A Reanalysis of the High/Scope Perry Preschool Program

James Heckman, Seong Hyeok Moon, Rodrigo Pinto,
Peter Savelyev, and Adam Yavitz[1]

University of Chicago

April 24, 2009

[1] James Heckman is Henry Schultz Distinguished Service Professor of Economics at the University of Chicago, Professor of Science and Society, University College Dublin, Alfred Cowles Distinguished Visiting Professor, Cowles Foundation, Yale University, and Senior Fellow, American Bar Foundation. Seong Hyeok Moon, Rodrigo Pinto, and Peter Savelyev are graduate students, and Adam Yavitz is a researcher, at the University of Chicago. A version of this paper was presented at a seminar at the High/Scope Perry Foundation, Ypsilanti, Michigan, December 2006; at a conference at the Minneapolis Federal Reserve in December 2007; at a conference on the role of early life conditions at the Michigan Poverty Research Center, University of Michigan, December 2007; at a Jacobs Foundation conference at Castle Marbach, April 2008; at the Leibniz Network Conference on Noncognitive Skills in Mannheim, Germany, May 2008; and at an Institute for Research on Poverty conference, Madison, Wisconsin, June 2008. We benefited from comments received at two brown bag lunches at the Statistics Department, University of Chicago, hosted by Stephen Stigler on early drafts of this paper. We thank all workshop participants. In addition, we thank Joseph Altonji, Ricardo Barros, Dan Black, Steve Durlauf, Chris Hansman, Paul LaFontaine, Devesh Raval, Azeem Shaikh, Jeff Smith, and Steve Stigler for helpful comments. Our collaboration with Azeem Shaikh on related work greatly strengthened the analysis of this paper. This research was supported in part by the Committee for Economic Development; by a grant from the Pew Charitable Trusts and the Partnership for America's Economic Success; the JB & MK Pritzker Family Foundation; Susan Thompson Buffett Foundation; Mr. Robert Dugger; and NICHD R01HD043411. The views expressed in this presentation are those of the authors and not necessarily those of the funders listed here. Supplementary materials for this paper may be found at `http://jenni.uchicago.edu/Perry/reanalysis/`.

Abstract

This paper presents a new analysis of the influential High/Scope Perry Preschool program, an early childhood intervention in the lives of disadvantaged children with long-term followup that was evaluated by the method of random assignment. Perry provided preschool education and home visits to disadvantaged children during their preschool years. Both treatments and controls were followed from age 3 through age 40.

We develop a framework for analyzing the experiment as implemented. Previous analyses of the data assume that the planned experimental protocol was actually implemented. In fact, it was compromised. Correcting for compromised randomization, we find statistically significant and economically important program effects for both males and females. The estimated treatment effects survive adjustments for multiple-hypothesis testing and small-sample inference.

We find statistically significant treatment effects for employment, education, and criminal activity that emerge early for females and later for males. There are strong favorable treatment effects for females for educational outcomes, early employment, and other early adult-life economic outcomes, as well as for arrests. There are strong favorable treatment effects for males on a number of key outcomes, including arrests, imprisonment, earnings at age 27, employment at age 40, and other age-40 economic outcomes. We examine the external validity of the Perry experiment.

Keywords: early childhood intervention; randomization; field experiment; multiple hypothesis testing, external validity.

JEL code: I21, C93.

Contents

1 Introduction

The High/Scope Perry Preschool program, conducted in the 1960s, was an early childhood intervention that provided preschool to low-IQ, disadvantaged African-American children living in Ypsilanti, Michigan, a town near Detroit. The study was evaluated by the method of random assignment. Participants were followed through age 40. There are plans for an age-50 followup. The beneficial long-term effects reported for the Perry program constitute a cornerstone of the argument for early intervention efforts throughout the world.

Many analysts discount the reliability of the Perry study. For example, Herrnstein and Murray (1994) and Hanushek and Lindseth (2009), among others, claim that the sample size in the study is too small to make valid inferences about the program. Others express the fear that previous analyses selectively report statistically significant estimates, biasing upward the reported statistical significance of the findings (Heckman, 2005). Unnoticed in the literature is a potentially more devastating critique: the proposed randomization protocol for the Perry project was compromised. This compromise casts doubt on the validity of evaluation methods that do not account for it and calls into question the validity of simple statistical procedures applied to analyze the Perry study. In addition, there is the question of how representative the Perry population is of the general African-American population. The case for universal pre-K is often based on the Perry study, even though the project only targeted a disadvantaged segment of the population.[1]

This paper demonstrates that: (a) Statistically significant Perry treatment effects survive analyses that account for the small sample size of the study. (b) Correcting for the effect of selectively reporting statistically significant responses, there are substantial impacts of the program for both males and females. Experimental results are stronger for females at younger adult ages and for males at older adult ages. (c) Accounting for the compromised randomization of the program often *strengthens* the case for statistically significant and economically important estimated treatment effects for the Perry program as compared to effects reported in the previous literature. (d) Perry participants are representative of a low-ability, disadvantaged African-American population. (e) There is some evidence that the dynamics of the local economy in which Perry was conducted may explain gender differences by age in earnings and employment status.

