial treatment and control groups, and this level of imbalance is assumed to be within the satisficing threshold δ_2 of the Perry staffers (by construction). Thus, $\delta_2 \geq \min\{\tau^2_{2,1}, \tau^2_{2,2}, \tau^2_{2,3}\}$.[22] In addition, $m_2 = 4$, since there must have been 4 children of working mothers in the initial treatment group, consisting of the 3 participants who remain in the final treatment group and the 1 participant who was transferred to the control group. Since 3 of the initial 4 participants remained in the final treatment group, $\min(\eta_2, m_2) = \min(\eta_2, 4) = 3$, implying that $\eta_2 = 3$, the only solution that satisfies the equality. We next present another example.

[22] In our application, $\delta_2 \geq 1.6037804$.

Example 2: Wave 3 (A Case with 1 or 2 Transfers in the Last Stage)

Wave 3	$D_i = 0$	$D_i = 1$	Total
$M_i = 0$	7	9	16
$M_i = 1$	5	0	5
Total	12	9	21

In this example, we show the contingency table of M_i and D_i for the 21 participants $i \in \mathcal{S}_3$ in cohort 3. The sizes of the larger and smaller undesignated groups would have been $\lceil |\mathcal{S}_3|/2 \rceil = 11$ and $\lfloor |\mathcal{S}_3|/2 \rfloor = 10$, respectively. However, either of these two groups could have been the initial treatment group. Since there are 12 members in the final control group and 9 in the final treatment group, there are two possible cases: if the initial treatment group had 10 members, there would have been $10 - 9 = 1$ transfer; but if it had 11 members, there would have been $11 - 9 = 2$ transfers. Since the number of transfers involving children of working mothers is either 1 or 2, the number of possibilities for the initial treatment group is $\binom{5}{1} + \binom{5}{2} = 5 + 10 = 15$, since all the 5 children of working mothers in this cohort are in the control group. Let $\tau^2_{3,1}, \ldots, \tau^2_{3,15}$ be the Hotelling statistics for those 15 possibilities. Then, $\delta_3 \geq \min\{\tau^2_{3,1}, \ldots, \tau^2_{3,15}\}$.[23] In addition, $m_3 \in \{1, 2\}$, since m_3 is the sum of the number of transfers (either 1 or 2) and the number of remaining children in the final treatment group (0 in this cohort). As no working mother remained in the treatment group, $\min(\eta_3, m_3) = 0$, implying that $\eta_3 = 0$, which is the only number consistent with this equality. Thus, the Perry staffers were unable to provide special home visitation accommodations for any of these 21 participants.

In Appendix Section 1, we present another example detailing partial identification of model parameters for wave 4. We next present a more general framework for partially identifying the satisficing thresholds and capacity constraints on the special home visits.

[23] In our application, $\delta_3 \geq 1.1309983$.

Partial Identification of the Satisficing Thresholds and Capacity Constraints in General

Wave c	$D_i = 0$	$D_i = 1$	Total
$M_i = 0$	$\omega_{0,0}$	$\omega_{0,1}$	$\omega_{0,*}$
$M_i = 1$	$\omega_{1,0}$	$\omega_{1,1}$	$\omega_{1,*}$
Total	$\omega_{*,0}$	$\omega_{*,1}$	$\lvert\mathcal{S}_c\rvert$

The above contingency table shows that there are $\omega_{m,d}$ participants with $(M_i, D_i) = (m, d) \in \{0,1\}^2$ among the participants $\mathcal{S}_c$ in cohort c.[24] The total number of children with non-working mothers is $\omega_{0,*} = \omega_{0,0} + \omega_{0,1}$ and that of working mothers is $\omega_{1,*} = \omega_{1,0} + \omega_{1,1}$. The total number of participants in the final control group is $\omega_{*,0} = \omega_{0,0} + \omega_{1,0}$ and that in the final treatment group is $\omega_{*,1} = \omega_{0,1} + \omega_{1,1}$. The partial identification of the satisficing thresholds and capacity constraints would vary depending on whether $|\mathcal{S}_c|$ is even or odd and also depending on whether $\omega_{*,1} = \lceil |\mathcal{S}_c|/2 \rceil$ or $\omega_{*,1} < \lceil |\mathcal{S}_c|/2 \rceil$. We discuss these cases in Appendix Section 1.

3.3 Applicability of Our Procedure to Other Experiments

Our model and bounding procedures are applicable for analyzing other experiments with appropriate modifications, including randomized trials conducted in developing countries. Bruhn and McKenzie (2009) survey 25 leading researchers, half of whom conducted 5 or more experiments. According to Bruhn and McKenzie (2009), these leading researchers had access to baseline data during the randomization phase for 71% of the experiments they conducted. One of the survey respondents admitted that they

> "*regressed variables like education on assignment to treatment, and then re-did the assignment if these coefficients were 'too big.*"'

In the authors' survey of 25 leading researchers, 52% admit to "subjectively deciding whether to redraw" and 15% admit to "using a statistical rule to decide whether to redraw" the treatment

[24] Note that all cohorts are of equal size so that $\omega_{m,d} \equiv \omega_{m,d,c}$ for all $(m,d) \in \{0,1\}^2$ but we suppress the subscript c for simplicity.

assignment vector in at least one of the experiments they conducted.[25] The authors conclude that

> "*this reveals common use of methods to improve baseline balance, including several rerandomization methods not discussed in print.*"

With appropriate modifications, our model of satisficing thresholds directly applies to experiments conducted in such a subjective and incompletely documented manner. Suitable adjustments include replacing Hotelling's statistic in our model with regression coefficients or other metrics actually used to measure covariate balance between the treatment and control groups. Our methods for partially identifying underlying randomization rules can be used when the subjective statistical thresholds are not documented. If rerandomization criteria are specified before carrying out treatment assignment, there exist simpler methods for conducting inference when the experimental design involves a precisely-specified and strictly-followed rerandomization procedure that is not as complex or compromised as was the Perry experiment (see, e.g., Li et al., 2018; Morgan and Rubin, 2012, 2015).[26]

Additionally, in some social experiments, post-randomization transfer of some participants from the control to the treatment group can occur if additional funding for the intervention becomes available. For example, wait-list control groups are used in some clinical studies. While this is the reverse of what occurred in the Perry experiment, our model (with appropriate modifications) can be applied without loss of generality. Overall, our approach can be adapted to analyze other experiments across multiple disciplines.

[25] These percentages are calculated by weighting each survey respondent by the number of experiments in which the respondent has participated.

[26] Morgan and Rubin (2012) state that they "only advocate rerandomization if the decision to rerandomize or not is based on a pre-specified criterion." Their inferential methods require knowledge of such pre-specified criteria. Even if experimenters do not specify the specific criteria for acceptable randomizations beforehand, analysts of such experiments can use our methods to conduct randomization-based hypothesis tests if it is known that the experimenters used a satisficing threshold with respect to a known metric. Although rerandomization methods have the property that they reduce variance of the null distribution asymptotically in certain settings (Li et al., 2018; Morgan and Rubin, 2012, 2015), this property is not guaranteed in a finite-sample setting.

4 Estimators of Treatment Effects

This section discusses our estimators of Perry treatment effects. Let D_i represent the treatment status of participant i, and let Z_i be his or her vector of the four pre-program covariates used during the randomization phase, i.e., Stanford-Binet IQ, index of socioeconomic status, gender, and mother's working status. In addition, let Y_i denote an outcome of interest of participants i in a relevant subsample $\mathcal{P}$ containing $N_{\mathcal{P}} = |\mathcal{P}|$ participants. We are interested in estimating the average treatment effect $\bar{\tau}$ in the subsample $\mathcal{P}$ given by

$$\bar{\tau} = \frac{1}{N_{\mathcal{P}}} \sum_{i \in \mathcal{P}} \left(Y_i^1 - Y_i^0 \right), \tag{4}$$

where Y_i^d is the counterfactual outcome of participant i when his or her treatment status D_i is fixed at $d \in \{0, 1\}$, using the observed data

$$Y_i = (1 - D_i)\, Y_i^0 + D_i\, Y_i^1. \tag{5}$$

We estimate the average treatment effects[27] of the Perry intervention on various outcomes using three different estimators, and test hypotheses about these treatment effects using randomization-based inference and other traditional inferential methods. The estimators we use are the unconditional difference-in-means (UDIM) estimator, the conditional ordinary least squares (COLS) estimator, and the augmented inverse probability weighting (AIPW) estimator. For completeness, we describe each of them in turn and what problems they address.

We might estimate the treatment effect parameter defined in equation (4) by the unconditional difference-in-means (UDIM) estimator $\hat{\Pi}_{\text{udim}}$ as follows:

$$\hat{\Pi}_{\text{udim}} = \frac{\sum_{i \in \mathcal{P}} R_i\, D_i\, Y_i}{\sum_{i \in \mathcal{P}} R_i\, D_i} - \frac{\sum_{i \in \mathcal{P}} R_i\, (1 - D_i)\, Y_i}{\sum_{i \in \mathcal{P}} R_i\, (1 - D_i)}, \tag{6}$$

[27] We focus mainly on the average treatments because many of our outcomes of interest are binary in nature. However, in the Appendix we also consider distributional treatment effects for important continuous outcomes.

where R_i is a binary indicator of whether we have the outcome Y_i on record.[28] This estimator assumes that the treatment status D_i is unconditionally independent of the counterfactual outcomes (Y_i^1, Y_i^0).[29]

In fact, the randomization procedure used in the Perry experiment only justifies conditional independence: $(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i \mid Z_i$. Exploiting this property and controlling for Z_i in a regression of Y_i on D_i and Z_i, we obtain the conditional ordinary least squares (COLS) estimator $\hat{\Pi}_{\text{cols}}$ given by

$$\hat{\Pi}_{\text{cols}} = e_i' \left(\sum_{i \in \mathcal{P}} R_i X_i X_i' \right)^{-1} \left(\sum_{i \in \mathcal{P}} R_i X_i Y_i \right), \tag{7}$$

where $X_i = (D_i, Z_i, 1)$ and e_i is the standard unit vector $(1, 0, 0, 0, 0, 0)$.[30]

The UDIM and COLS estimators assume that non-response is determined at random. However, R_i, an indicator of whether Y_i is missing, could depend on the treatment status D_i and the pre-program covariates Z_i. The augmented inverse probability weighting (AIPW) estimator allows for this possibility by using a weaker assumption that $Y_i \perp\!\!\!\perp R_i \mid D_i,\ Z_i$, i.e., the outcome is independent of non-response status conditional on the treatment status and pre-program covariates. The AIPW estimator of the treatment effect is given by

$$\hat{\Pi}_{\text{aipw}} = \frac{1}{N_{\mathcal{P}}} \sum_{i \in \mathcal{P}} \left(\hat{\pi}_i^1 - \hat{\pi}_i^0 \right), \tag{8}$$

where

$$\hat{\pi}_i^d = \hat{Y}_i^d + \frac{\mathbb{I}\{R_i = 1,\ D_i = d\}}{\hat{\lambda}_i^d \, \hat{\phi}_i^d} \left(Y_i^d - \hat{Y}_i^d \right). \tag{9}$$

In this expression, $\hat{Y}_i^d$ is the gender-specific ordinary least squares-based estimator of the condi-

[28] The indicator R_i equals 0 if Y_i is missing and equals 1 if Y_i is not missing. In addition, the estimator given by equation (6) can be studentized using its cluster-robust asymptotic standard error, allowing for correlation between error terms of the participant-siblings in the experiment.

[29] An additional assumption is that R_i (whether Y_i is missing) is determined completely at random.

[30] This estimator can also be studentized using its cluster-robust asymptotic standard error, allowing for correlation between error terms of the participant-siblings in the experiment. Note that the COLS estimator also assumes that R_i is determined at random.

tional expectation $\mathbb{E}[Y_i \mid Z_i, D_i = d, R_i = 1]$ for $d \in \{0,1\}$,[31] $\hat{\phi}_i^d$ is an estimator of $\Pr(D_i = d \mid Z_i)$, the i-th participant's propensity of being in the experimental state d, and $\hat{\lambda}_i^d$ is an estimator of $\Pr(R_i^1 = 1 \mid Z_i, D_i = d)$, the propensity of having a non-missing outcome after fixing the treatment status D_i, for $d \in \{0,1\}$.[32] The AIPW estimator adjusts the outcome based on pre-program covariates and corrects for covariate imbalance and various forms of non-response.[33] It is asymptotically unbiased due to a double robustness property: the estimator is robust to misspecification of either the propensity score models or the models for counterfactual outcomes, but not both (Kang and Schafer, 2007; Lunceford and Davidian, 2004; Robins et al., 1994).[34] For this reason, we prefer the AIPW estimator over the UDIM and COLS estimators. However, we present estimates from all of these procedures in the Appendix as a form of sensitivity analysis.[35]

[31] Specifically, $\hat{Y}_i^d = (Z_i, 1)'(\sum_{k \in \mathcal{G}_i^d}(Z_k, 1)(Z_k, 1)')^{-1}(\sum_{k \in \mathcal{G}_i^d}(Z_k, 1)Y_k)$, where G_i indicates whether the i-th participant is male and $\mathcal{G}_i^d = \{k : D_k = d, R_k = 1, G_k = G_i\}$, for $d \in \{0,1\}$.

[32] All of the propensity scores are estimated using a logit specification and the penalized maximum likelihood method of Greenland and Mansournia (2015), which circumvents the issue of separation in small samples.

[33] The AIPW estimator also assumes conditional independence of the counterfactual outcomes and the treatment status, i.e., $(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i \mid Z_i$, which is valid because of the random assignment of the treatment status conditional on pre-program variables. Note that the propensity score model used in the AIPW estimator is a direct consequence of the law of conditional probability: $\Pr(R_i = 1, D_i = d \mid Z_i) = \Pr(R_i^1 = 1 \mid Z_i, D_i = d)\Pr(D_i = d \mid Z_i)$ for $d \in \{0,1\}$. In the econometrics literature, the AIPW estimator is better known as a type of efficient influence function (EIF) estimator (Cattaneo, 2010). The estimator given by equation (8) can be studentized using the empirical sandwich standard error under the assumption that the propensity score and regression models are both correctly specified (Lunceford and Davidian, 2004). For studentization, we use a cluster-robust version of this asymptotic standard error, given by the following formula: $\frac{1}{N_{\mathcal{P}}}[\sum_{j \in J}(\sum_{i \in \mathcal{F}_j} \hat{\pi}_i^1 - \hat{\pi}_i^0 - \hat{\Pi}_{\text{aipw}})^2]^{1/2}[|J|/(|J|-1)]^{1/2}$, where $\mathcal{F}_j$ represents a cluster of participant-siblings in the set J of clusters. Our studentized test statistics are based on the asymptotic standard error mainly for computational ease, but studentization based on the bootstrap standard error would be superior in theory.

[34] The double-robustness property (consistency despite certain forms of misspecification) is easier to understand by rewriting equation (9) as follows: $\hat{\pi}_i^d = Y_i^d + (\hat{\lambda}_i^d\,\hat{\phi}_i^d)^{-1}(\mathbb{I}\{R_i = 1,\, D_i = d\} - \hat{\lambda}_i^d\,\hat{\phi}_i^d)(Y_i^d - \hat{Y}_i^d)$ for $d \in \{0,1\}$. If the propensity score models are correctly specified, the average value of $\hat{\lambda}_i^d\,\hat{\phi}_i^d$ consistently estimates the probability that $\mathbb{I}\{R_i = 1,\, D_i = d\} = 1$, in which case the sample average of the whole second term in the rewritten expression for $\hat{\pi}_i^d$ tends to zero. If, on the other hand, the counterfactual outcome model is correctly specified, then the average value of $\hat{Y}_i^d$ consistently estimates the expectation of Y_i^d, again in which case the sample average of the whole second term in the rewritten expression for $\hat{\pi}_i^d$ converges to zero. Thus, the AIPW estimator remains consistent for the average treatment effect if either the propensity score models or the counterfactual outcome models are misspecified but not both. See Robins et al. (1994), Lunceford and Davidian (2004), and Kang and Schafer (2007).

[35] The AIPW estimator can become unstable if both the propensity score models and the counterfactual outcome models are misspecified (Kang and Schafer, 2007). Thus, we do not solely rely on the AIPW estimator but use it in conjunction with the UDIM and COLS estimators.

5 Hypotheses of Interest, Test Statistics, and Inference

We seek to test hypotheses regarding the counterfactual outcomes of Perry participants. The conventional way to analyze randomized experiments is to posit a null hypothesis that the average effect of treatment is zero and to proceed testing it with large-sample methods using asymptotic or bootstrap distributions. Given the relatively small size of our sample, reliance on large sample methods could be problematic. We show later that this concern is justified.[36]

In some settings permutation tests can be used to test the null hypothesis that the outcomes in the control group have the same distribution as those in the treatment group without relying on large-sample theory. Permutation tests exploit the property that treatment and control labels within the same strata are exchangeable under the null hypothesis of a common outcome distribution. If randomization of the treatment status did not involve explicit stratification based on baseline covariates, permutation tests need to make restrictive assumptions on the strata within which treatment and control labels are exchangeable. This approach is used by Heckman et al. (2010a).[37]

An alternative to that approach is to use knowledge of the randomization protocol to draw inferences about treatment effects. Once a precise null model of treatment effects is proposed for the purpose of hypothesis testing, we can determine the distribution of estimators generated by the randomization scheme under the null hypothesis. It is then possible to measure the incompatibility of the observed data with respect to the null distribution, which allows us to assess the statistical significance of the estimated treatment effects.

In this section, we first formulate our hypotheses of interest. We then discuss conventional inferential procedures and introduce our worst-case (least favorable) approximate randomization tests. After theoretically justifying our new tests, we assess how their empirical performance compares with that of traditional methods using a Monte Carlo study.

[36] In a set of 53 studies of randomized controlled trials published in some leading economics journals, Young (2019) also finds that experimental results obtained using asymptotic theory are misleading, relative to results based on randomization tests.

[37] However, unless the permutation method reflects the method used for random assignment of the treatment, permutation tests do not in general allow us to test hypotheses about counterfactual outcomes of the individual Perry participants.

5.1 Hypotheses of Interest

The conventional approach specifies a joint distribution F for the vector (Y^1, Y^0, Z), comprising the outcome variable Y^1 under treatment, outcome variable Y^0 under absence of treatment, and background variables Z, at the population level and test the hypothesis $\mathcal{H}_\mathcal{C}$ that Y^1 and Y^0 have equal means, i.e.,

$$\mathcal{H}_\mathcal{C} : \mathbb{E}_F[Y^1 - Y^0] = 0, \tag{10}$$

using the observed data $(Y_i, D_i, Z_i)_{i\in\mathcal{P}}$ under the assumption that (Y_i^1, Y_i^0, Z_i) is distributed according to F for all $i \in \mathcal{P}$. Because each participant in our sample is assigned to either the treatment group or the control group, we only observe either Y_i^1 and Y_i^0 for each participant $i \in \mathcal{P}$. It is of interest to conduct tests about the missing counterfactual outcomes within our sample, even though tests of population-level parameters are more commonly employed.

In this paper, instead of appealing to some hypothetical long-run sampling experiment, we are interested in knowing the properties of the sample in hand. For example, we are interested in testing whether the average treatment effect *within* the sample is zero, i.e.,

$$\mathcal{H}_\mathcal{N} : \frac{1}{N_\mathcal{P}} \sum_{i\in\mathcal{P}} \left(Y_i^1 - Y_i^0\right) = 0. \tag{11}$$

The hypothesis $\mathcal{H}_\mathcal{N}$ is usually attributed to Neyman (1923). A special case of $\mathcal{H}_\mathcal{N}$ is the sharp null hypothesis of no treatment effects *whatsoever* for all participants:

$$\mathcal{H}_\mathcal{F} : \tau_i \equiv Y_i^1 - Y_i^0 = 0 \tag{12}$$

for all $i \in \mathcal{P}$.[38] This is Fisher's (1925; 1935) null hypothesis and is a joint test of zero individual treatment effects. It is trivially equivalent to Neyman's (1923) hypothesis if there is no heterogeneity in the treatment effects $\tau_i \equiv Y_i^1 - Y_i^0$ of the Perry participants $i \in \mathcal{P}$ or if the individual effects

[38] While this formulation states that each individual treatment effect τ_i is zero, the analyst may fix each τ_i at a desired value for hypothesis testing. Such a hypothesis is often called *sharp* because it specifies one set of counterfactual outcomes for the participants.

exhibit uniformity such as $\tau_i \geq 0$ for all $i \in \mathcal{P}$. The advantage of Fisher's hypothesis $\mathcal{H}_{\mathcal{F}}$ is that it provides a testable model in which all the counterfactual outcomes are specified.[39] Such hypothesis testing can be conducted using our knowledge of the randomization protocol without relying on large-sample theory, as we explain below. With all the counterfactual outcomes specified, we can learn about the randomization distribution of our measure of the treatment effect. Then we can see where in that distribution the observed estimate of the treatment effect falls. This would help us understand the extent to which the observed data can be rationalized using the specified null model. See Athey and Imbens (2017) and Abadie et al. (2017) for background on this topic.

Note that Neyman's hypothesis $\mathcal{H}_{\mathcal{N}}$ nests Fisher's sharp null hypothesis $\mathcal{H}_{\mathcal{F}}$. In general there are many configurations of the individual treatment effects that are all consistent with Neyman's hypothesis $\mathcal{H}_{\mathcal{N}}$. Thus, to truly test $\mathcal{H}_{\mathcal{N}}$ just using our knowledge of the randomization protocol, we would need to test each one of all the sharp null hypotheses like $\mathcal{H}_{\mathcal{F}}$ that imply $\mathcal{H}_{\mathcal{N}}$.[40] However, a non-rejection of Fisher's null hypothesis $\mathcal{H}_{\mathcal{F}}$ would imply non-rejection of Neyman's null hypothesis $\mathcal{H}_{\mathcal{N}}$, and so testing other sharp null hypotheses may not be necessary if we are unable to reject $\mathcal{H}_{\mathcal{F}}$. Of course, a rejection of $\mathcal{H}_{\mathcal{F}}$ would not imply a rejection of $\mathcal{H}_{\mathcal{N}}$. However, in the next subsection, we construct tests of Fisher's null hypothesis $\mathcal{H}_{\mathcal{F}}$, given in equation (12), that serve as approximate tests of Neyman's null hypothesis $\mathcal{H}_{\mathcal{N}}$ in equation (11) and the conventional population-level null hypothesis $\mathcal{H}_{\mathcal{C}}$ in equation (10).

We next discuss our inferential methods and test statistics. We first review some conventional methods used in empirical studies. We then discuss our methods for randomization-based design-specific inference and why their use is preferable.

[39] Note that we observe either Y_i^1 or Y_i^0 for each participant $i \in \mathcal{P}$. Thus, under the null model (12), the other counterfactual outcome can be imputed according to the fact that $Y_i^1 = Y_i^0$. In general, if τ_i is hypothesized to be equal to a number τ_i°, the counterfactual outcomes (Y_i^1, Y_i^0) under the null model are equal to $(Y_i + \tau_i^\circ, Y_i)$ if $D_i = 0$ and is equal to $(Y_i, Y_i - \tau_i^\circ)$ if $D_i = 1$ for all $i \in \mathcal{P}$.

[40] When the outcomes under consideration are binary and the experiment involves a completely randomized design, there are are strategies to test Neyman's hypothesis in a computationally feasible way (see, e.g., Li and Ding, 2016; Rigdon and Hudgens, 2015).

5.2 Test Statistics and Inference

5.2.1 Conventional Measures of Statistical Significance

For tests of population-level parameters such as $\mathcal{H}_{\mathcal{C}}$ in equation (10), the most commonly reported measure of statistical significance, the computation of which is facilitated by statistical software packages, is the one-sided asymptotic p-value (based on the analytic standard error) given by

$$p_{A,A} = \Phi(-|\hat{\theta}/\hat{\sigma}_A|), \tag{13}$$

where $\hat{\theta}$ is the estimated treatment effect,[41] $\hat{\sigma}_A$ is its analytic asymptotic standard error,[42] and Φ is the standard normal distribution function.[43] One-sided tests are based on the hypothesis that treatment effects are non-negative. For completely randomized experiments, the value $p_{A,A}$ can also be thought of as the p-value based on a large-sample approximation of the distribution of the estimate, say difference-in-means, over all possible randomizations under the null hypothesis $\mathcal{H}_{\mathcal{N}}$ (Neyman, 1923). Li et al. (2018) derive an asymptotic theory of the difference-in-means estimator in experiments involving rerandomization (using a pre-specified balancing rule using the Mahalanobis distance), for which the asymptotic distribution of the estimator is a linear combination of normal and truncated normal variables.

Resampling methods are also used to quantify statistical uncertainty. For example, the bootstrap standard error $\hat{\sigma}_B$, which is the standard deviation of the bootstrap distribution of the treatment effect estimator, is also often reported alongside estimates in research papers. A bootstrap standard

[41] There are three choices for the treatment effect estimate $\hat{\theta}$ in our case: the unconditional difference-in-means (UDIM) estimate, the conditional ordinary least squares (COLS) estimate, or the augmented inverse probability weighting (AIPW) estimator.

[42] The analytic asymptotic standard error in the case of our UDIM and COLS estimators is the cluster-robust standard error. In the case of the AIPW estimator, the cluster-robust version of the empirical sandwich error given in the previous section serves as the analytic asymptotic standard error.

[43] The two-sided asymptotic p-value, obtained by doubling $p_{A,A}$, is more more frequently reported. However, since doubling involves only a trivial computation, we use the one-sided p-value to compare and contrast this p-value with other p-values presented in this section. Often the standard normal distribution function Φ is also replaced by that of Student's t-distribution when it is straightforward to compute degrees of freedom. However, since the normal distribution approximates the t-distribution well even with as few as 30 degrees of freedom, we retain the use of the normal distribution. Using Φ also allows better comparability between tests using different estimators.

error-based asymptotic p-value is given by

$$p_{A,B} = \Phi(-|\hat{\theta}/\hat{\sigma}_B|), \tag{14}$$

replacing the analytic standard error $\hat{\sigma}_A$ with the bootstrap standard error $\hat{\sigma}_B$ in equation (13).

There are at least two more bootstrap-based p-values that are less frequently used in economics. One is the "type-2 p-value" of Singh and Berk (1994). It is closely connected to Efron's (1979a; 1979b; 1981) percentile method for constructing confidence intervals, which is widely used in statistics and social sciences. Singh and Berk (1994) propose a simple way to obtain a measure of statistical significance using the bootstrap distribution. Their measure, which is termed the "type-2 p-value," is based on the bootstrap distribution of the nonstudentized estimate and is approximated by

$$p_{B,N} = (1 + R_B)^{-1} \left[1 + \sum_{r=1}^{R_B} \mathbb{I}\{\theta_r^* \leq 0\} \right], \tag{15}$$

where R_B is the number of stratified bootstraps based on strata defined by gender and cohort of the participants and θ_r^* is the bootstrapped treatment effect estimate for $r \in \{1, \ldots, R_B\}$,[44] if $\hat{\theta} \geq 0$. If $\hat{\theta} < 0$, the inequality in the equation is reversed so that the type-2 p-value becomes $p_{B,N} = (1 + R_B)^{-1} \left[1 + \sum_{r=1}^{R_B} \mathbb{I}\{\theta_r^* \geq 0\} \right]$.[45] The value $p_{B,N}$ represents the fraction of the bootstrapped estimates that do not have the same sign as the original estimate. If $\hat{\theta} \geq 0$, note that the type-2 p-value measures the probability mass under the bootstrap distribution below zero, not the right tail of the null distribution above the estimate. If $\hat{\theta} < 0$, the type-2 p-value measures the probability mass under the bootstrap distribution above zero, not the right tail of the null distribution below the estimate. Thus, the type-2 p-value is not a p-value in the classical sense. However, this statistic is useful because of its connection to Efron's (1979a; 1979b; 1981) percentile method for

[44] In our application, we choose $R_B = 2500$.

[45] While we reverse the inequality in equation (15) based on the sign of $\hat{\theta}$, knowing the sign of the estimate and the associated directional p-value would be sufficient to easily compute the p-value in the reverse direction if one wishes. However, we acknowledge that reversing inequalities in our p-value definitions based on the sign of the original estimate could increase type III directional errors, which involve correctly rejecting the null hypothesis but for the wrong reason, i.e., by inferring the wrong sign for the treatment effect under the null hypothesis. Significance levels could be chosen appropriately to account for this by, for example, using half the usual nominal desired level.

constructing confidence intervals: the type-2 p-value is also the "aimed coverage probability of the smallest one sided confidence interval based on Efron's percentile method which includes" zero (Singh and Berk, 1994). Inverting the hypothesis test based on a general version of $p_{B,N}$ provides a percentile method-based confidence bound for the treatment effect.

Another bootstrap-based p-value is the studentized bootstrap (or percentile-t) p-value $p_{B,S}$ approximated by

$$p_{B,S} = (1 + R_B)^{-1} \left[1 + \sum_{r=1}^{R_B} \mathbb{I}\{(\theta_r^* - \hat{\theta})/\sigma_r^* \geq \hat{\theta}/\hat{\sigma}_A\} \right], \tag{16}$$

where θ_r^* and σ_r^* are the bootstrapped estimate and its associated analytic standard error for the r-th bootstrap replicate for $r \in \{1, \cdots, R_B\}$, respectively, and $\hat{\theta}/\hat{\sigma}_A$ is the original estimate divided by its analytic standard error.[46] Hall (1988) advocates this procedure because it is akin to looking up studentized tables (using the bootstrap distribution of $(\theta_r^* - \hat{\theta})/\sigma_r^*$) instead of ordinary normal tables (as done in equation (13) to compute $p_{A,A}$) for the test statistic $\hat{\theta}/\hat{\sigma}_A$, since the standard error $\hat{\sigma}_A$ is being estimated.[47] Note that we reverse the inequality in equation (16) if the estimate $\hat{\theta}$ is not positive. We henceforth use such a reversion of inequality in the definitions of the other p-values discussed next and present only the formulas for the case where $\hat{\theta}$ is positive.[48] There are other bootstrap-based inferential procedures that we do not pursue in this paper. For example, Imbens and Menzel (2018) develop a bootstrap procedure that accounts not only for sampling uncertainty but also the uncertainty resulting from the stochastic nature of the treatment assignment. In their procedure, they make use of a least favorable copula for the counterfactual outcomes, which happens to be the isotone copula for inferences regarding the average treatment effect but takes various forms for other parameters (see Heckman et al., 1997).

[46] Note that we use $\hat{\sigma}_A$ instead of $\hat{\sigma}_B$ for studentization to reduce computational burden. While it would be superior to use $\hat{\sigma}_B$ for studentization, it involves a computationally intensive double bootstrap procedure.

[47] Hall (1988) however notes that his choice of the studentized bootstrap method over other bootstrap-based approaches "is not unequivocal."

[48] While we reverse the inequality in equation (16) based on the sign of $\hat{\theta}$, knowing the sign of the estimate and the associated directional p-value would be sufficient to easily compute the p-value in the reverse direction if one wishes. However, we acknowledge that reversing inequalities in our p-value definitions based on the sign of the original estimate could increase type III directional errors (regarding the sign of the effect) under the null hypothesis. Significance levels could be chosen appropriately to account for this.

Permutation tests are often used when researchers are interested in testing whether treatment and control groups have a common outcome distribution without relying on large-sample theory. Such tests rely on the property that the treatment and control labels are exchangable within each stratum of the experiment under the null hypothesis of a common distribution. In their permutation tests, Heckman et al. (2010a) use strata defined by wave, gender, and indicator for above-median socioeconomic status, assuming that experimental labels within each stratum are exchangeable. Heckman et al. (2011) improve on the methodology of Heckman et al. (2010a) by (i) exploiting a symmetry generated by the Perry randomization protocol[49] and (ii) allowing the aforementioned strata to be further divided according to a binary variable that is only partially observed in their model.[50] We use a simplified version of the permutation tests used in these previous papers so as to compare permutation inference with our new methods. Our permutation p-value based on the nonstudentized test statistic is approximated by

$$p_{P,N} = (1 + R_P)^{-1} \left[1 + \sum_{p=1}^{R_P} \mathbb{I}\{\theta_p^* \geq \hat{\theta}\} \right], \tag{17}$$

[49]The symmetry exploited by Heckman et al. (2011) is equivalent to fact that Q_c, which represents the result of a fair coin flip to determine which of the two undesignated groups becomes the initial treatment group, is equally likely to be 0 or 1 in our model.