We develop and apply small-sample permutation procedures that are tailored to test hypotheses for samples generated from the less-than-ideal randomization conducted in the Perry experiment. We correct estimated treatment effects for imbalances that arose in implementing the randomization protocol and from post-randomization reassignment. We address the potential problem that arises from arbitrarily selecting "significant" results from a set of possible outcomes using recently developed stepdown multiple-hypothesis testing procedures. We do multiple inference on joint hypotheses within blocks of economically interpretable

[1] See, e.g., The Pew Center on the States (2009) for one statement about the benefits of universal pre-K.

outcomes. The procedures we use minimize the probability of falsely rejecting any true null hypotheses. We test hypotheses on groups of conceptually similar outcomes measured at the same age. The methods developed in this paper are applicable to numerous real-world experiments where the randomization protocol departs from an ideal randomization procedures.[2]

This paper proceeds as follows. Section 2 describes the Perry experiment. Section 3 discusses the statistical challenges confronted in analyzing the Perry experiment. Section 4 presents our methodology. The main empirical analysis is presented in Section 5. Section 6 examines the representativeness of the Perry sample and the external validity of the experiment. Section 7 compares this study with previous studies of the Perry Preschool experiment. Section 8 discusses the key identification assumption used in this paper, and alternative approaches. Section 9 concludes. Supplementary material is provided in the Web Appendix.[3]

2 Perry: Experimental Design and Background

The High/Scope Perry program was a pre-kindergarten educational program for low-IQ African-American children. It was evaluated by the method of randomized assignment. The experiment was conducted during the early- to mid-1960s in the district of the Perry Elementary School, a public school in Ypsilanti, Michigan. The sample size is small: 123 children allocated over five entry cohorts. Data were collected at age 3, the entry age, and through annual surveys until age 15, with additional follow-ups conducted at ages 19, 27, and 40. Program attrition remains low through age 40. Numerous measures were collected on economic, criminal, and educational outcomes over this span as well as on cognition and personality. Program intensity was low compared to many subsequent early childhood development programs.[4] Beginning at age 3, and lasting two years, treatment consisted of a 2.5-hour educational preschool on weekdays during the school year, supplemented by weekly home visits by teachers.[5] High/Scope's innovative curriculum, developed over the course of the Perry experiment, was based on the Piagetian principle of active learning, guiding students through the formation of key developmental factors using open-ended questions (Schweinhart et al. 1993, pp. 34–36; Weikart et al. 1978, pp. 5–6, 21–23). A more complete description of the curriculum of the Perry program is given in Web Appendix A.

[2] This problem is pervasive in the literature. For example, in the Abecedarian program, randomization was also compromised as some initially enrolled in the experiment were later dropped (Campbell and Ramey, 1994). In the SIME-DIME experiment, the randomization protocol was never clearly described. See Kurz and Spiegelman, 1972.

[3] `http://jenni.uchicago.edu/Perry/reanalysis`

[4] For example, the Abecedarian program. (See, e.g., Campbell et al., 2002.) Cunha, Heckman, Lochner, and Masterov, 2006 and Reynolds and Temple, 2008 discuss a variety of these programs and compare their intensity.

[5] An exception is that the first entry cohort received only one year of treatment, beginning at age four.

Eligibility Criteria The program admitted five entry cohorts in the early 1960s, drawn from the population surrounding Perry Elementary School. Candidate families for the study were identified from a survey of the families of the students attending the elementary school, by neighborhood group referrals, and through door-to-door canvassing. The eligibility rules for participation were that the participants (1) be African-American; (2) have an IQ between 70 and 85 at study entry,[6] and (3) be disadvantaged as measured by parental employment level, parental education, and housing density (people/room). The Perry study targeted families who were more disadvantaged than other African-American families in the U.S. but were representative of a large segment of the disadvantaged African-American population. We discuss the issue of the external validity of the program in Section 6.

Among children in the Perry Elementary School neighborhood, Perry program families were particularly disadvantaged. Table 1 shows that compared to other families with children in the Perry School catchment area, Perry program families were younger, had lower levels of parental education, and had fewer working mothers. Further, Perry program families had fewer educational resources, larger families, and greater participation in welfare, compared to the families with children in another neighborhood elementary school in Ypsilanti (the Erickson School). Moreover, the Perry Elementary School catchment children were as a whole substantially more disadvantaged than the Erickson catchment children, who were predominantly middle-class and white.

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, Bond, and McNeil (1978, p. 16), "virtually all eligible children were enrolled in the project," so this concern appears to be of second order importance for the Perry study.

Randomization Protocol The randomization protocol used in the Perry Project was complex. Following Weikart et al. (1978, p. 16), for each designated eligible entry cohort, children were assigned to treatment and control groups in the following way, illustrated graphically in Figure 1:

1. In any entering cohort, younger siblings of previously enrolled families are assigned the same treatment status as their older siblings.[7]
2. Those remaining were ranked by their entry IQ score.[8] Odd- and even-ranked subjects were assigned to two separate groups.

[6] Measured by the Stanford-Binet IQ test (1960s norming).

[7] The rationale for excluding younger siblings from the randomization process was that enrolling children in the same family in the treatment group and others in the control group would weaken the observed treatment effect due to within-family spillovers.

[8] Ties were broken by a toss of a coin.

Table 1: Comparing Families of Participants with Other Families with Children in the Perry Elementary School Catchment, Ypsilanti, MI.

		Perry School (Overall)[a]	**Perry Preschool**[b]	**Erickson School**[c]
Mother	Average Age	35	31	32
	Mean Years of Education	10.1	9.2	12.4
	% Working	60%	20%	15%
	Mean Occupational Level[d]	1.4	1.0	2.8
	% Born in South	77%	80%	22%
	% Educated in South	53%	48%	17%
Father	% Fathers Living in the Home	63%	48%	100%
	Mean Age	40	35	35
	Mean Years of Education	9.4	8.3	13.4
	Mean Occupational Level[d]	1.6	1.1	3.3
Family & Home	Mean SES[e]	11.5	4.2	16.4
	Mean # of Children	3.9	4.5	3.1
	Mean # of Rooms	5.9	4.8	6.9
	Mean # of Others in Home	0.4	0.3	0.1
	% on Welfare	30%	58%	0%
	% Home Ownership	33%	5%	85%
	% Car Ownership	64%	39%	98%
	% Members of Library[f]	25%	10%	35%
	% with Dictionary in Home	65%	24%	91%
	% with Magazines in Home	51%	43%	86%
	% with Major Health Problems	16%	13%	9%
	% Who Had Visited a Museum	20%	2%	42%
	% Who Had Visited a Zoo	49%	26%	72%
	N	**277**	**45**	**148**

Source: Weikart, Bond, and McNeil (1978). **Notes:** (a) These are data based on parents who attended parent-teacher meetings at the Perry school or that were tracked down at their homes by Perry personnel (Weikart, Bond, and McNeil, 1978, pp. 12–15); (b) The Perry Preschool subsample consists of the full sample (treatment and control) from the first two waves; (c) The Erickson School was an "all-white school located in a middle-class residential section of the Ypsilanti public school district." (ibid., p. 14); (d) Occupation level: 1 = unskilled; 2 = semiskilled; 3 = skilled; 4 = professional; (e) See the base of Figure 3 for the definition of socio-economic status (SES) index; (f) Any member of the family.

Figure 1: Perry Randomization Protocol

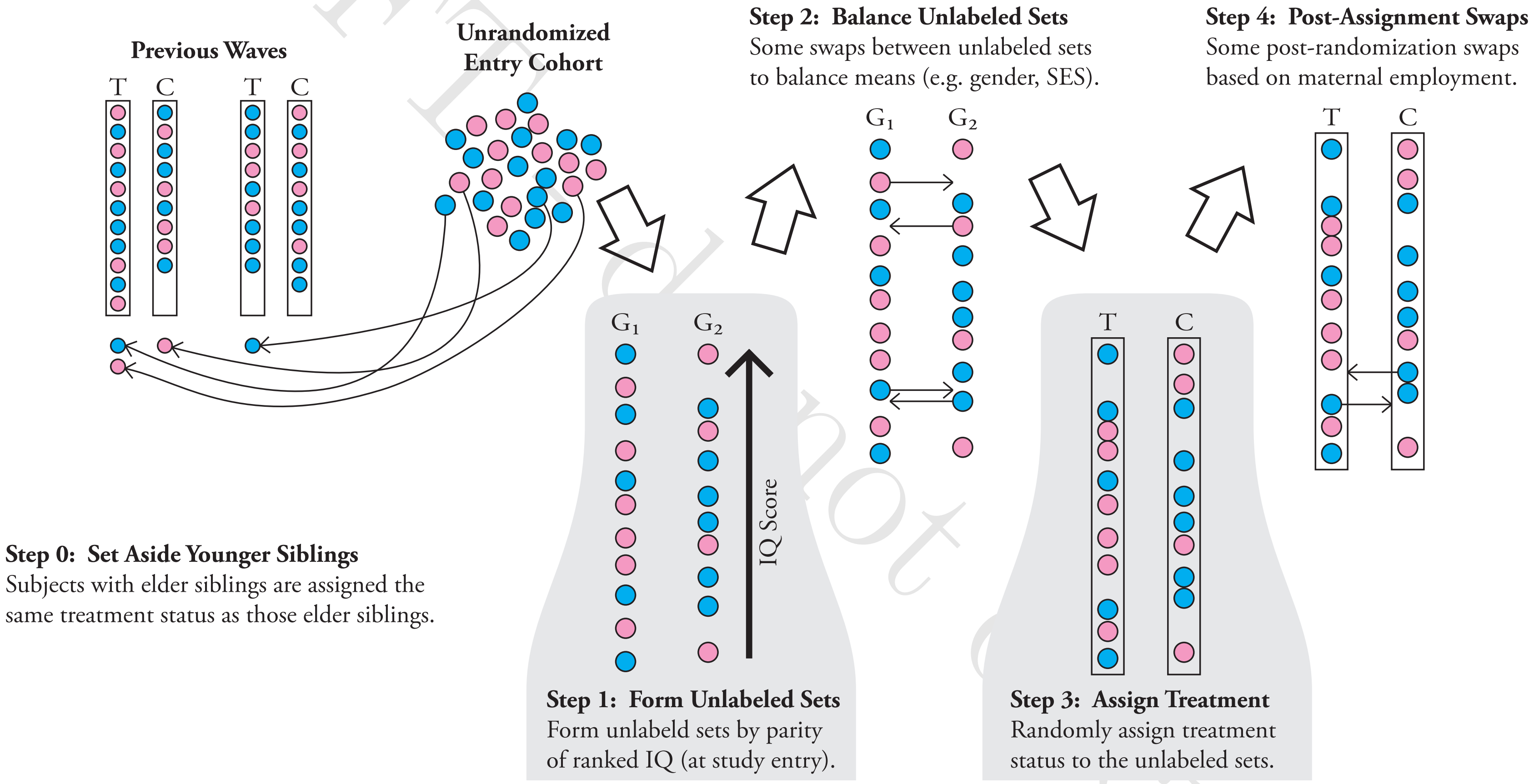

Previous Waves
T
C
T
C
Unrandomized Entry Cohort
Step 0: Set Aside Younger Siblings
Subjects with elder siblings are assigned the same treatment status as those elder siblings.
G_1
G_2
IQ Score
Step 1: Form Unlabeled Sets
Form unlabeld sets by parity of ranked IQ (at study entry).
Step 2: Balance Unlabeled Sets
Some swaps between unlabeled sets to balance means (e.g. gender, SES).
G_1
G_2
T
C
Step 3: Assign Treatment
Randomly assign treatment status to the unlabeled sets.
Step 4: Post-Assignment Swaps
Some post-randomization swaps based on maternal employment.
T
C