[50]The partially observed binary variable U_i used in Heckman et al. (2011) equals 1 if the mother of participant i was unavailable for home visits and 0 otherwise. This variable is observed only for children of non-working mothers and the children of working mothers in the final treatment group, for whom $U_i = 0$. For the children of working mothers in the control group, this variable is not observed and could be either 0 or 1. To deal with this difficulty, Heckman et al. (2011) conduct two types of permutation tests. In the first version, the authors set U_i to 0 for all children of working mothers in the final control group and conduct a permutation test accordingly. In the second version, the authors perform the following procedure: (i) randomly sample the vector of U_i from the space of possibilities; (ii) conduct a permutation test given the sampled vector of U_i and obtain the corresponding permutation p-value; and (iii) repeat steps (i) and (ii) several times and then take the maximum p-value over the set of p-values computed for the randomly sampled vectors of U_i. Our worst-case inferential methods, described later, are similar in spirit to the approach of Heckman et al. (2011), but there is a key difference between their approach and our approach. Their maximum p-value is over a discrete space of possibilities, while our worst-case p-value is over a continuous space of possibilities, enabling us to use extreme value theory to construct bounds for the worst-case p-value. The other key difference between our approaches is regarding the variable U_i. While $U_i = 0$ for non-working mothers in both papers, we do not view U_i as binary for working mothers. Consistent with our review of the randomization protocol, we assume that children of working mothers are able to participate in the program if special arrangements, such as weekend home visits, are made for them. In our model, the special arrangements have capacity constraints, so only a limited number of slots are available for such arrangements because of financial and logistical constraints. Our model is limited in the sense that it does not allow for heterogeneity among working mothers in their availability for special arrangements. We assume that the Perry staffers choose with equal probability which working mothers get special arrangements.

where R_P is the number[51] of block permutations within cohorts of eldest participant-siblings (whose treatment statuses determine that of their younger participant-siblings) and θ_p^* is the treatment effect estimate for the p-th permutation, with $p \in \{1, \ldots, R_P\}$. Chung and Romano (2013, 2016) show that the rejection probability of hypothesis tests based on nonstudentized test statistics, such as the test using $p_{P,N}$ in equation (17), can be higher than the nominal level when testing equality of means instead of equality of distributions, even when the samples are large, unless the treatment and controls groups have equal sizes or variances. They show that using studentized test statistics can, under some assumptions, make permutation tests achieve asymptotic validity when the distributions are not assumed to be equal while also retaining the exact level under the assumption of equality of distributions. Based on their findings, we also generate permutation p-value based on the studentized test statistic using the following approximation:

$$p_{P,S} = (1 + R_P)^{-1} \left[1 + \sum_{p=1}^{R_P} \mathbb{I}\{\theta_p^*/\sigma_p^* \geq \hat{\theta}/\hat{\sigma}_A\} \right], \tag{18}$$

where θ_p^* and σ_p^* are the estimate and its associated analytic standard error for the p-th permutation, where $p \in \{1, \cdots, R_P\}$, respectively, and $\hat{\theta}/\hat{\sigma}_A$ is the original estimate divided by its analytic standard error.

5.2.2 Worst-Case Approximate Randomization Tests and Design-Specific Inference

We now discuss our worst-case approximate randomization tests. Fisher's sharp null hypothesis $\mathcal{H}_\mathcal{F}$ specifies all the counterfactual outcomes, which are imputed according to the hypothesis using the observed data. If we knew the exact randomization protocol of the Perry experiment, we could use the actual randomization as the "reasoned basis" for inference and as "the physical basis of the validity of the test" (Fisher, 1935). In other words, we could measure where the observed test statistic falls along its exact randomization distribution, i.e., the distribution of the test statistic over all possible treatment status vectors that could have been hypothetically generated by the random-

[51] We use $R_P = 2500$ in our paper.

ization protocol. The more extreme the observed test statistic falls along the null distribution, the more incompatible the observed data would be with the sharp null hypothesis. However, we do not know the exact randomization protocol: even in our stylized model of the randomization protocol, the satisficing thresholds and capacity constraints are only partially identified. To account for this ambiguity, we could in theory conduct randomization tests[52] over the set of all possible randomization protocols. Thus, we could conduct worst-case approximate randomization tests using the least favorable distribution among all the possible randomization distributions. This results in the following upper bound for the worst-case *p*-value:

$$p_w = \sup_{\gamma \in \Xi} \mathbb{P}_{\Lambda_\gamma}\{T(\tilde{D}_\gamma) \geq T(D)\}, \tag{19}$$

where $\gamma = (\delta_0, \ldots, \delta_4, \eta_{0,1}, \eta_2, \eta_3, \eta_4)$ is the vector of satisficing thresholds and capacity constraints, Ξ is the partially identified set of γ, $\mathbb{P}_{\Lambda_\gamma}$ represents probability under the probability space Λ_γ of randomizations generated by the protocol parameterized by γ, $\tilde{D}_\gamma$ represents a random realization of the treatment status vector generated by the probability space Λ_γ, D denotes the observed treatment status vector, and $T(\cdot)$ is the chosen estimator of the test statistic such that $T(\cdot)$ maps a treatment status vector to a real number measuring the magnitude of the outcome difference between the treatment and control groups. The sample space Ω_γ of the probability space Λ_γ generating the treatment status vector is given by

$$\Omega_\gamma = \left(\bigtimes_{c=0}^{4} \mathbb{U}_c(\delta_c) \right) \times \Omega_{Q,V_\gamma}, \tag{20}$$

where $\bigtimes_{c=0}^{4} \mathbb{U}_c(\delta_c)$ is the Cartesian product of the sets of admissible partitions of $\mathcal{S}_c$ (in the initial step of the protocol) across all cohorts $c \in \{0, \ldots, 4\}$, and Ω_{Q,V_γ} is the Cartesian product of the sample spaces for all other random variables, characterizing the outcomes Q_c of the coin flips and vectors of variables $V_{i,c}$ used for determining which children of working mothers are transferred

[52] These tests are approximate because our model simplifies the actual randomization procedure and can at best be considered a useful approximation of the true model of the protocol.

from the treatment to control group in the last step across all cohorts $c \in \{0, \ldots, 4\}$, used in the randomization protocol parameterized by γ.[53]

There are two challenges in computing the worst-case p-value. First, approximating the probability $\mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_\gamma) \geq T(D)\}$ for a given value $\gamma^* \in \Xi$ is computationally demanding. Second, estimating or bounding p_w based on such tail probability estimates for a finite number of points on Ξ is also challenging. We tackle these two challenges sequentially.

Approximating Tail Probabilities of Randomization Distributions The first challenge is to approximate $\mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_{\gamma^*}) \geq T(D)\}$ for a given value $\gamma^* = (\delta_0^*, \ldots, \delta_4^*, \eta_{0,1}^*, \eta_2^*, \eta_3^*, \eta_4^*) \in \Xi$. Our approach is to break up the sample space of Λ_{γ^*} into two parts, compute the tail probability (measuring how extreme the estimated treatment effect is in the distribution of possible treatment effects) on each of these two parts, and then use the law of total probability and Monte Carlo simulations to get the desired final result. To do so, we introduce additional notation. Let $\delta_c^\dagger$ be the lower bound of the partially identified set for the true value of the satisficing threshold δ_c for $c \in \{0, \ldots, 4\}$. Then, for any given value $\delta_c^* \geq \delta_c^\dagger$, observe that

$$\mathbb{U}_c(\delta_c^*) = \mathcal{X}_c \cup \mathcal{Y}_c(\delta_c^*), \tag{21}$$

where

$$\mathcal{X}_c = \{(\mathcal{A}, \mathcal{B}) \in \mathbb{U}_c(\infty) : \tau_c^2(\mathcal{A}, \mathcal{B}) \leq \delta_c^\dagger\} \tag{22}$$

and

$$\mathcal{Y}_c(\delta_c^*) = \{(\mathcal{A}, \mathcal{B}) \in \mathbb{U}_c(\infty) : \delta_c^\dagger < \tau_c^2(\mathcal{A}, \mathcal{B}) \leq \delta_c^*\}, \tag{23}$$

for all $c \in \{0, \ldots, 4\}$. In other words, it is possible to use $\mathbb{U}_c(\infty)$, which is the set with an infinite satisficing threshold such that all partitions of $\mathcal{S}_c$ are acceptable, to construct $\mathbb{U}_c(\delta_c^*)$ as the union of two sets $\mathcal{X}_c$ and $\mathcal{Y}_c(\delta_c^*)$. The set $\mathcal{X}_c$ has elements with Hotelling statistics below the lower bound

[53] Specifically, $\Omega_{Q,V_\gamma} = \{0,1\}^5 \times \left(\bigtimes_{c\in\{(0,1),2,3,4\}} \bigtimes_{m=1}^{M_c} \{v \in \{0,1\}^m : ||v||_1 = \min(\eta_c, m)\}\right)$, where $M_{0,1} = \sum_{i\in\mathcal{S}_0 \bigcup \mathcal{S}_1} M_i$ and $M_c = \sum_{i\in\mathcal{S}_c} M_i$ for all $c \in \{2,3,4\}$.

$\delta_c^\dagger$ of the partially identified set for the satisficing threshold. The other set $\mathcal{Y}_c(\delta_c^*)$ has elements with Hotelling statistics between $\delta_c^\dagger$ and δ_c^*. Let $\Omega_{\gamma^*}^{\mathcal{X}} = \bigtimes_{c=0}^4 \mathcal{X}_c$ and $\Omega_{\gamma^*}^{\mathcal{Y}} = \bigtimes_{c=0}^4 \mathcal{Y}_c(\delta_c^*)$ be the Cartesian products of those sets across cohorts. Notice that

$$\Omega_{\gamma^*} = (\Omega_{\gamma^*}^{\mathcal{X}} \cup \Omega_{\gamma^*}^{\mathcal{Y}}) \times \Omega_{Q,V_{\gamma^*}} = (\Omega_{\gamma^*}^{\mathcal{X}} \times \Omega_{Q,V_{\gamma^*}}) \cup (\Omega_{\gamma^*}^{\mathcal{Y}} \times \Omega_{Q,V_{\gamma^*}}). \tag{24}$$

Let $\Lambda_{\gamma^*}^{\mathcal{X}}$ and $\Lambda_{\gamma^*}^{\mathcal{Y}}$ be the uniform probability spaces over the sample spaces $\Omega_{\gamma^*}^{\mathcal{X}} \times \Omega_{Q,V_{\gamma^*}}$ and $\Omega_{\gamma^*}^{\mathcal{Y}} \times \Omega_{Q,V_{\gamma^*}}$, respectively. In addition, let

$$x(\gamma^*) = \frac{|\,\Omega_{\gamma^*}^{\mathcal{X}} \times \Omega_{Q,V_{\gamma^*}}\,|}{|\,\Omega_{\gamma^*}\,|} = \frac{|\,\Omega_{\gamma^*}^{\mathcal{X}}\,| \cdot |\,\Omega_{Q,V_{\gamma^*}}\,|}{|\,\Omega_{\gamma^*}^{\mathcal{X}} \cup \Omega_{\gamma^*}^{\mathcal{Y}}\,| \cdot |\,\Omega_{Q,V_{\gamma^*}}\,|} = \frac{|\,\Omega_{\gamma^*}^{\mathcal{X}}\,|}{|\,\Omega_{\gamma^*}^{\mathcal{X}} \cup \Omega_{\gamma^*}^{\mathcal{Y}}\,|}, \tag{25}$$

which is the proportion of elements in the sample space Ω_{γ^*} belonging to $\Omega_{\gamma^*}^{\mathcal{X}} \times \Omega_{Q,V_{\gamma^*}}$. Then, by the law of total probability, it follows that

$$\mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_{\gamma^*}) \geq T(D)\} = x(\gamma^*) \cdot \mathbb{P}_{\Lambda_{\gamma^*}^{\mathcal{X}}}\{T(\tilde{D}_{\gamma^*}^{\mathcal{X}}) \geq T(D)\} + y(\gamma^*) \cdot \mathbb{P}_{\Lambda_{\gamma^*}^{\mathcal{Y}}}\{T(\tilde{D}_{\gamma^*}^{\mathcal{Y}}) \geq T(D)\}, \tag{26}$$

where $y(\gamma^*) = 1 - x(\gamma^*)$, $\tilde{D}_{\gamma^*}^{\mathcal{X}}$ is a random realization on the probability space $\Lambda_{\gamma^*}^{\mathcal{X}}$, and $\tilde{D}_{\gamma^*}^{\mathcal{Y}}$ is a random realization on the probability space $\Lambda_{\gamma^*}^{\mathcal{Y}}$. Since the sample spaces in the model are large, we use Monte Carlo draws from the probability spaces to stochastically approximate the tail probability $\mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_{\gamma^*}) \geq T(D)\}$.[54] However, we do not account for Monte Carlo error in these approximations, which is a limitation of our study.[55] While there are other ways to approximate the desired tail probabilities, we believe that our approach provides a feasible way to estimate $\mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_{\gamma^*}) \geq T(D)\}$ for multiple points γ^* on Ξ efficiently using a combination of rejection

[54] Specifically, we use 500,000 Monte Carlo draws from $\left(\bigtimes_{c=0}^4 \mathbb{U}_c(\infty)\right)$ to approximate $x(\gamma^*)$. We use 400 Monte Carlo draws from $\Lambda_{\gamma^*}^{\mathcal{X}}$ to approximate $\mathbb{P}_{\Lambda_{\gamma^*}^{\mathcal{X}}}\{T(\tilde{D}_{\gamma^*}^{\mathcal{X}}) \geq T(D)\}$. In addition, we use 2600 Monte Carlo draws from $\Lambda_{\gamma^\infty}^{\mathcal{Y}}$, where $\gamma^\infty = (\infty, \ldots, \infty, \eta_{0,1}^*, \eta_2^*, \eta_3^*, \eta_4^*)$, to approximate $\mathbb{P}_{\Lambda_{\gamma^*}^{\mathcal{Y}}}\{T(\tilde{D}_{\gamma^*}^{\mathcal{Y}}) \geq T(D)\}$.

[55] An ad hoc way to deal with Monte Carlo error in the approximations is to use large-sample theory to construct confidence bounds for the tail probability of interest. Those confidence bounds could be used conservatively instead of the approximations for the p-value for a given γ^*.

sampling and importance sampling schemes.

Estimating and Bounding the Worst-Case Tail Probability The second challenge is to estimate or bound the worst-case tail probability p_w. We use stochastic approximations for this purpose as well. It is computationally infeasible to compute the p-value for each of the infinite points in the partially identified set Ξ and take the maximum of those p-values. To deal with this challenge, we first write $\Xi = \bigcup_{l=1}^{L} \Xi_l$, where $\Xi_1, \ldots, \Xi_L$ are disjoint hyper-rectangles that form a partition of the set Ξ. In our application, $L = 20$, and each hyper-rectangle represents the partially identified set for $(\delta_0, \ldots, \delta_4)$ at fixed values of $(\eta_{0,1}, \eta_2, \eta_3, \eta_4)$.[56] Then, note that

$$p_w = \max\{p_w^1, \ldots, p_w^L\}, \tag{27}$$

where

$$p_w^l = \sup_{\gamma \in \Xi_l} \mathbb{P}_{\Lambda_\gamma}\{T(\tilde{D}_\gamma) \geq T(D)\} \tag{28}$$

for $l \in \{1, \ldots, L\}$. We approximate the supremum value p_w^l for each $l \in \{1, \ldots, L\}$ using the estimated p-values $p_{(1)}^l, \ldots, p_{(S)}^l$ arranged in descending order for $S = 900$ random points on the set Ξ_l.

We approximate p_w^l for each $l \in \{1, \ldots, L\}$ in three different ways. The first type of estimate $\tilde{p}_M^l$ is the worst-case maximum (max.) p-value, which is simply the maximum order statistic

$$\tilde{p}_M^l = \max_{1 \leq s \leq S} p_{(s)}^l, \tag{29}$$

since this converges almost surely to p_w^l as $S \to \infty$. The second type of estimate $\tilde{p}_R^l$ is the worst-

[56] Note that in our application, $\eta_{0,1}$, η_2, and η_3 are point-identified while η_4 is partially identified to be in the set $\{0, \ldots, 4\}$. Thus, $(\eta_{0,1}, \eta_2, \eta_3, \eta_4)$ has 5 possible values. In addition, since we do not know the gender and mother's working status at baseline for one of the 5 participants who dropped out of the study for extraneous reasons, there are 4 possible configurations for that person's gender and mother's working status. Thus, in total there are $L = 5 \times 4 = 20$ hyper-rectangles that make up Ξ.

case adjusted p-value given by

$$\tilde{p}_R^l = p_{(1)}^l + (p_{(1)}^l - p_{(2)}^l) = 2\,p_{(1)}^l - p_{(2)}^l, \tag{30}$$

which uses the difference between the first and second order statistics to provide a more conservative estimate than the worst-case maximum p-value. This is based on the work of Robson and Whitlock (1964), who show that $\tilde{p}_R^l$ is mean unbiased to the order S^{-2} and asymptotically median unbiased. The third type of worst-case p-value is what we term the worst-case de Haan p-value, which is based on de Haan's (1981) 90% asymptotic confidence bound for the true supremum based on the S randomly sampled p-values. The worst-case de Haan p-value $\tilde{p}_D^l$ is given by

$$\tilde{p}_D^l = p_{(1)}^l + (p_{(1)}^l - p_{(2)}^l) \cdot \max\{1, K_{dH}\}, \tag{31}$$

where K_{dH} is factor provided by de Haan (1981) for the 90% asymptotic confidence bound,[57] and $\max\{1, K_{dH}\}$ helps enforce monotonicity between the worst-case maximum, adjusted, and de Haan p-values, i.e., $\tilde{p}_D^l \geq \tilde{p}_R^l \geq \tilde{p}_M^l$. Finally, p_w can be approximated by the worst-case maximum value p_M given by

$$p_M = \max\{\tilde{p}_M^1, \ldots, \tilde{p}_M^L\}. \tag{32}$$

Replacing M in the above equation with R and D would provide the worst-case adjusted p-value and the worst-case de Haan p-value, respectively.