Figure 2: IQ at Entry by Entry Cohort and by Treatment Group

Class 1

IQ	Counts	
	Control	Treat.
88	2	1
86	1	
85		1
84		2
83		1
82	2	
80	1	1
79		1
77	1	2
76		1
73		1
71	1	
70	1	
69	3	
68	1	
67		1
66		1
63	2	
	15	13

Class 2

IQ	Counts	
	Control	Treat.
87	2	1
86	2	
85	1	
84		2
83		1
79		1
73		1
72		2
71	1	
70	1	
69	1	
64	1	
	9	8

Class 3

IQ	Counts	
	Control	Treat.
87	3	1
86	1	2
84	1	
83	1	1
82	1	1
81	1	2
80		2
79	1	1
75	1	1
73	1	1
71	1	
69	1	
68	1	
	14	12

Class 4

IQ	Counts	
	Control	Treat.
86		2
85	2	
84		2
83	3	2
82	2	1
81	1	
80	1	
79	1	1
78	2	1
77		1
76	2	
75		1
73		1
66		1
	14	13

Class 5

IQ	Counts	
	Control	Treat.
88		1
85	2	1
84	1	
83		3
82	2	
81		1
80	1	2
79	2	
78	1	1
76	2	1
75	1	1
71	1	
61		1
	13	12

Note: Stanford-Binet IQ at study entry (age 3) was used to measure baseline IQ.

Balancing on IQ produced an imbalance in family background measures. This was corrected in a second, "balancing", stage of the protocol.

3. Some individuals initially assigned to one group were swapped between the groups to balance gender and mean socio-economic (SES) score, "with Stanford-Binet scores held more or less constant."

4. A coin toss randomly selected one group as the treatment group and the other as the control group.

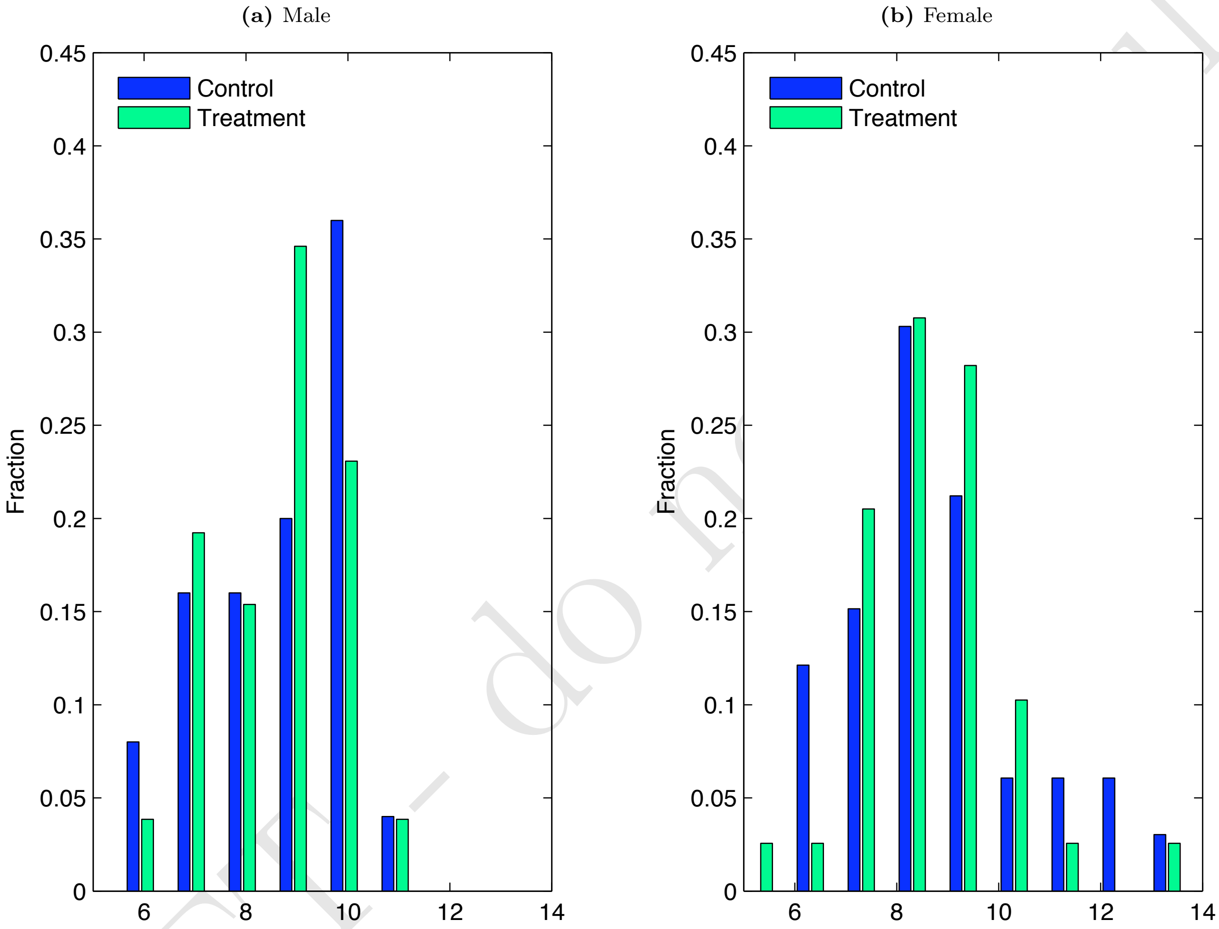
5. Some individuals provisionally assigned to treatment, whose mothers were employed at the time of the assignment, were swapped with control individuals whose mothers were not employed. The rationale for this swap was that it was difficult for working mothers to participate in home visits assigned to the treatment group.