In the above discussion, the test statistic $T(\cdot)$ used to compute the worst-case tail probability is left general. There is reason to suspect that the choice of the test statistic matters, as shown in the case of permutation tests by Chung and Romano (2013). Wu and Ding (2018) show that using studentized test statistics in certain randomization tests can control type I error asymptotically under weak null hypotheses, such as Neyman's null hypothesis, while preserving finite-sample validity under sharp null hypotheses. Their theory ignores covariates and is limited to completely

[57] Specifically, $K_{dH} = [0.9^{\alpha_{dH}} - 1]^{-1}$, where $\alpha_{dH} = -\ln(\sqrt{S})/\ln[(p_{(3)} - p_{(\sqrt{S})})/(p_{(2)} - p_{(3)})]$.

randomized factorial experiments and stratified or clustered experiments. However, they conjecture that "the strategy [of using studentized test statistics to make randomization tests asymptotically robust under weak null hypotheses while retaining their finite-sample validity under sharp null hypotheses] is also applicable for experiments with general treatment assignment mechanisms" (Wu and Ding, 2018). While we do not attempt to prove their conjecture in our experimental setting, we take it seriously given their results for certain randomization tests along with Chung and Romano's (2013) results for permutation tests. Thus, we provide worst-case p-values using both the nonstudentized statistic $\hat{\theta}$, which is the treatment effect estimate, and studentized statistic $\hat{\theta}/\hat{\sigma}_A$, which is the estimate divided by its asymptotic analytic standard error. We use $p_{M,N}$, $p_{R,N}$, and $p_{D,N}$ to denote the worst-case maximum, adjusted, and de Haan p-values using nonstudentized statistics, respectively, and $p_{M,S}$, $p_{R,S}$, and $p_{D,S}$ to denote the respective worst-case p-values using studentized statistics.

In the appendix, we provide sufficient information for those interested in conducting Holm (1979) tests of their own multiple hypotheses. Let $\rho_{(1)}, \ldots, \rho_{(K)}$ be the associated single p-values arranged in ascending order. Then, the Holm p-values adjusted for the multiplicity of the testing problem are given by $\varrho_{(k)} = \max_{j \leq k} \min(1, (K - j + 1)\, \rho_{(j)})$ for $k \in \{1, \ldots, K\}$.

5.2.3 Monte Carlo Experiments Assessing Conventional and Worst-Case Methods

We assess the empirical performance of various conventional and worst-case p-values using a series of Monte Carlo experiments. First, we consider a simple null model where $Y_i^0 \sim \mathcal{N}(0, 1)$ and $\tau_i = Y_i^1 - Y_i^0 = 0$ for all $i \in \mathcal{P}$. Note that the outcome in this model has a standard normal distribution for all participants and does not depend on covariates or the treatment status. In addition, we set the non-response probability to 20%, which in this simple model is independent of the outcome, treatment status, and covariates. We generate 200 datasets using Monte Carlo simulations under this model.[58] Figure 2 shows the rejection rates of various p-values at the 10 percent significance

[58] Although the specified null model for the Monte Carlo experiment does not depend on any covariates, the construction of many of our test statistics involves covariates. Note that the structure of the dataset (covariates and the original treatment status) other than the outcome in these 200 datasets remains the same as in the original Perry dataset.

level, i.e., the fraction of p-values that are below 0.10 in the 200 simulated datasets, for the pooled sample of participants. The rejection rates at the 10 percent significance level for the various asymptotic, bootstrap, and permutation p-values are far above the nominal levels. The studentized bootstrap p-value performs the worst, despite its attractive theoretical properties with respect to second order accuracy in large-samples (Hall, 1988). In contrast, our worst-case p-values have rejection rates that are close to or below the two nominal significance levels.

While the superior performance of our worst-case p-values in these Monte Carlo experiments seems reassuring, this is not the reason we prefer our worst-case approximate randomization tests. We prefer them on a theoretical basis, since our worst-case p-values have approximate finite-sample validity if our assumptions hold. Our worst-case p-values are conservative in the sense that they are approximations of the suprema of the randomization test-based p-values. We use this justification to make inferences about each economic outcome of interest instead of using a justification that relies on a hypothetical repeated sampling perspective. Our worst-case p-values are conservative not merely because their rejection rates in the 200 Monte Carlo experiments for a particular null model are near or lower than two nominal significance levels. Nevertheless, we present results of our Monte Carlo experiments for the sake of interested readers.

Figures 3 and 4 show the performance of the various p-values in the male and female samples. The overall patterns are similar to the pattern found for Monte Carlo experiments in the pooled sample, although the worst-case p-values seem to be a bit more conservative in the female sample. Since we only use 200 replications for our Monte Carlo experiments, we are unable to make strong conclusions about subtle differences among the various worst-case p-values. However, it seems clear that our methods perform much better than the conventional methods. This is easier to observe in Figure 5, which relates to rejection rates in the pooled sample at various significance levels. There seems to be a sharp difference between the curves representing rejection rates of conventional p-values and those of worst-case p-values, although we do not distinguish between the p-values further within each category. In the first panel of Figure 5, the rejection rates of the conventional p-values are generally above the line representing twice the significance level. In

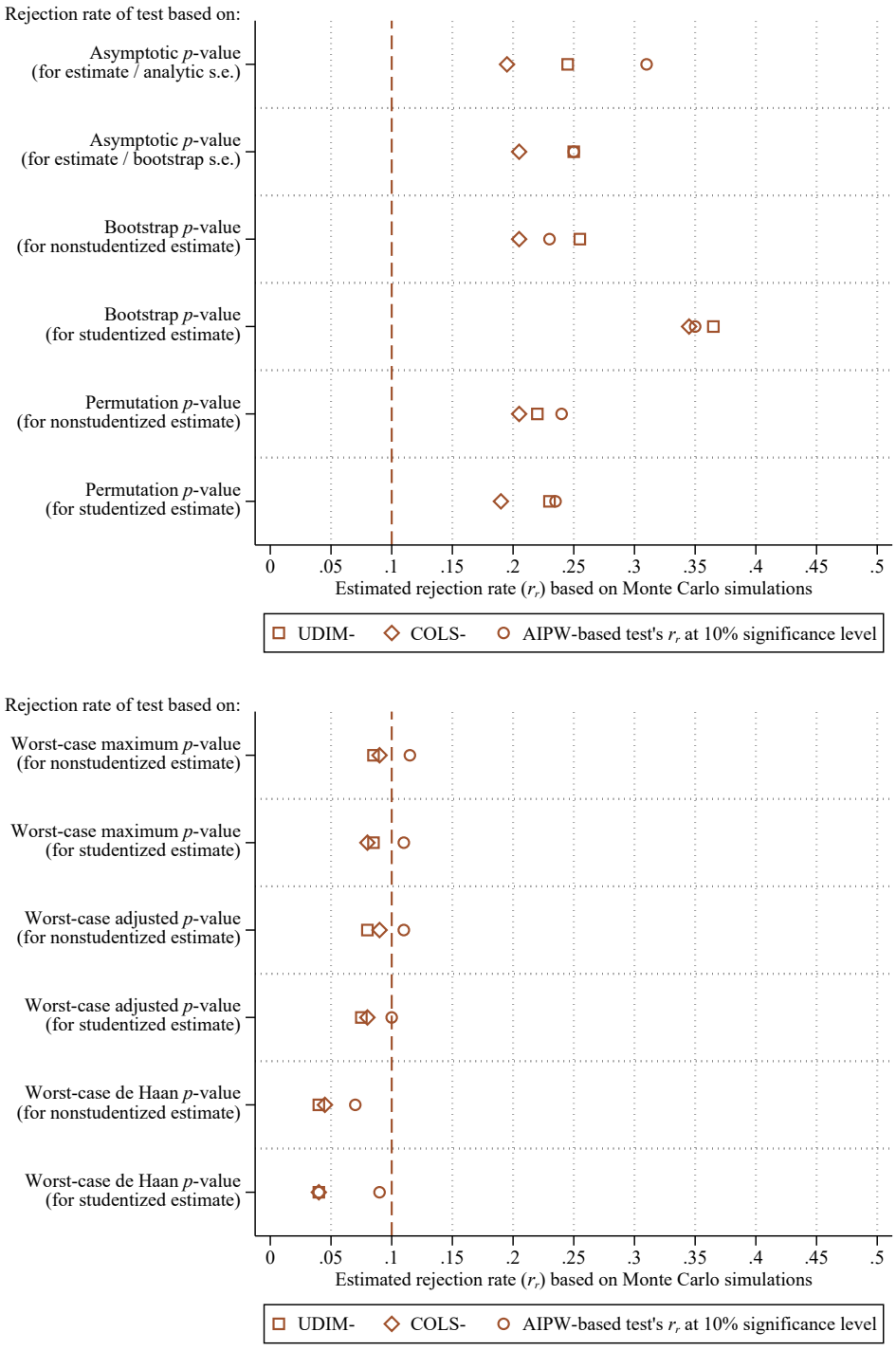
contrast, for the significance levels below 10%, rejection rates of the worst-case p-values are below the line representing a rejection rate equal to the nominal level, although this is not generally true for significance levels above 10%. We might suspect that the performance of the conventional p-values in the first panel is because of their one-sided nature. The second panel of Figure 5 shows the rejection rates of doubled p-values. Even for the doubled conventional p-values, the rejection rates are not always near the desired nominal levels. On the other hand, the rejection rates for the doubled worst-case p-values are way below their nominal levels and are overly conservative (from a repeated sampling perspective) in this particular Monte Carlo experiment. In Section 2 of the Appendix, we also report the results of Monte Carlo experiments under alternative null models. Our results for these experiments display similar patterns as before with respect to differences in the rejection rates between conventional and worst-case p-values.[59]

There is also a stark contrast between results from our worst-case inferential approach and that of Heckman et al. (2011), who also attempt to use least favorable null distributions but over a discrete set of possibilities. For example, the most stringent p-values they report for the effects on the California Achievement Test (CAT) reading, arithmetic, language, language mechanics, and spelling scores at age 14 in the male sample are 0.036, 0.086, 0.012, 0.023, 0.012, respectively. After adjusting for multiple hypothesis testing, their most stringent p-values are no more than 0.086, based on which they conclude that these effects are statistically significant. In contrast, using our approach, the worst-case maximum p-values using the nonstudentized UDIM test statistic are 0.178, 0.116, 0.077, 0.062, 0.122, respectively. Using the studentized AIPW test statistic, our worst-case maximum p-values are 0.349, 0.291, 0.177, 0.133, 0.273, respectively,[60] suggesting that the effects on the CAT scores for males are not statistically significant.

[59] Specifically, two of our alternative null models are $Y_i^0 \sim \text{Bernoulli}(0.5)$ and $Y_i^0 \sim \text{Bernoulli}(0.1)$ with $\tau_i = 0$. The other two null models we consider have treatment effect heterogeneity but zero average treatment effect, i.e., $\tau_i \sim 2 \cdot \mathcal{N}(0,1)$ and $\tau_i \sim \text{Uniform}(-4,4)$ with $Y_i^0 \sim \mathcal{N}(0,1)$. The rejection rates of worst-case p-values in the presence of treatment effect heterogeneity seem higher than those under the absence of heterogeneity. However, the rejections rates of worst-case de Haan p-values for models with treatment effect heterogeneity are by and large still near the nominal levels for significance levels at or below 10%.

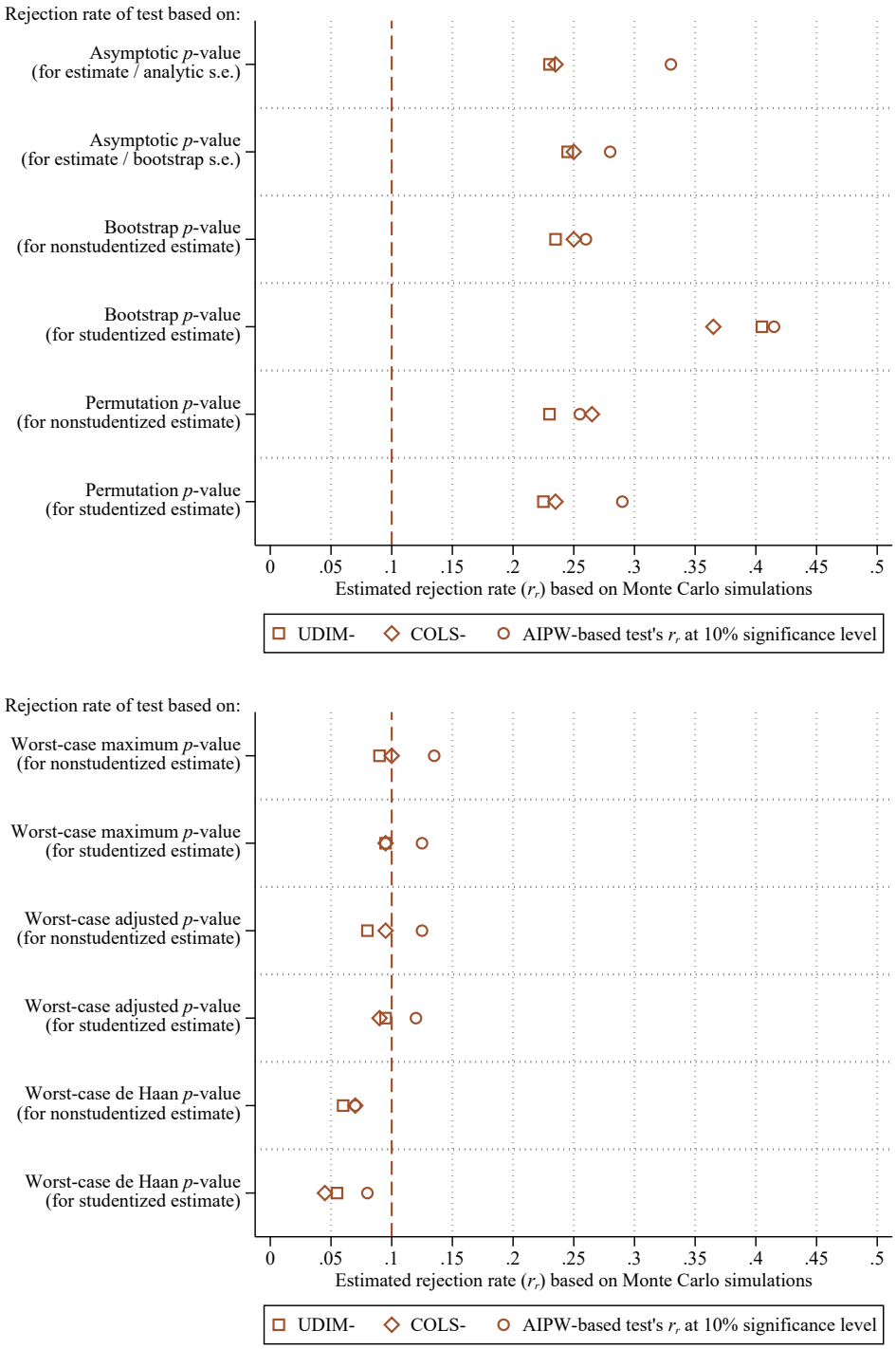
[60] The corresponding worst-case de Haan p-values are 0.382, 0.322, 0.210, 0.147, 0.302, respectively.

Figure 2: Monte Carlo-Based Rejection Rates of Various *P*-Values in the **Pooled** Sample Under the Null Hypothesis that $Y_i^0 \sim \mathcal{N}(0,1)$ and $\tau_i = Y_i^1 - Y_i^0 = 0$ with an Attrition Probability of 20%



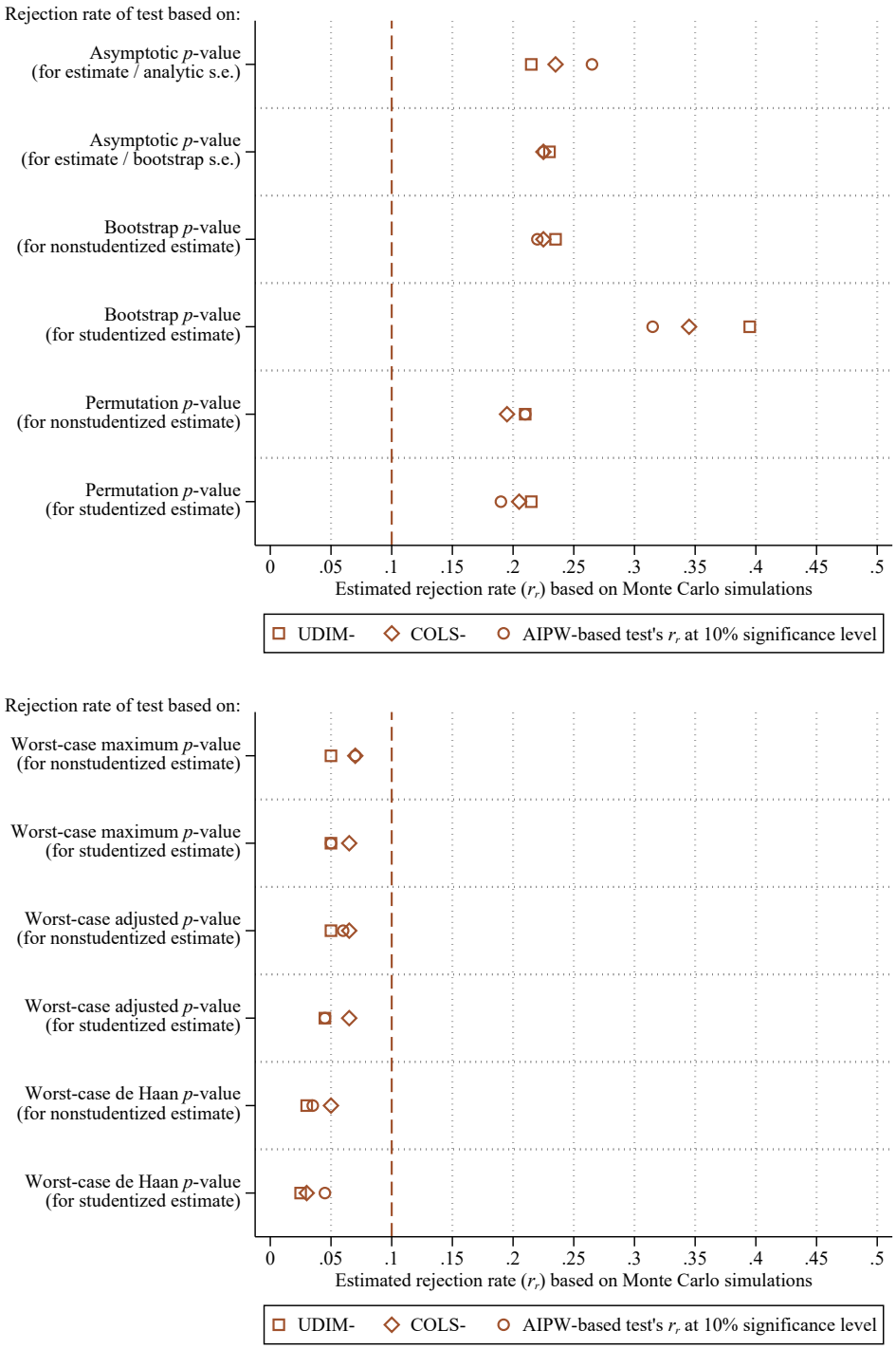
Note: This graph shows the empirical rejection rates at the 5% and 10% significance levels for various *p*-values based on 200 Monte Carlo simulations. The test statistic relevant to the *p*-value is given in parentheses. The label for each marker on the graph lists the estimate used for the test statistic and, in brackets, the significance level used for the hypothesis test.

Figure 3: Monte Carlo-Based Rejection Rates of Various *P*-Values in the **Male** Sample Under the Null Hypothesis that $Y_i^0 \sim \mathcal{N}(0,1)$ and $\tau_i = Y_i^1 - Y_i^0 = 0$ with an Attrition Probability of 20%



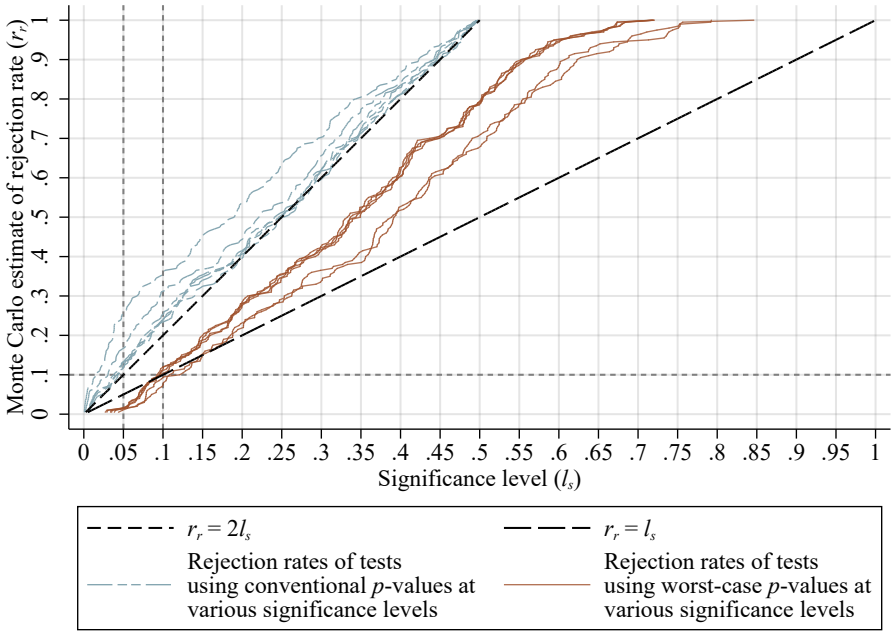
Note: This graph shows the empirical rejection rates at the 5% and 10% significance levels for various *p*-values based on 200 Monte Carlo simulations. The test statistic relevant to the *p*-value is given in parentheses. The label for each marker on the graph lists the estimate used for the test statistic and, in brackets, the significance level used for the hypothesis test.

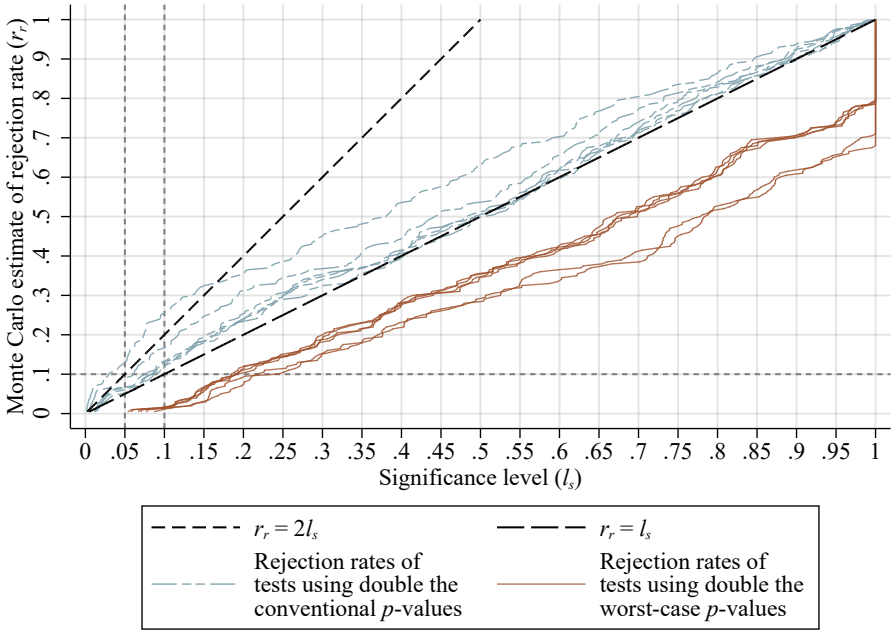
Figure 4: Monte Carlo-Based Rejection Rates of Various *P*-Values in the **Female** Sample Under the Null Hypothesis that $Y_i^0 \sim \mathcal{N}(0,1)$ and $\tau_i = Y_i^1 - Y_i^0 = 0$ with an Attrition Probability of 20%



Note: This graph shows the empirical rejection rates at the 5% and 10% significance levels for various *p*-values based on 200 Monte Carlo simulations. The test statistic relevant to the *p*-value is given in parentheses. The label for each marker on the graph lists the estimate used for the test statistic and, in brackets, the significance level used for the hypothesis test.

Figure 5: Monte Carlo-Based Rejection Rates of Various *P*-Values in the **Pooled** Sample Under the Null Hypothesis that $Y_i^0 \sim \mathcal{N}(0,1)$ and $\tau_i = Y_i^1 - Y_i^0 = 0$ with an Attrition Probability of 20%





Note: This graph shows the estimated rejection rates at various significance levels from 0 to 1 for two categories of p-values based on 200 Monte Carlo simulations. Test statistics for the p-values mentioned in this graph are based on the AIPW estimator. Conventional p-values include asymptotic p-values (using the estimate divided by analytic or bootstrap standard error as the test statistic) as well as bootstrap and permutation p-values (using nonstudentized and studentized estimates as test statistics). Worst-case p-values include the worst-case maximum, adjusted, and de Haan p-values (using nonstudentized and studentized estimates as test statistics). Since these graphs aim to contrast the two categories of p-values (conventional and worst-case), the p-values within each category are not distinguished further.

6 Results and Discussion

6.1 Crime

Using administrative data on the criminal activity of the participants, we illustrate the importance of long-term follow-up and the importance of accounting for essential features of the experimental setup. Table 1 provides estimates and measures of statistical significance of treatments effects on cumulative convictions for violent misdemeanors and felonies at ages 30, 40, and 50, and those between ages 20 and 50. For the pooled sample of participants, the AIPW estimate of the treatment effect on cumulative violent misdemeanor convictions is -0.53 at age 30 and -0.69 at age 50. Each of these effects brings the mean of the treatment group almost to zero at the respective age. The treatment effects on violent misdemeanor convictions are statistically significant at the 2.6% level (but not necessarily at lower significance levels) regardless of the method used for inference from among those discussed in the previous section.

The choice of inferential method becomes more important in analyzing treatment effects on cumulative convictions for violent felonies. At age 30, we are unable to detect statistically significant effects. At age 40, the magnitude of the treatment effect is higher at about -0.21, which represents more than a four-tenths reduction in the control mean. However, using the simple difference-in-means estimate and its conventional p-values can be misleading in this case. Using the conventional p-values, the effect at age 40 seems significant at the 10% level. However, the worst-case p-values, especially those associated with the AIPW estimate, are much higher. The worst-case de Haan p-values for the UDIM, COLS, and AIPW estimates are about 0.090, 0.154, and 0.175, respectively. Thus, if participants had not been followed up after age 40, it would have been misleading to conclude that the effects on violent felony convictions were significant. However, the long-term follow-up till late midlife has allowed us to track the criminal activity of the participants more completely. At age 50, the effect on cumulative violent felony convictions is much higher at about -0.36, representing a reduction of more than half of the control mean.

Table 1: Treatment Effects on the Crime Outcomes of the **Pooled** Participants

	Statistic or p-value	Test statistic	Cumulative violent misdemeanor convictions age 30	age 40	age 50	ages 20–50	Cumulative violent felony convictions age 30	age 40	age 50	ages 20–50
Summary	(i) Number of observations		123	120	102	102	123	120	102	102
	(ii) Mean of the control group		0.5231	0.6825	0.7200	0.6600	0.2846	0.4762	0.6400	0.5800
	(iii) Mean of the treatment group		0.0517	0.0877	0.1538	0.1538	0.1897	0.1930	0.2115	0.0962
Estimates	(iv) UDIM (difference in means)		−0.4714	−0.5948	−0.5662	−0.5062	−0.0950	−0.2832	−0.4285	−0.4838
	(v) COLS (conditional OLS estimate)		−0.5783	−0.7009	−0.7142	−0.6362	−0.0565	−0.2169	−0.3723	−0.4341
	(vi) AIPW (augmented IPW estimate)		−0.5300	−0.6491	−0.6926	−0.6167	−0.0561	−0.2052	−0.3639	−0.4188
Asymptotic p-values	(01) $p^1_{A,A}$	UDIM / Analytic s.e.	**0.0109**	**0.0033**	**0.0048**	**0.0087**	0.2301	**0.0333**	**0.0129**	**0.0018**
	(02) $p^2_{A,A}$	COLS / Analytic s.e.	**0.0097**	**0.0038**	**0.0047**	**0.0093**	0.3248	**0.0676**	**0.0206**	**0.0026**
	(03) $p^3_{A,A}$	AIPW / Analytic s.e.	**0.0064**	**0.0021**	**0.0023**	**0.0050**	0.3174	**0.0664**	**0.0126**	**0.0010**
	(04) $p^1_{A,B}$	UDIM / Bootstrap s.e.	**0.0021**	**0.0005**	**0.0044**	**0.0086**	0.2263	**0.0332**	**0.0132**	**0.0013**
	(05) $p^2_{A,B}$	COLS / Bootstrap s.e.	**0.0017**	**0.0006**	**0.0043**	**0.0090**	0.3217	**0.0708**	**0.0251**	**0.0026**
	(06) $p^3_{A,B}$	AIPW / Bootstrap s.e.	**0.0020**	**0.0010**	**0.0078**	**0.0144**	0.3217	**0.0778**	**0.0233**	**0.0023**
Bootstrap p-values	(07) $p^1_{B,N}$	Nonstudentized UDIM	**0.0004**	**0.0004**	**0.0016**	**0.0036**	0.2252	**0.0344**	**0.0144**	**0.0012**
	(08) $p^2_{B,N}$	Nonstudentized COLS	**0.0004**	**0.0004**	**0.0016**	**0.0036**	0.3156	**0.0712**	**0.0276**	**0.0016**
	(09) $p^3_{B,N}$	Nonstudentized AIPW	**0.0004**	**0.0004**	**0.0012**	**0.0028**	0.3264	**0.0868**	**0.0284**	**0.0028**
	(10) $p^1_{B,S}$	Studentized UDIM	**0.0004**	**0.0004**	**0.0004**	**0.0004**	0.1548	**0.0020**	**0.0004**	**0.0004**
	(11) $p^2_{B,S}$	Studentized COLS	**0.0004**	**0.0004**	**0.0004**	**0.0004**	0.2792	**0.0156**	**0.0012**	**0.0004**
	(12) $p^3_{B,S}$	Studentized AIPW	**0.0004**	**0.0004**	**0.0004**	**0.0008**	0.2688	**0.0196**	**0.0028**	**0.0004**
Permutation p-values	(13) $p^1_{P,N}$	Nonstudentized UDIM	**0.0036**	**0.0008**	**0.0012**	**0.0032**	0.2648	**0.0392**	**0.0104**	**0.0016**
	(14) $p^2_{P,N}$	Nonstudentized COLS	**0.0004**	**0.0004**	**0.0004**	**0.0004**	0.3604	**0.0704**	**0.0172**	**0.0020**
	(15) $p^3_{P,N}$	Nonstudentized AIPW	**0.0016**	**0.0004**	**0.0012**	**0.0020**	0.3556	**0.0792**	**0.0236**	**0.0040**
	(16) $p^1_{P,S}$	Studentized UDIM	**0.0036**	**0.0004**	**0.0036**	**0.0072**	0.2624	**0.0384**	**0.0148**	**0.0020**
	(17) $p^2_{P,S}$	Studentized COLS	**0.0028**	**0.0004**	**0.0040**	**0.0076**	0.3552	**0.0680**	**0.0216**	**0.0028**
	(18) $p^3_{P,S}$	Studentized AIPW	**0.0024**	**0.0008**	**0.0036**	**0.0080**	0.3488	**0.0708**	**0.0148**	**0.0024**
Worst-case max. p	(19) $p^1_{M,N}$	Nonstudentized UDIM	**0.0122**	**0.0051**	**0.0099**	**0.0113**	0.4086	**0.0720**	**0.0347**	**0.0124**
	(20) $p^2_{M,N}$	Nonstudentized COLS	**0.0093**	**0.0025**	**0.0051**	**0.0083**	0.4956	0.1488	**0.0438**	**0.0137**
	(21) $p^3_{M,N}$	Nonstudentized AIPW	**0.0135**	**0.0122**	**0.0099**	**0.0119**	0.4873	0.1633	**0.0586**	**0.0175**
	(22) $p^1_{M,S}$	Studentized UDIM	**0.0122**	**0.0053**	**0.0122**	**0.0151**	0.4057	**0.0708**	**0.0411**	**0.0148**
	(23) $p^2_{M,S}$	Studentized COLS	**0.0122**	**0.0103**	**0.0122**	**0.0157**	0.4922	0.1443	**0.0518**	**0.0184**
	(24) $p^3_{M,S}$	Studentized AIPW	**0.0099**	**0.0133**	**0.0134**	**0.0154**	0.4820	0.1543	**0.0473**	**0.0155**
Worst-case adjusted p	(25) $p^1_{R,N}$	Nonstudentized UDIM	**0.0128**	**0.0053**	**0.0099**	**0.0118**	0.4109	**0.0731**	**0.0361**	**0.0133**
	(26) $p^2_{R,N}$	Nonstudentized COLS	**0.0094**	**0.0025**	**0.0053**	**0.0084**	0.5025	0.1511	**0.0451**	**0.0137**
	(27) $p^3_{R,N}$	Nonstudentized AIPW	**0.0142**	**0.0128**	**0.0099**	**0.0125**	0.4907	0.1647	**0.0589**	**0.0178**
	(28) $p^1_{R,S}$	Studentized UDIM	**0.0128**	**0.0053**	**0.0128**	**0.0158**	0.4101	**0.0708**	**0.0418**	**0.0156**
	(29) $p^2_{R,S}$	Studentized COLS	**0.0128**	**0.0108**	**0.0128**	**0.0157**	0.4956	0.1458	**0.0529**	**0.0197**
	(30) $p^3_{R,S}$	Studentized AIPW	**0.0099**	**0.0143**	**0.0134**	**0.0162**	0.4847	0.1554	**0.0473**	**0.0199**
Worst-case de Haan p	(31) $p^1_{D,N}$	Nonstudentized UDIM	**0.0132**	**0.0065**	**0.0099**	**0.0161**	0.4679	0.1118	**0.0463**	**0.0152**
	(32) $p^2_{D,N}$	Nonstudentized COLS	**0.0096**	**0.0027**	**0.0067**	**0.0096**	0.5418	0.2043	**0.0702**	**0.0240**
	(33) $p^3_{D,N}$	Nonstudentized AIPW	**0.0142**	**0.0131**	**0.0107**	**0.0334**	0.5400	0.1894	**0.0829**	**0.0202**
	(34) $p^1_{D,S}$	Studentized UDIM	**0.0154**	**0.0065**	**0.0147**	**0.0199**	0.5318	**0.0900**	**0.0919**	**0.0203**
	(35) $p^2_{D,S}$	Studentized COLS	**0.0154**	**0.0154**	**0.0165**	**0.0220**	0.5839	0.1542	**0.0642**	**0.0260**
	(36) $p^3_{D,S}$	Studentized AIPW	**0.0154**	**0.0181**	**0.0179**	**0.0258**	0.5078	0.1754	**0.0535**	**0.0716**