Even after the swaps at stage 3 were made, pre-program measures were still somewhat imbalanced between treatment and control groups. See Figure 2 for IQ and Figure 3 for SES.

3 Statistical Challenges in Analyzing the Perry Program

Drawing valid inference from the Perry study requires meeting statistical challenges from three sources: small sample size, the complexity of the treatment assignment protocol actually used, and a large set of outcome measures relative to sample size.

Figure 3: SES Index, by Gender and Treatment Status



Notes: The socio-economic status (SES) index is a weighted linear combination of 3 variables: (a) average highest grade completed by whatever parent(s) were present, with coefficient 1/2; (b) father's employment status (or mother's, if the father was absent): 3 for skilled, 2 for semi-skilled, and 1 for unskilled or none, all with coefficient 2; (c) number of rooms in the home divided by number of people living in the household, with coefficient 2. The skill level of the parent's job is rated by the study coordinators and is not clearly defined. An SES index of 11 or lower was required to enter the study (Weikart, Bond, and McNeil, 1978, pp 14). This criterion was not always adhered to: out of the full sample, 7 individuals have parental SES above the cutoff. (6 out of 7 are in the treatment group, and 6 out of 7 are in the last two waves.)

Small Sample Size The small sample size of the Perry study and the non-normality of many outcome measures calls into question the validity of classical tests, such as those based on the t, F, and χ^2 statistics. Classical statistical tests rely on central limit theorems when the data are not normal and produce inferences based on p-values that are only asymptotically valid. Classical testing procedures can be unreliable when sample sizes are small and the data have non-normal distributions.[9] In the case of the Perry study, there are approximately 25 observations per gender per treatment assignment group, and the distribution of observed measures is often highly skewed.[10] Our paper addresses the problem of small sample size by using permutation-based inference. We discuss this procedure in Section 4.

The Treatment Assignment Protocol The protocol actually implemented in the Perry program was not the one initially proposed. Treatment and control status were reassigned after the initial random assignment. This reassignment creates two potential problems.

First, it can induce correlation between treatment assignment and baseline characteristics of participants. If these baseline measures affect outcomes, then treatment assignments correlate with outcomes through the induced common dependence. This relationship between outcomes and treatment assignments violates the assumption of independence between treatment assignment D and outcomes Y, even in the absence of treatment effects.

Second, even if the treatment assignment is statistically independent of the baseline variables, compromised randomization can still result in biased inference. A compromised randomization protocol can cause the distribution of treatment assignments to differ from the distribution that would result from the initially proposed randomization protocol. If this occurs, incorrect inference can result if the data are analyzed assuming that no compromise in randomization has occurred. Specifically, analyzing the Perry study assuming that a fair coin decides the treatment assignment of each participant — as if an idealized, non-compromised randomization had occurred — misspecifies the actual treatment assignment mechanism and hence the probability of assignment to treatment. This can produce incorrect critical values and improper control of Type-I error. Web Appendix C presents a Monte-Carlo study of this point. In Section 4.4, we describe how to account for the compromised randomization using permutation-based inference conditioned on baseline measures.

These potential problems are in addition to a distinct *third* problem, arising from the imbalance in the covariates between treated and controls resulting from the swaps performed at stage 3 of the randomization protocol. The imbalance is documented in Figures 2 and 3 requires conditioning on covariates to restore balance.

[9] See Micceri (1989) for a survey.

[10] Crime measures are a case in point.

Table 2: Percentage of Test Statistics Greater than Indicated Significance Level*

	All Data	Male Data Only	Female Data Only
Percentage of p-values smaller than 1%	7%	3%	7%
Percentage of p-values smaller than 5%	23%	13%	22%
Percentage of p-values smaller than 10%	34%	21%	31%

* Based on 715 outcomes in the Perry Study. (See Schweinhart et al. (2005) for a description of the data.) 269 outcomes from the period before the age-19 interview. 269 from the age-19 interview. 95 outcomes from the age-27 interview. 55 outcomes from the age-40 interview.

Multiple Outcomes The large number of outcomes available in the Perry study creates the possibility that analysts may selectively report statistically significant outcomes, without correcting for the effects of such preliminary screening. This practice is sometimes termed "cherry picking".[11] Multiple hypothesis testing procedures can avoid bias in inference arising from selectively reporting "statistically significant" results by adjusting inference to take into account the overall set of outcomes from which the statistically significant results are selected.

The following informal calculations show that this concern may be overstated for the Perry study. Table 2 summarizes the inference for 715 Perry study outcomes by reporting the percentage of hypotheses rejected at various significance levels.[12] If there was no experimental treatment effect, and outcomes were statistically independent, we would expect only 1% of the hypotheses to be rejected at the 1% level, but instead 7% overall are rejected (3% for men and 7% for women). At the 5% significance level, we obtain a 23% overall rejection rate (13% for men and 22% for women). Far more than 10% of the hypotheses are statistically significant when the 10% level is used. These results suggest that treatment effects are present both for the full sample as well as for the male and female subsamples.

The assumption of independence among the outcomes used to make these informal calculations is strong. We use modern methods for testing multiple hypotheses while accounting for possible dependence among outcomes in order to turn this suggestive analysis into sharper inference about the Perry program. In particular, we use a stepdown multiple-testing procedure that controls for the family-wise error rate (FWER) — the probability of rejecting at least one true null hypothesis among a set of hypotheses we seek to test jointly. This procedure, and its combination with the permutation-testing and conditional inference approaches above is described in Section 4.5.