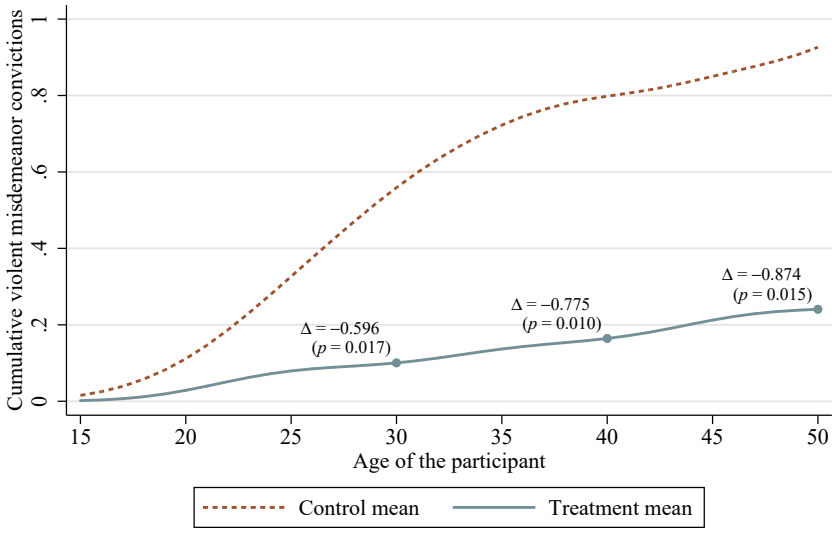
Note: Row (i) provides the number of non-missing observations for each variable. Rows (ii) and (iii) contain the means of the control and treatment groups, respectively. Rows (iv), (v), and (vi), i.e., UDIM, COLS, and AIPW, contain the unconditional difference-in-means (UDIM) estimates of treatment effects, conditional ordinary least squares (COLS) estimates (conditional on pre-program covariates, i.e., participant's IQ, SES, gender, and mother's working status at baseline), and the augmented inverse probability weighting (AIPW) estimates (accounting for non-response and imbalance in pre-program covariates between the experimental groups), respectively. Rows (01) through (36) contain various p- values. The superscripts 1, 2, and 3 of these p-values are associated with the UDIM, COLS, and AIPW estimates, respectively. Rows $(01)-(03)$ provide the one-sided asymptotic p-values based on studentized test statistics using analytic standard error, while rows $(04)-(06)$ provide those using the bootstrap standard error. Rows $(07)-(09)$ provide the bootstrap p-values based on nonstudentized test statistics, while rows $(10)-(12)$ provide those based on studentized test statistics. Rows $(13)-(15)$ provide the permutation p-values based on nonstudentized test statistics, while rows $(16)-(18)$ provide those based on studentized test statistics. Rows $(19)-(21)$ provide the worst-case maximum p-values based on nonstudentized test statistics, while rows $(22)-(24)$ provide those based on studentized test statistics. Rows $(25)-(27)$ provide the worst-case adjusted Robson-Whitlock p-values based on nonstudentized test statistics, while rows $(28)-(30)$ provide those based on studentized statistics. Rows $(31)-(33)$ provide the worst-case de Haan p-values based the nonstudentized test statistics, while rows $(34)-(36)$ provide those based on studentized test statistics.

The effects on violent felony convictions at age 50, i.e., related to cumulative crime up to age 50 as well as life-course-persistent crime after teenage years (between ages 20 and 50), are significant at the 10% level. This important example illustrates the importance of long-term follow-up and reporting an entire menu of measures of statistical significance.

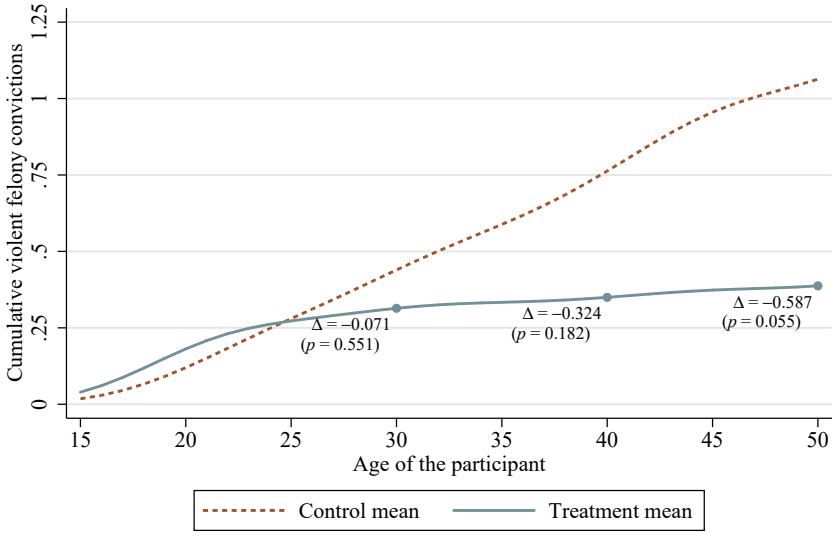
These crime effects in the pooled sample are largely made up of the impacts on male participants. Figure 6 shows the life course trajectories of cumulative criminal convictions for violent misdemeanors and felonies in the untreated and treated male samples. Similar to the pattern for the pooled sample in Table 1, the treatment effect on the cumulative violent felony convictions does not appear statistically significant in the male sample until age 50, at which point the AIPW estimate of the effect is -0.587 with a worst-case maximum p-value of 0.055. On the other hand, the effect on cumulative violent misdemeanor convictions for males, which increases from -0.596 at age 30 to -0.874 at age 50, is statistically significant at the 2% level throughout using the worst-case maximum p-value.

In addition to understanding cumulative crime outcomes, it is important to analyze effects on crime after the teenage years. Moffitt (2018) develops a developmental taxonomy that distinguishes "adolescent-limited males," who show antisocial behavior mainly during adolescence and are thought to be "common and normative," from "life-course-persistent males," who display pervasive and persistent antisocial behavior and are "hypothesized to be rare, with pathological risk factors and poor life outcomes." Table 2 shows the AIPW estimates of treatment effects and the associated p-values for selected outcomes relating to life-course-persistent crime between ages 20 and 50 for males. All of these effects, with one exception, are statistically significant at the 10% level using the worst-case maximum p-value for the studentized test statistic. In addition to this p-value, we also present four others for the sake of comparison: the asymptotic p-value for the estimate divided by the analytic standard error, and the bootstrap, permutation, and worst-case de Haan p-values for the studentized AIPW test statistic. Appendix Section 4 shows that the results in Table 2 are by and large robust to alternative estimation and inference procedures. Appendix Section 4 also reports non-significant effects on other crime outcomes we consider at various ages.

Figure 6: Cumulative Violent Criminal Convictions over the Life Course for Males



Note: Δ = augmented inverse probability weighting estimate (AIPW) of the treatment effect; p = worst-case maximum p-value based on approximate randomization test using studentized AIPW; the control and treatment means are smoothed estimates using the Gaussian kernel with bandwidth of 3.



Note: Δ = augmented inverse probability weighting estimate (AIPW) of the treatment effect; p = worst-case maximum p-value based on approximate randomization test using studentized AIPW; the control and treatment means are smoothed estimates using the Gaussian kernel with bandwidth of 3.

We find statistically significant effects on arrests for crimes classified as property, violent, and drug-related misdemeanors, especially on arrests for violent misdemeanors, committed by males. These effects on arrests also translate into effects on convictions for the misdemeanors. The AIPW estimate suggests that the treated males spend about 109 fewer days in jail on average (about a four-fifths reduction) for misdemeanors than the untreated men between ages 20 and 50. The are also

Table 2: Selected Treatment Effects on Life-Course-Persistent Crime of Male Participants

Variable	*Untreated mean*	*Treated mean*	*AIPW estimate*	*Asymptotic p-value*	*Bootstrap p-value*	*Permutation p-value*	*Worst-case max. p*	*Worst-case de Haan p*
Cumulative violent misdemeanor arrests	1.2000	0.5172	−0.9693	0.0051	0.0020	0.0076	0.0246	0.0317
Cumulative classified misdemeanor arrests	3.1000	1.6207	−1.4667	0.0197	0.0136	0.0272	0.0559	0.0773
Cumulative violent misdemeanor convictions	0.8333	0.2414	−0.7870	0.0048	0.0016	0.0084	0.0245	0.0625
Cumulative classified misdemeanor convictions	2.4667	0.9310	−1.4766	0.0021	0.0012	0.0056	0.0185	0.0252
Two or more violent misdemeanor convictions	0.2000	0.0345	−0.2293	0.0030	0.0032	0.0076	0.0250	0.0468
Two or more classified misdemeanor convictions	0.5000	0.2414	−0.2893	0.0093	0.0072	0.0160	0.0369	0.0436
Cumulative days jailed for any misdemeanors	138.57	36.103	−109.10	0.0141	0.0012	0.0072	0.0373	0.0438
Cumulative violent felony arrests	1.1917	0.3793	−0.6479	0.0125	0.0008	0.0120	0.0420	0.0830
Cumulative violent felony convictions	0.9333	0.1724	−0.6806	0.0018	0.0008	0.0044	0.0207	0.0542
Cumulative fines for violent felonies	164.54	0.0000	−142.99	0.0076	0.0008	0.0032	0.0122	0.0227
Months sentenced for violent felonies	53.306	13.931	−38.649	0.0285	0.0008	0.0548	0.1206	0.1513
One or more violent felony convictions	0.3000	0.0690	−0.1788	0.0257	0.0084	0.0240	0.0658	0.1381
Two or more violent felony arrests	0.3333	0.1034	−0.2425	0.0053	0.0020	0.0056	0.0363	0.0553
Two or more violent felony convictions	0.3000	0.0345	−0.2601	0.0009	0.0004	0.0044	0.0220	0.0299

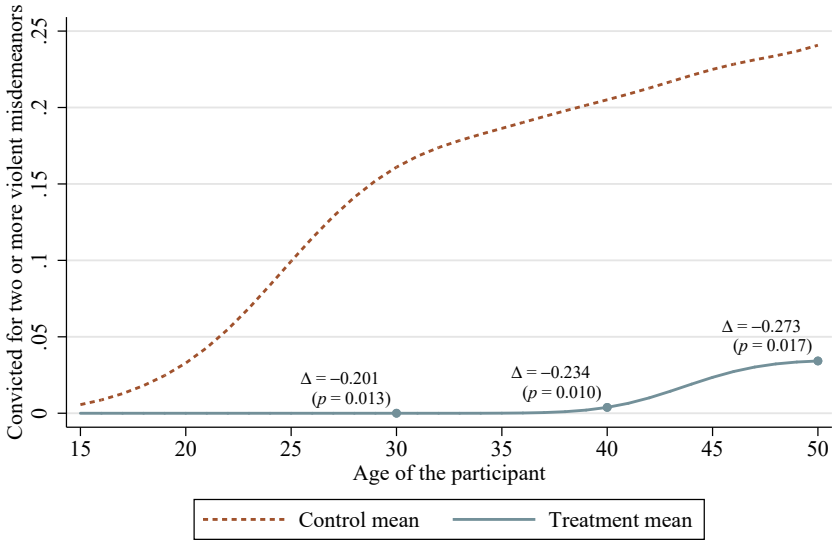
Note on the variables: The above variables relate to life-course-persisent criminal activity from ages 20 through 50. The variable *cumulative violent misdemeanor/felony arrests/convictions* refers to the cumulative number of arrests/convictions for violent misdemeanors/felonies. *Classified* misdemeanors or felonies are those classified as property, violent, or drug-related crime. *Fine* refers to fine in 2017 dollars. The variable *months sentenced for violent felonies* refers to minimum months of prison sentence for violent felonies. *One/two or more violent misdemeanor/felony arrests/convictions* refers to a binary indicator of one/two or more arrests/convictions of the specified kind.

Note on the columns: The columns labeled *untreated mean* and *treated mean* contain the means of the participants in the control and treatment groups, respectively. The column labeled *AIPW estimate* contains the augmented inverse probability weighting (AIPW) treatment effect estimates. The column labeled *asymptotic p-value* contains the corresponding one-sided asymptotic p-value based on studentized test statistic using the analytic standard error. The columns labeled *bootstrap p-value* and *permutation p-value* contain p-values based on the studentized bootstrap and permutation tests, respectively. The columns labeled *worst-case max. p* and *worst-case de Haan p* contain worst-case maximum and de Haan p-values based on approximate randomization tests using studentized test statistics, respectively.

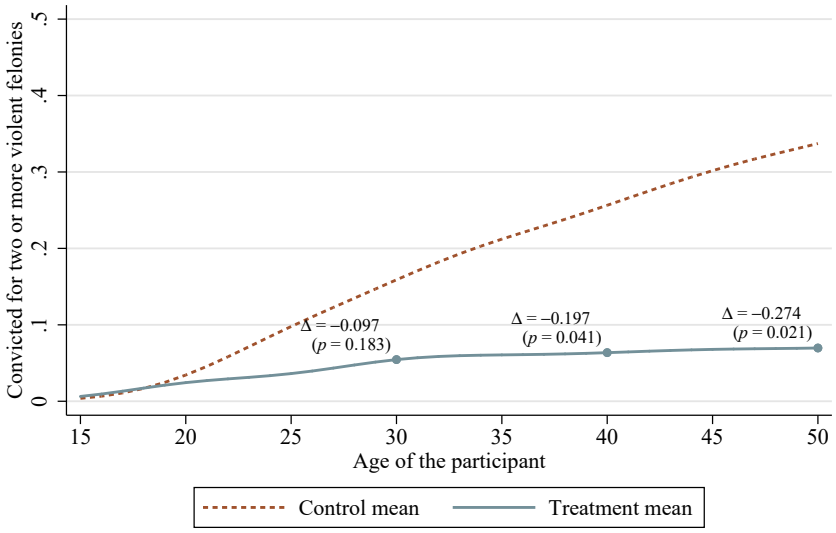
significant effects on arrests, convictions, and fines for violent felonies. The average cumulative number of convictions for violent felonies in the male control group is 0.93, whereas it is only 0.17 for treated men. The treatment effect is estimated at -0.68 after accounting for non-response and covariate imbalance, with worst-case maximum and de Haan p-values of 0.021 and 0.054, respectively. The effect on the number of months sentenced for violent felonies is not statistically significant at the 10% level but only at the 12.1% level when using the worst-case maximum p-value. Nevertheless, we report it here because it is economically significant: the treated men are

sentenced for three fewer years in prison than the untreated men, who on average are sentenced for more than four years in prison. Overall, between ages 20 and 50, 30% of the control group men have at least a conviction, whereas only about 7% of the treatment group are convicted at least once. There are also treatment effects on being convicted more than once. Figure 7 shows the fraction of male participants with two or more convictions in each group longitudinally. By age 50, about

Figure 7: Probability of Two or More Violent Criminal Convictions over the Life Course for Males



Note: Δ = augmented inverse probability weighting estimate (AIPW) of the treatment effect; p = worst-case maximum p-value based on approximate randomization test using studentized AIPW; the control and treatment means are smoothed estimates using the Gaussian kernel with bandwidth of 3.



Note: Δ = augmented inverse probability weighting estimate (AIPW) of the treatment effect; p = worst-case maximum p-value based on approximate randomization test using studentized AIPW; the control and treatment means are smoothed estimates using the Gaussian kernel with bandwidth of 3.

23% of the men in the control group have two or more convictions for violent misdemeanors, while only about 3% of the treated men have such a profile. The numbers related to violent felonies are approximately 33% and 7%, respectively. In the cases of both violent misdemeanors and felonies, the AIPW estimates of the treatment effects are sightly higher than the raw mean differences and are statistically significant at the 4% regardless of the inference method.[61]

Appendix Section 4 shows that most of the effects on men for outcomes listed in Table 2 are detectable and statistically significant even when the outcome is measured cumulatively through ages 30, 40, and 50, instead of considering only the life-course-persistent criminal activity of men between ages 20 and 50. However, this is not the case for women. Table 3 reports the statistically

Table 3: Treatment Effects on Selected Crime Outcomes of Female Participants at Age Forty

Variable	*Untreated mean*	*Treated mean*	*AIPW estimate*	*Asymptotic p-value*	*Bootstrap p-value*	*Permutation p-value*	*Worst-case max. p*	*Worst-case de Haan p*
Cumulative violent misdemeanor arrests	0.6000	0.0400	−0.5281	0.0330	0.0044	0.0300	0.0847	0.1312
Cumulative violent misdemeanor convictions	0.4400	0.0000	−0.4713	0.0543	0.0096	0.0568	0.0684	0.1429
One or more violent misdemeanor arrests	0.2400	0.0400	−0.1591	0.0161	0.0076	0.0300	0.1060	0.1144
One or more violent misdemeanor convictions	0.1200	0.0000	−0.1284	0.0274	0.0044	0.0704	0.0906	0.1123
Two or more violent misdemeanor arrests	0.1600	0.0000	−0.1545	0.0128	0.0020	0.0388	0.0754	0.1063
Two or more violent misdemeanor convictions	0.1200	0.0000	−0.1284	0.0274	0.0044	0.0704	0.0906	0.1123
Cumulative classified felony arrests	0.4000	0.0400	−0.3255	0.0310	0.0024	0.0468	0.0746	0.0907
Cumulative felony arrests of any kind	0.4800	0.0400	−0.4064	0.0302	0.0024	0.0444	0.0638	0.1117
One or more classified felony arrests	0.2400	0.0400	−0.1830	0.0176	0.0036	0.0376	0.0732	0.0829
One or more felony arrests of any kind	0.2400	0.0400	−0.1830	0.0176	0.0036	0.0376	0.0732	0.0829

Note on the variables: The above variables relate to criminal activity through age 40. The variable *cumulative violent misdemeanor/felony arrests/convictions* refers to the cumulative number of arrests/convictions for violent misdemeanors/felonies. *Classified* misdemeanors or felonies are those classified as property, violent, or drug-related crime. *One/two or more violent misdemeanor/felony arrests/convictions* refers to a binary indicator of one/two or more arrests/convictions of the specified kind.

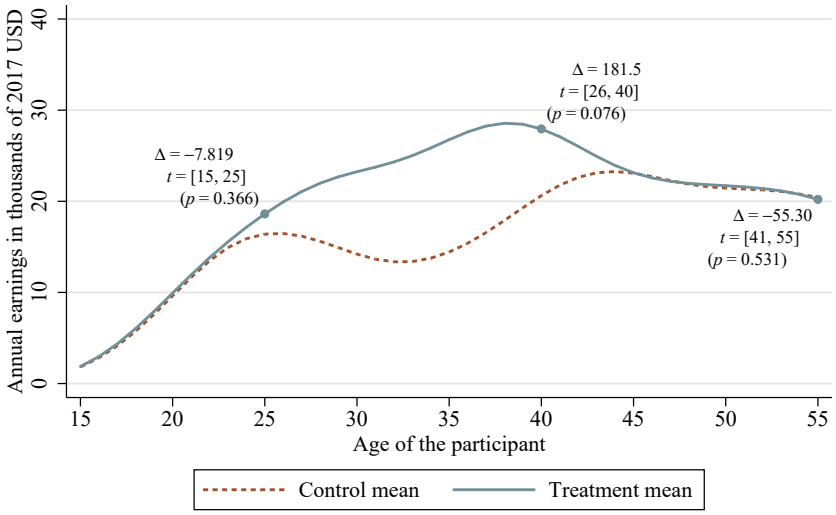
Note on the columns: The columns labeled *untreated mean* and *treated mean* contain the means of the participants in the control and treatment groups, respectively. The column labeled *AIPW estimate* contains the augmented inverse probability weighting (AIPW) treatment effect estimates. The column labeled *asymptotic p-value* contains the corresponding one-sided asymptotic p-value based on studentized test statistic using the analytic standard error. The columns labeled *bootstrap p-value* and *permutation p-value* contain p-values based on the studentized bootstrap and permutation tests, respectively. The columns labeled *worst-case max. p* and *worst-case de Haan p* contain worst-case maximum and de Haan p-values based on approximate randomization tests using studentized test statistics, respectively.

[61] These estimates do not account for the fact that the length of incarceration for the first conviction could impact the possibility of two or more convictions. However, our estimates can be treated as lower bounds (in magnitude) for the effect if we assume the following: the treated men with just one conviction are not any more prone to being convicted again in the absence of imprisonment than their control group counterparts.

significant effects on the cumulative crime of women through age 40. While the magnitudes of these effects remain similar at age 50, a drop in the female sample size from 50 observations at age 40 to 43 observations at age 50 makes it harder to ascertain the statistical significance of effects at age 50. Even at age 40, the magnitudes of the effects on women are smaller than those on men. Nevertheless, there appear to be effects on arrests and convictions for violent misdemeanors and a few other effects on felonies through age 40 in the female sample, although the conventional p-values overstate the statistical significance of these effects. For example, Heckman et al. (2013) report a p-value of 0.016 for the treatment effect on cumulative violent misdemeanor arrests at age 40. However, as reported in Table 3, the worst-case maximum and de Haan p-values for this outcome are much higher at 0.085 and 0.131, respectively.

6.2 Employment

Figure 8: Mean Annual Earnings over the Life Course for Males



Note: Δ = augmented inverse probability weighting estimate (AIPW) of the treatment effect on the cumulative annual earnings in the time period $t = [a, b]$, where a and b are starting and ending ages; p = worst-case maximum p-value based on approximate randomization test using studentized AIPW; the control and treatment means are smoothed estimates using the Gaussian kernel with bandwidth of 3.

We find that significantly fewer untreated Perry men are employed in their late twenties and thirties compared to the treated men, possibly because of incarceration of many of those untreated

Table 4: Treatment Effects on Selected Employment Outcomes of Male Participants

Variable	*Untreated mean*	*Treated mean*	*AIPW estimate*	*Asymptotic p-value*	*Bootstrap p-value*	*Permutation p-value*	*Worst-case max. p*	*Worst-case de Haan p*
Earnings in thousands (age 15 to 25)	112.65	111.00	−7.8186	0.3855	0.3211	0.3647	0.3661	0.4516
Earnings in thousands (age 26 to 30)	82.002	117.69	31.753	0.0840	0.0356	0.0972	0.2208	0.2463
Earnings in thousands (age 31 to 35)	61.346	117.63	72.551	0.0100	0.0024	0.0152	0.0742	0.1030
Earnings in thousands (age 36 to 40)	87.941	153.27	77.197	0.0082	0.0020	0.0144	0.0422	0.0855
Earnings in thousands (age 26 to 40)	231.29	388.60	181.50	0.0129	0.0016	0.0196	0.0756	0.0887
Earnings in thousands (age 41 to 55)	336.31	333.99	−55.297	0.3292	0.3439	0.3475	0.5306	0.5762
Earnings in thousands (age 15 to 55)	680.25	833.59	118.39	0.2649	0.2087	0.2859	0.3692	0.4509
Growth rate of earnings (age 26 to 30)	−0.0492	0.0961	0.2294	0.0028	0.0008	0.0020	0.0236	0.0394
Frac. of time employed (age 31 to 35)	0.3278	0.5050	0.2809	0.0046	0.0016	0.0088	0.0432	0.0613
Frac. of time employed (age 36 to 40)	0.4490	0.6134	0.2322	0.0063	0.0016	0.0120	0.0461	0.0622
Frac. of time employed (age 26 to 40)	0.4154	0.5607	0.2047	0.0065	0.0012	0.0112	0.0524	0.1285

Note on the variables: *Earnings in thousands* refers to cumulative earnings (in thousands of 2017 USD) during the specified time period. *Frac. of time employed* refers to the total fraction of time spent employed during the specified time period. *Growth rate of earnings* refers to the growth rate of average earnings during the specified time period.

Note on the columns: The columns labeled *untreated mean* and *treated mean* contain the means of the participants in the control and treatment groups, respectively. The column labeled *AIPW estimate* contains the augmented inverse probability weighting (AIPW) treatment effect estimates. The column labeled *asymptotic p-value* contains the corresponding one-sided asymptotic p-value based on studentized test statistic using the analytic standard error. The columns labeled *bootstrap p-value* and *permutation p-value* contain p-values based on the studentized bootstrap and permutation tests, respectively. The columns labeled *worst-case max. p* and *worst-case de Haan p* contain worst-case maximum and de Haan p-values based on approximate randomization tests using studentized test statistics, respectively.

men.[62] As Table 4 shows, the average fraction of time spent employed by untreated men from age 26 through age 40 is about 42%. The associated treatment effect is about 20 percentage points, with a corresponding worst-case maximum p-value of 0.052. This difference is also reflected in their earnings profiles. As shown in Figure 8, the treated men earn on average about $181,500 more cumulatively between ages 26 and 40 than the untreated men. During this period, average earnings in the male control group has a negative annual growth rate of about 5%, while the male treatment group has a positive growth rate of about 10%. However, the earnings profiles of control men and treatment men seem similar before age 25 and also after age 40.[63] In fact, the AIPW estimates of treatment effects on the cumulative earnings in these time periods are negative, although not statistically significant. This is in contrast with a treatment effect of $60,296 (in 2006 USD) that Heckman et al. (2010b) estimate for cumulative earnings between ages 41 and 65 using kernel

[62] We cannot firmly conclude as we lack the complete prison term data containing the exact dates of incarceration.

[63] In late thirties, the average earnings of the untreated men starts to increase, leading to similarity of the earnings profiles of the treatment and control men starting in their early forties. This is possibly because of re-entry of previously incarcerated control men into the workforce, although we do not have a way of confirming this. It is of note that the earnings peak of the control men is not as high as that of the treated men, which occurs in late thirties.

matching and extrapolation.[64] Because the treatment and control groups for males seem to differ in average earnings only in the late twenties and thirties but not in other time periods, the aggregated treatment effect on the cumulative earnings between ages 15 and 55 (about $118,400 in 2017 USD) is not statistically significant. For women, Table 3 presents some suggestive evidence of treatment effects on the employment rate from teenage years through age 40. However, we do not find any significant effects on earnings of the women. Appendix Section 5 reports the full set of results on employment and earnings for the pooled, male, and female samples.

Table 5: Treatment Effects on Selected Employment Outcomes of Female Participants

Variable	*Untreated mean*	*Treated mean*	*AIPW estimate*	*Asymptotic p-value*	*Bootstrap p-value*	*Permutation p-value*	*Worst-case max. p*	*Worst-case de Haan p*
Frac. of time employed (age 15 to 25)	0.1760	0.2979	0.1024	0.0322	0.0556	0.0552	0.1287	0.1955
Frac. of time employed (age 26 to 30)	0.5179	0.6709	0.0976	0.1745	0.1427	0.1747	0.2684	0.4773
Frac. of time employed (age 31 to 35)	0.3964	0.6930	0.2696	0.0125	0.0112	0.0260	0.0635	0.0876
Frac. of time employed (age 36 to 40)	0.5874	0.7660	0.1574	0.0537	0.0332	0.0720	0.1272	0.1404
Frac. of time employed (age 26 to 40)	0.5006	0.7100	0.1749	0.0391	0.0272	0.0532	0.1001	0.1149
Frac. of time employed (age 41 to 55)	0.4714	0.5545	0.0684	0.3015	0.3123	0.3179	0.3323	0.4970
Frac. of time employed (age 15 to 55)	0.4028	0.5425	0.1165	0.0662	0.0792	0.0864	0.1295	0.1601

Note on the variables: *Frac. of time employed* refers to the total fraction of time spent employed during the specified time period.

Note on the columns: The columns labeled *untreated mean* and *treated mean* contain the means of the participants in the control and treatment groups, respectively. The column labeled *AIPW estimate* contains the augmented inverse probability weighting (AIPW) treatment effect estimates. The column labeled *asymptotic p-value* contains the corresponding one-sided asymptotic p-value based on studentized test statistic using the analytic standard error. The columns labeled *bootstrap p-value* and *permutation p-value* contain p-values based on the studentized bootstrap and permutation tests, respectively. The columns labeled *worst-case max. p* and *worst-case de Haan p* contain worst-case maximum and de Haan p-values based on approximate randomization tests using studentized test statistics, respectively.

6.3 Health

The Perry participants were administered a battery of biomedical tests for the first time in the latest follow-up at around age 55. Appendix Section 6 provides a complete set of results for all the health outcomes we consider, including body fat, blood pressure, peak flow, pulse, cortisol, cholesterol, hemoglobin levels, arterial inflammation, kidney function, various diseases, smoking, alcohol consumption, drug use, eating habits, and hospitalization.

[64] Other methods of extrapolation in the supplemental material of Heckman et al. (2010b) yield treatment effect estimates as high as $231,655. Heckman et al. (2010b) extrapolated earnings beyond age 40 because their data only had earnings profiles through age 40.

Table 6: Treatment Effects on Selected Late-Midlife Health Outcomes of Male Participants

Variable	*Untreated mean*	*Treated mean*	*AIPW estimate*	*Asymptotic p-value*	*Bootstrap p-value*	*Permutation p-value*	*Worst-case max. p*	*Worst-case de Haan p*
High total cholesterol	0.9444	0.7083	−0.2906	0.0035	0.0012	0.0104	0.0414	0.0417
High C-reactive protein	0.5417	0.3462	−0.3244	0.0061	0.0156	0.0176	0.0532	0.0620
Weekly homecooking rate	4.3333	7.7241	4.0018	0.0089	0.0024	0.0132	0.0652	0.1061
Monthly bedridden rate	0.0322	0.0149	−0.0255	0.0337	0.0092	0.0584	0.0935	0.1194

Note on the variables: *High total cholesterol* indicates whether total cholesterol concentration in mg/dl is 220 or higher. *High C-reactive protein* is a binary indicator of whether the C-reactive protein in mg/L is 3 or higher. *Weekly homecooking rate* is the number of meals the participant prepares at home per week. *Monthly bedridden rate* is the percentage of days the participant was mostly in bed due to illness in the month preceding the late-midlife interview.

Note on the columns: The columns labeled *untreated mean* and *treated mean* contain the means of the participants in the control and treatment groups, respectively. The column labeled *AIPW estimate* contains the augmented inverse probability weighting (AIPW) treatment effect estimates. The column labeled *asymptotic p-value* contains the corresponding one-sided asymptotic p-value based on studentized test statistic using the analytic standard error. The columns labeled *bootstrap p-value* and *permutation p-value* contain p-values based on the studentized bootstrap and permutation tests, respectively. The columns labeled *worst-case max. p* and *worst-case de Haan p* contain worst-case maximum and de Haan p-values based on approximate randomization tests using studentized test statistics, respectively.

Table 7: Treatment Effects on Selected Late-Midlife Health Outcomes of Female Participants

Variable	*Untreated mean*	*Treated mean*	*AIPW estimate*	*Asymptotic p-value*	*Bootstrap p-value*	*Permutation p-value*	*Worst-case max. p*	*Worst-case de Haan p*
Hair cortisol	89.292	39.014	−59.278	0.0054	0.0016	0.0236	0.0405	0.0505
Regular exercise indicator	0.2500	0.4348	0.2261	0.0376	0.0252	0.0528	0.0979	0.1333
Diabetes indicator	1.0000	0.8261	−0.2229	0.0020	0.0008	0.0080	0.0473	0.0723
Substance rehabilitation indicator	0.1500	0.0000	−0.1773	0.0082	0.0012	0.0316	0.0440	0.0653
Prolonged uninsured status	0.2000	0.0435	−0.1719	0.0300	0.0104	0.0568	0.0885	0.0946

Note on the variables: *Hair cortisol* is a biomarker for chronic stress measured in in pg/mg. *Regular exercise indicator* is a binary indicator of whether the participant engages in very energetic sports or activities (e.g., gym, biking, swimming) more than once a week. *Diabetes indicator* is an indicator of whether the participant was ever diagnosed with diabetes or has high total cholesterol or high glycated hemoglobin. *Substance rehabilitation indicator* indicates whether the participant was treated for drug use or drinking since the second last interview.