[11] This issue was first raised in the context of the Perry experiment in the comments of Heckman (2005). An attempt to solve this problem is presented in Anderson (2008).

[12] Inference is based on a permutation-testing method where the t-statistic of the difference in means between treatment and control groups is used as the test statistic.

4 Methods

This section formally describes statistical techniques for inference in small experiments such as the Perry study. In particular, we account for the three problems in small-sample inference discussed in Section 3: compromised randomization, imbalance in covariates between treatments and controls, and multiple-hypothesis testing. We first review the standard model of treatment effects. We then discuss randomized experiments and the consequences of compromised randomization and covariate imbalance. Next we develop the statistical background to describe the conditions under which permutation-based inference produces valid inference for the Perry study. Finally, we discuss the multiple-hypothesis testing procedure used in this paper.

4.1 Randomized Experiments

Randomization is used to avoid selection bias. Under the null hypothesis of no treatment effect, treatment and control outcomes are independent of treatment assignment. A standard model of program evaluation describes the observed outcome for participant i, that is Y_i, by $Y_i = DY_{i,1} + (1 - D_i)Y_{i,0}$, where $(Y_{i,0}, Y_{i,1})$ are potential outcomes corresponding to treatment and control status for participant i, respectively, and D_i is the assignment indicator: $D_i = 1$ if treatment occurs, $D_i = 0$ otherwise.

An evaluation problem arises in standard observational studies because either $Y_{i,1}$ or $Y_{i,0}$ is observed, but not both. *Selection bias* can arise from participant self-selection into the treatment group. Randomized experiments attempt to eliminate this type of bias by inducing independence between $(Y_{i,0}, Y_{i,1})$ and D_i. Notationally, $(Y_0, Y_1) \perp\!\!\!\perp D$, where Y_0, Y_1, and D are vectors of the pooled variables across participants.[13] Web Appendix B discusses this point in greater detail.

Compromised randomization precludes inference under the assumption $(Y_0, Y_1) \perp\!\!\!\perp D$ (where "$\perp\!\!\!\perp$" denotes independence) and may also induce selection bias. The following statistical description of the Perry randomization protocol helps to clarify the basis for inference under complex experimental design and compromised randomization.

4.2 Randomization and Population Distributions

Denote the set of participants by $\mathcal{I} = \{1, \ldots, I\}$, where $I = 123$ for the Perry program. We denote the random variable representing treatment assignments by $D = (D_i : i \in \mathcal{I})$. The set $\mathcal{D}$ is the support of the

[13] Heckman and Smith (1995) and Heckman and Vytlacil (2007) discuss *randomization bias* and *substitution bias*. The Perry program is not subject to these biases. Randomization bias occurs when random assignment causes the type of person participating in a program to differ from the type that would participate in the program as it normally operates based on participant decisions. Substitution bias arises when members of an experimental control group gain access to close substitutes for the experimental treatment. During the pre-Head Start era of the early 1960s, there were no government alternative programs for Perry, so the problem of substitution bias is unimportant for the analysis of the Perry study.

vector of random assignments, namely $\mathcal{D} = [0,1] \times \cdots \times [0,1]$, 123 times, in short, $\mathcal{D} = [0,1]^{123}$. Assignment is produced by a randomization protocol described by a deterministic function $\mathbf{M}$. The arguments of $\mathbf{M}$ are variables which affect treatment assignment.

Define R as a random variable that describes the outcome of a randomization device (e.g., the flip of a coin in the Perry study). Prior to determining the realization of R, two groups, are formed on the basis of $\mathbf{X}$ values. Then R is determined by a flip of a coin. The distribution R does not depend on the composition of the two groups. After randomization, individuals are swapped across assigned treatment groups based on some $\mathbf{X}$ values (e.g., mother's working status). $\mathbf{M}$ captures all three aspects of the treatment assignment mechanism. More formally, $\mathbf{M}$ is a map:

$$\mathbf{M}(R, \mathbf{X}) : \text{supp}(R) \times \text{supp}(\mathbf{X}) \rightarrow \mathcal{D}. \tag{1}$$

For the Perry study, baseline variables $\mathbf{X}$ consist of data on the following measures: IQ, enrollment cohort, socio-economic status (SES) index, family structure, gender, and maternal employment status, all measured at study entry.

A consequence of randomization is that, under the protocol $\mathbf{M}$, treatment assignments with the same $\mathbf{X}$ are exchangeable random variables: they share the same treatment assignment distribution $D \mid X$.[14] By construction, R is independent of (Y_0, Y_1). Assuming that D is generated by $(\mathbf{X}, R)$ via $\mathbf{M}$, and that we observe $\mathbf{X}$, then D is independent of (Y_0, Y_1) given $\mathbf{X}$.[15] More formally, as a consequence of our assumptions about the randomization protocol and the observability of $\mathbf{X}$, we obtain the following assumption:

Assumption A-1. $(Y_1, Y_0) \perp\!\!\!\perp D \mid X$.

This assumption justifies matching as a method to correct for irregularities in the randomization protocol.

Characterizing the Distribution of Outcomes Outcome Y is generated by a function ψ:

$$Y = \psi(D, X, Z, \epsilon_Y), \tag{2}$$

where ϵ_Y denotes unobserved variables that determine Y, and Z are additional measured variables that may affect Y that are not used in the randomization protocol $\mathbf{M}$. By assumption, the Z variables are independent of D conditional on X: $Z \perp\!\!\!\perp D \mid X$. Usually, Z can be understood as a vector of baseline variables not used in $\mathbf{M}$ that operate on Y.