Note on the columns: The columns labeled *untreated mean* and *treated mean* contain the means of the participants in the control and treatment groups, respectively. The column labeled *AIPW estimate* contains the augmented inverse probability weighting (AIPW) treatment effect estimates. The column labeled *asymptotic p-value* contains the corresponding one-sided asymptotic p-value based on studentized test statistic using the analytic standard error. The columns labeled *bootstrap p-value* and *permutation p-value* contain p-values based on the studentized bootstrap and permutation tests, respectively. The columns labeled *worst-case max. p* and *worst-case de Haan p* contain worst-case maximum and de Haan p-values based on approximate randomization tests using studentized test statistics, respectively.

Tables 6 and 7 show the significant treatment effects on the health measures for male and female participants, respectively. Treated male participants have lower incidence of dyslipidemia (excessive total cholesterol) and arterial inflammation (high C-reactive protein level) than those in the control group. Male participants in the program group cook and eat homemade food more often, and they are also less likely to be bedridden. The female treatment group has lower hair cortisol (lower long-term stress) on average than the female control group. Treated female participants have lower rates of diabetes, substance usage treatment, and prolonged uninsured status than those in the control group. The treated women are also more likely exercise regularly.

6.4 Cognitive and Noncognitive Skills

Heckman et al. (2013) suggest that a boost in childhood noncognitive skills, especially a reduction in externalizing behavior for males, mediate later life outcomes. We analyze whether the Perry program has long-lasting effects on cognitive and noncognitive skills using test measures and reports collected at the last follow-up. We construct Empirical Bayes scores of positive personality skills using self-ratings by the participants as well as ratings by external household members on the Ten Item Personality Inventory (Gosling et al., 2003). Appendix Section 8 provides details on the construction of these scores and alternative measures, in addition to reporting a full set of results on all the noncognitive outcomes considered. We estimate that the treatment improved positive personality skills by about half a standard deviation, as reported in Table 8.

Table 8: Treatment Effects on Cognitive and Non-cognitive Outcomes

Variable	*Sample*	*Untreated mean*	*Treated mean*	*AIPW estimate*	*Asymptotic p-value*	*Bootstrap p-value*	*Permutation p-value*	*Worst-case max. p*	*Worst-case de Haan p*
Positive personality	Pooled	−0.2114	0.2165	0.5231	0.0045	0.0028	0.0152	0.0444	0.0545
Executive functioning	Pooled	−0.1936	0.1869	0.3056	0.0422	0.0168	0.0504	0.0859	0.1078
Positive personality	Male	−0.1490	0.1694	0.5249	0.0296	0.0172	0.0528	0.0854	0.1418
Executive functioning	Male	−0.2385	0.2448	0.4532	0.0268	0.0092	0.0312	0.0708	0.0919
Positive personality	Female	−0.2981	0.2683	0.5206	0.0217	0.0224	0.0408	0.1740	0.2033
Executive functioning	Female	−0.1261	0.1139	0.0973	0.3410	0.2968	0.3492	0.4163	0.4596

Note on the variables: *Positive personality* refers to a positive personality Empirical Bayes score estimated using sums of reverse coded external ratings (by a household member) and self-ratings of how reserved, critical, disorganized, anxious, and conventional the participants are. *Executive functioning* refers to an executive functioning Empirical Bayes latent score estimated using general performance on Raven's and Stroop tests and also using test items with high difficulty and discrimination levels. The underlying latent variables for both of these outcomes are normalized to have mean 0 and variance 1. See Appendix for more details on the construction of these scores.

Note on the columns: The column labeled *sample* identifies the gender of the Perry participants in the subsample under consideration. *Pooled* refers to the pooled sample of male and female individuals. The columns labeled *untreated mean* and *treated mean* contain the means of the participants in the control and treatment groups, respectively. The column labeled *AIPW estimate* contains the augmented inverse probability weighting (AIPW) treatment effect estimates. The column labeled *asymptotic p-value* contains the corresponding one-sided asymptotic p-value based on studentized test statistic using the analytic standard error. The columns labeled *bootstrap p-value* and *permutation p-value* contain p-values based on the studentized bootstrap and permutation tests, respectively. The columns labeled *worst-case max. p* and *worst-case de Haan p* contain worst-case maximum and de Haan p-values based on approximate randomization tests using studentized test statistics, respectively.

We also use items on cognitive tests, specifically Raven's and Stroop tests, administered to the participants in constructing an Empirical Bayes score of executive functioning. Appendix Section 8 provides the details and also alternative cognitive measures. Although we do not find statistically significant effects on the number of correct responses on the tests, Table 8 reports evidence that the treatment group has a higher level of executive functioning than the control group when the

overall performance on Raven's and Stoop tests is taken into account. These effects seem to come primarily from the male sample. This enhanced executive functioning, together with the improved socioeconomic skills, is a potential cause of the lower male criminal activity, although we do not conduct a formal mediation analysis.

6.5 Childhood Home Environment and Parental Attachment

Table 9: Treatment Effects on Childhood Home Environment and Parental Attachment

Variable	*Sample*	*Untreated mean*	*Treated mean*	*AIPW estimate*	*Asymptotic p-value*	*Bootstrap p-value*	*Permutation p-value*	*Worst-case max. p*	*Worst-case de Haan p*
Verbally abused	Male	0.3333	0.1379	−0.2194	0.0290	0.0136	0.0424	0.0683	0.1071
Verbally abused	Female	0.2500	0.1739	−0.1268	0.1736	0.1344	0.2208	0.3248	0.4188
Felt neglected	Male	0.2333	0.0000	−0.2298	0.0017	0.0004	0.0124	0.0370	0.0393
Felt neglected	Female	0.1000	0.0435	−0.1214	0.1120	0.0824	0.1716	0.1718	0.2513
Abducted by parent	Male	0.1000	0.0000	−0.1154	0.0238	0.0008	0.0288	0.0841	0.1379
Abducted by parent	Female	0.1000	0.0435	−0.0819	0.1234	0.0784	0.1772	0.3091	0.3443
Neglected or abducted	Male	0.2333	0.0000	−0.2298	0.0017	0.0004	0.0124	0.0370	0.0393
Neglected or abducted	Female	0.1500	0.0435	−0.1515	0.0678	0.0420	0.0892	0.1684	0.1870
Attached to mother	Male	0.8966	1.0000	0.1475	0.0174	0.0488	0.0184	0.0690	0.0819
Attached to mother	Female	0.8500	0.9130	0.0291	0.3302	0.3120	0.2896	0.7361	0.7972
Attached to father	Male	0.7333	0.5357	−0.1047	0.2060	0.2016	0.2104	0.3629	0.4206
Attached to father	Female	0.5000	0.7727	0.3345	0.0081	0.0128	0.0188	0.0766	0.0827
Attached to parents	Male	0.6552	0.5357	0.0035	0.4892	0.4656	0.4744	0.5042	0.5578
Attached to parents	Female	0.4500	0.7727	0.3603	0.0047	0.0104	0.0116	0.0791	0.1012

Note on the variables: *Verbally abused* refers to a binary indicator of whether the participant was verbally abused by an adult before age 18. *Felt neglected* refers to an indicator of whether the participant felt neglected before age 18. *Abducted by parent* indicates whether the participant was abducted or hid by a parent to keep away from another parent before age 18. *Neglected or abducted* indicates whether at least one of the two previous variables equals one. *Attached to mother* and *attached to father* indicate whether the participant felt close to the biological mother and father through age 15, respectively. *Attached to parents* indicates whether the participant felt attached to both biological parents through age 15. All of these self-reports by the participants were collected at the last follow-up at around age 55.

Note on the columns: The column labeled *sample* identifies the gender of the Perry participants in the subsample under consideration. The columns labeled *untreated mean* and *treated mean* contain the means of the participants in the control and treatment groups, respectively. The column labeled *AIPW estimate* contains the augmented inverse probability weighting (AIPW) treatment effect estimates. The column labeled *asymptotic p-value* contains the corresponding one-sided asymptotic p-value based on studentized test statistic using the analytic standard error. The columns labeled *bootstrap p-value* and *permutation p-value* contain p-values based on the studentized bootstrap and permutation tests, respectively. The columns labeled *worst-case max. p* and *worst-case de Haan p* contain worst-case maximum and de Haan p-values based on approximate randomization tests using studentized test statistics, respectively.

At the last follow-up participants were asked some questions about on their childhood. This was motivated by the influential analyses of ACE (Adverse Childhood Experiences) of Felitti et al. (1998), which have been shown to be strongly associated with later life outcomes. As Table 9 shows, significantly fewer treated male participants report having been verbally abused by an adult,

feeling neglected, and having been abducted by one parent to hide from another, all before age 18, than the control group males. Relative to them, a higher fraction of the men in the treatment group also report having been close to the biological mother before age 15. Interestingly, the estimate of the treatment effect on attachment to biological father is negative among males, although it is not statistically significant. In contrast, there is a statistically significant treatment effect on the fraction of women with close attachment to both parents before age 15, primarily as a result of higher levels of attachment to the father among the treated women. These treatment effects on the relationships and attachment between the participants and their parents provide a suggestive context for understanding the lower crime rates of the treated men.

7 Conclusion

This paper reports the first comprehensive and rigorous longitudinal analysis of the treatment effects of Perry through the late midlife. We model our incomplete knowledge about the specific details of the Perry experimental design to conduct inference. We formalize our ambiguity about the randomization protocol and develop worst-case randomization tests using the least favorable randomization null distributions of test statistics. We find that the methods of estimation and inference used in many previous studies (to study participant outcomes through age 40) produce some misleading results, although a substantial number of previously reported treatment effects remain. Our framework can be applied (with appropriate modifications) to other compromised or incompletely documented randomized experiments, especially those using rerandomization designs without prespecified balancing rules.

We find long-lasting impacts of the Perry program that reduced life-course-persistent crime among males and increased their earnings during middle adulthood. We also find treatment effects on health, cognitive, and noncognitive skill measures taken after midlife. Improvements in childhood home environments and parental attachment likely play an important role as the source of the lifetime treatment effects for men who experience enhanced executive functioning. A companion

paper (Heckman and Karapakula, 2019) reports effects on the family lives of the participants, in addition to documenting inter- and intra-generational spillover effects of the Perry program. Our paper not only confirms many of the previously-reported medium-term treatment effects on the Perry Preschoolers but also finds impacts in various life domains post-midlife, thus documenting the long-term efficacy of targeted preschool programs.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2017). Sampling-based vs. design-based uncertainty in regression analysis. *arXiv preprint arXiv:1706.01778*.

Athey, S. and G. W. Imbens (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*, Volume 1, pp. 73–140. Elsevier.

Banerjee, A., S. Chassang, S. Montero, and E. Snowberg (2017). A theory of experimenters. Working Paper 23867, National Bureau of Economic Research.

Banerjee, A., S. Chassang, and E. Snowberg (2016). Decision theoretic approaches to experiment design and external validity. Working Paper 22167, National Bureau of Economic Research.

Bruhn, M. and D. McKenzie (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics 1*(4), 200–232.

Campbell, F. A., C. T. Ramey, E. Pungello, J. Sparling, and S. Miller-Johnson (2002). Early childhood education: Young adult outcomes from the Abecedarian Project. *Applied Developmental Science 6*(1), 42–57.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics 155*(2), 138–154.

Chung, E. and J. P. Romano (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics 41*(2), 484–507.

Chung, E. and J. P. Romano (2016). Multivariate and multiple permutation tests. *Journal of Econometrics 193*(1), 76–91.

Cunha, F., J. J. Heckman, L. Lochner, and D. V. Masterov (2006). Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education 1*, 697–812.

de Haan, L. (1981). Estimation of the minimum of a function using order statistics. *Journal of the American Statistical Association 76*(374), 467–469.

Efron, B. (1979a). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*(1), 1–26.

Efron, B. (1979b). Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review 21*(4), 460–480.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics 9*(2), 139–158.

Elango, S., J. L. García, J. J. Heckman, and A. Hojman (2015). Early childhood education. In *Economics of Means-Tested Transfer Programs in the United States*, Volume 2, pp. 235–297. University of Chicago Press.

Felitti, V. J., R. F. Anda, D. Nordenberg, D. F. Williamson, A. M. Spitz, V. Edwards, and J. S. Marks (1998). Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: The Adverse Childhood Experiences (ACE) study. *American Journal of Preventive Medicine 14*(4), 245–258.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver and Boyd.

Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd.

Gosling, S. D., P. J. Rentfrow, and W. B. Swann Jr (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality 37*(6), 504–528.

Greenland, S. and M. A. Mansournia (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine 34*(23), 3133–3143.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, 927–953.

Heckman, J., R. Pinto, and P. Savelyev (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review 103*(6), 2052–86.

Heckman, J. J. and G. Karapakula (2019). Intergenerational and intragenerational externalities of the Perry Preschool Program. *HCEO Working Paper 2019-033*.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Yavitz (2010a). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics 1*(1), 1–46.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Yavitz (2010b). The rate of return to the Highscope Perry Preschool Program. *Journal of Public Economics 94*(1-2), 114–128.

Heckman, J. J., R. Pinto, A. M. Shaikh, and A. Yavitz (2011). Inference with imperfect randomization: The case of the perry preschool program. Technical report, National Bureau of Economic Research.

Heckman, J. J., J. Smith, and N. Clements (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies 64*(4), 487–535.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Imbens, G. and K. Menzel (2018). A causal bootstrap. Technical report, National Bureau of Economic Research.

Kang, J. D. Y. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science 22*(4), 523–539.

Li, X. and P. Ding (2016). Exact confidence intervals for the average causal effect on a binary outcome. *Statistics in Medicine 35*(6), 957–960.

Li, X., P. Ding, and D. B. Rubin (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences 115*(37), 9157–9162.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine 23*(19), 2937–2960.

Moffitt, T. E. (2018). Male antisocial behaviour in adolescence and beyond. *Nature Human Behaviour*, 1.

Morgan, K. L. and D. B. Rubin (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics 40*(2), 1263–1282.

Morgan, K. L. and D. B. Rubin (2015). Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association 110*(512), 1412–1421.

Neyman, J. S. (1923). Próba uzasadnienia zastosowań rachunku prawdopodobieństwa do doswiadczeń polowych (On the application of probability theory to agricultural experiments: Essay on principles). *Roczniki Nauk Rolniczych (Annals of Agricultural Sciences) 10*, 1–51. Reprinted in *Statistical Science 5*(4), 465–472, as a translation by D. M. Dabrowska and T. P. Speed (1990) from section 9 (29–42) of the original Polish article.

Obama, B. (2013). The 2013 State of the Union Address. *The White House Office of the Press Secretary*.

Rigdon, J. and M. G. Hudgens (2015). Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine 34*(6), 924–935.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*(427), 846–866.

Robson, D. and J. Whitlock (1964). Estimation of a truncation point. *Biometrika 51*(1/2), 33–39.

Schweinhart, L. J. (2013). Long-term follow-up of a preschool experiment. *Journal of Experimental Criminology 9*(4), 389–409.

Schweinhart, L. J., H. V. Barnes, D. P. Weikart, W. Barnett, and A. Epstein (1993). Significant benefits: The High/Scope Perry Preschool Study through age 27 (Monographs of the High/Scope Educational Research Foundation, 10). *Ypsilanti, MI: High Scope Educational Research Foundation*.

Schweinhart, L. J., J. R. Berrueta-Clement, W. S. Barnett, A. S. Epstein, and D. P. Weikart (1985). The promise of early childhood education. *The Phi Delta Kappan 66*(8), 548–553.

Schweinhart, L. J., J. Montie, Z. Xiang, W. S. Barnett, C. R. Belfield, and M. Nores (2005). Lifetime effects: The High/Scope Perry Preschool Study through age 40 (Monographs of the High/Scope Educational Research Foundation, 14). *Ypsilanti, MI: High Scope Educational Research Foundation*.

Schweinhart, L. J. and D. P. Weikart (1980). *Young Children Grow Up: The Effects of the Perry Preschool Program on Youths Through Age 15*. Ypsilanti, MI: High Scope Educational Research Foundation.

Singh, K. and R. H. Berk (1994). A concept of type-2 p-value. *Statistica Sinica*, 493–504.

Weikart, D. P., J. T. Bond, and J. T. McNeil (1978). *The Ypsilanti Perry Preschool Project: Preschool years and longitudinal results through fourth grade*. Number 3. Ypsilanti, MI: High Scope Educational Research Foundation.

Wu, J. and P. Ding (2018). Randomization tests for weak null hypotheses. *arXiv preprint arXiv:1809.07419*.

Young, A. (2019). Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics 134*(2), 557–598.

Zigler, E. and D. P. Weikart (1993). Reply to Spitz's comments. *American Psychologist 48*(8), 915–916.