[14] See Appendix D for a formal discussion.

[15] Heckman, Pinto, Shaikh, and Yavitz (2009) relax the assumption that all components of $\mathbf{X}$ are observed. Components of $\mathbf{X}$ that are not observed and that partly determine $(Y_1 - Y_0)$ are a source of bias for treatment effects.

In practice, conditioning on Z can be important for controlling imbalance in variables that are not used to assign treatment but that affect outcomes. For example, birth weight (a variable not used in the Perry randomization protocol) may be low on average in the control group and high in the treatment group, and birthweight may affect outcomes. In this case an estimated treatment effect could arise in any sample due to this imbalance, and not because of the treatment itself. Such imbalance may arise from step 3 of the randomization protocol.

Matching assumption (A-1) can be written as $(Y_1(Z), Y_0(Z)) \perp\!\!\!\perp D \mid X$. One could enrich the conditioning information set by adding Z as well:

Assumption A-2. $(Y_1(Z), Y_0(Z)) \perp\!\!\!\perp D \mid X, Z$.

Assumption (A-2) departs from traditional inference for randomized experiments by using information beyond that used in the experimental design.[16]

Exchangeability The null hypothesis of no-treatment effect is equivalent to the statement that control and treated outcome distributions are the same:

Hypothesis H-1. $(Y_1 \overset{d}{=} Y_0) \mid X$,

where $\overset{d}{=}$ denotes equality in distribution. A consequence of Hypothesis (H-1) is the conditional exchangeability of observations. Let $Y = (Y_i; i \in \mathcal{I})$ be the ordered random vector of outcomes. A parallel notation for the conditioning variables is $X = (X_i; i \in \mathcal{I})$. For each element i, the vector Y can only take values $Y_{i,0}$ or $Y_{i,1}$. The outcome for participant i obeys the relationship $Y_i = D_i Y_{i,1} + (1 - D_i) Y_{i,0}$. If Hypothesis (H-1) is true, the distribution of the elements of Y that share the same value of variables X_i is the same irrespective of the treatment label. Thus, a permutation of these elements does not change the distribution of Y.[17] More precisely:

$$(Y_i; i \in \mathcal{I}) \overset{d}{=} (Y_{\pi(i)}; i \in \mathcal{I}) \tag{3}$$

and

$$\forall\, \pi : \mathcal{I} \to \mathcal{I} : \quad \text{such that } \pi \text{ is a bijection and } (\pi(i) = j) \Rightarrow (X_i = X_j). \tag{4}$$

Under Assumption (A-1), the joint distribution of (Y, D) is invariant under permutation of elements that

[16] Biased selection can occur in the context of randomized experiments if the randomization uses information that is not available to the program evaluator and is statistically dependent on the potential incomes. For example, suppose that the protocol $\mathbf{M}$ is based in part on an unobserved variable U not in R that is correlated with ϵ_Y in (2):

$$\mathbf{M}(R, \mathbf{X}, U) : supp(R) \times supp(\mathbf{X}) \times supp(U) \to \mathcal{D}. \tag{1'}$$

Under (1′), Assumption A-1 is replaced by: **Assumption A-1′.** $(Y_1, Y_0) \perp\!\!\!\perp D \mid X, U$. Heckman, Pinto, Shaikh, and Yavitz (2009) examine this case.

[17] See Appendix 4 for proof of exchangeability.

share the same pre-program variables X. Thus, from (A-1), one can augment (3) by adjoining D_i to Y_i:

$$((Y_i, D_i); i \in \mathcal{I}) \overset{d}{=} ((Y_{\pi(i)}, D_i); i \in \mathcal{I}). \tag{3'}$$

Equalities in distribution (3′) and (4) are consequences of Assumption (A-1) and Hypothesis (H-1). Together, they justify the permutation inference used in this paper.

Summarizing the discussion in this subsection, assumption (A-1) and hypothesis (H-1) imply that $Y \perp\!\!\!\perp D \mid X$, the hypothesis of no-treatment-effect we seek to test. This is demonstrated by the following argument where A_j denotes a set associated with j:

$$\begin{aligned}
\Pr((D,Y) \in (A_D, A_Y)|X) &= E(\mathbf{1}[D \in A_D] \cdot \mathbf{1}[Y \in A_Y]|X) \\
&= E(\mathbf{1}[Y \in A_Y]|D \in A_D, X) \cdot \Pr(D \in A_D|X) \\
&= E(\mathbf{1}[(Y_1 \cdot D + Y_0 \cdot (1-D)) \in A_Y]|D \in A_D, X) \cdot \Pr(D \in A_D|X) \\
&= E(\mathbf{1}[Y_0 \in A_Y]|D \in A_D, X) \cdot Pr(D \in A_D|X) \text{ by (H-1)} \\
&= E(\mathbf{1}[Y_0 \in A_Y]|X) \cdot \Pr(D \in A_D|X) \text{ by (A-1)} \\
&= \Pr(Y \in A_Y|X) \ \cdot \ \Pr(D \in A_D|X).
\end{aligned}$$

4.3 Permutation Testing Procedure

The permutation-based inference used in this paper addresses the problem posed by small sample size in a way that permits us to simultaneously account for compromised randomization when Assumptions (A-1) and (H-1) are valid.

Theoretical Basis Permutation procedures test the invariance of outcomes Y to the treatment indicators arrayed in D by using permutations that swap the positions of the elements of the outcome Y. We use the g to index permutation function π, where the permutation of elements of Y according to π_g is represented by gY. Notationally, gY is defined as:

$$gY = \left(\widetilde{Y}_i; i \in \mathcal{I} \mid \ \widetilde{Y}_i = Y_{\pi_g(i)}, \text{where } \pi_g \text{ is a permutation function (i.e., } \pi_g : \mathcal{I} \to \mathcal{I} \text{ is a bijection)}\right).$$

Our procedure tests whether $Y \perp\!\!\!\perp D \mid X$ using the *Randomization Hypothesis*:[18]

$$(Y, D) \overset{d}{=} (gY, D)|X \ \ \forall g \in \mathscr{G}. \tag{5}$$

Equality in distribution (5) is a consequence of assumption (A-1) and hypothesis (H-1). The set $\mathscr{G}$ contains all permutations g such that (5) holds. Intuitively, hypothesis (5) states that if there are no treatment effects and the randomization protocol is such that the distribution of Y is invariant over some strata of variables X, then the permutation of elements of Y within this strata does not change the joint distribution of the vectors Y and D.[19]

Advantages of Permutation-Based Inference Permutation tests involve testing a null hypothesis using permutations of the data. If the null hypothesis is true, the distribution of the data is invariant to permutations. Our procedure relies on the assumption of exchangeability of observations under the null hypothesis. Permutation-based inferences are often termed data-dependent because the computed p-values are conditioned on the observed data. These tests are also *distribution-free* because they do not rely on assumptions about the parametric distribution from which the data have been sampled. Because permutation tests give accurate p-values even when the sampling distribution is skewed, they are often used when sample sizes are small and sample statistics are unlikely to be normal. Hayes (1996) shows the advantage of permutation tests over the classical approaches for the analysis of small samples and non-normal data.

Under the *Randomization Hypothesis* statistics based on assignments D and outcomes Y are *distribution-invariant* or *exchangeable* under reassignments based on the permutations $g \in \mathscr{G}$. For example, under the null hypothesis of no treatment effect, the distribution of a statistic such as the difference in means between treatments and controls will not change if treatment status is permuted across observations according to g.

Single-Hypothesis Permutation Testing Our test compares the test statistic computed on the sample data with test statistics computed on resampled data where treatment and control labels are permuted for the outcomes in each resampling. The p-value for our test is the fraction of the statistics greater than the statistic in the original (unpermuted) data.[20] A level-α critical value for this test would be the $100 \times \alpha$ percentile of the permutation distribution.[21]

[18] See Lehmann and Romano (2005, Chapter 9).

[19] Web Appendix D discusses further aspects of our permutation methodology.

[20] For a one-sided hypothesis test where, for example, the test statistic is the treatment-control difference-in-means, the null hypothesis is no treatment effect, and the alternative is that treatment effects are positive.

[21] Web Appendix E provides a formal explanation of this general procedure.

4.4 Accounting for Compromised Randomization

In this paper, the problem of compromised randomization is solved by assuming *conditional* exchangeability of assignments given X. Thus, even though assignments might not be exchangeable across all background measures, they are assumed to be exchangeable conditional on the measures. A byproduct of our approach is the correction for imbalance in covariates between treatments and controls.

Conditional inference is implemented using a permutation-based test that relies on restricted classes of permutations, denoted by $\mathscr{G}_X$. We partition the sample into subsets, where each subset consists of participants with common background measures. Such subsets are sometimes called *orbits* or *blocks*. Under the null hypothesis of no-treatment effect, treatment and control outcomes have the same distribution within an orbit.[22] Equivalently, under the null hypothesis, treatment assignments D are exchangeable (therefore permutable) with respect to the outcome Y for participants who share common pre-program values X. Thus, the valid permutations $g \in \mathscr{G}_X$ swap labels *within* conditioning orbits.

We adapt standard permutation methods to account for the explicit Perry randomization protocol. Features of the randomization protocol, such as identical treatment assignments for siblings, generate a distribution of treatment assignments that cannot be described (or replicated) by simple random assignment.[23]

Conditional Inference in Small Samples Invoking conditional exchangeability decreases the number of valid permutations of the values of Y or D by permuting only within orbits. The small Perry sample size prohibits very fine partitions of the available conditioning variables. In general, nonparametric conditioning in small samples introduces the serious practical problem of small or even empty orbits. To circumvent this problem and obtain restricted permutation orbits of reasonable size, we assume a linear relationship between some of the baseline measures in X and outcomes Y. We partition the data based on orbits formed by measures that do not have a linear relationship with outcome measures. Removing the effects of some conditioning variables, we are left with larger subsets within which permutation-based inference is feasible.

More precisely, suppose that the data on pre-program variables X take on J distinct values, say, $\{a_1, a_2, \ldots, a_J\}$. Partition the index set $\mathcal{I}$ into J disjoint sets where each set indexed by j is defined by the participants that share the same value a_j; $j = 1, \ldots, J$ for pre-program variables X. We assume a linear relationship between Y and some X given the remaining conditioning variables.[24] Divide the vector X into two parts: those variables $X^{[L]}$ which are assumed to have a linear relationship with Y, and $X^{[P]}$, whose relationship with

[22] The baseline variables can affect outcomes, but may (or may not) affect the distribution of assignments produced by the compromised randomization.

[23] Web Appendix D provides relevant theoretical background, as well as operational details, about implementing the permutation framework.

[24] Linearity is not strictly required, but we use it in our empirical work. In place of linearity, we could use a more general parametric functional form with unknown parameters.

Y is unconstrained, so that $X = [X^{[L]}, X^{[P]}]$. We use a parallel notation for $a_j = [a_j^{[L]}, a_j^{[P]}]\,; j \in \{1, \ldots, J\}$. The relationship is assumed to be $Y \equiv h(X^{[L]}, X^{[P]}, \epsilon_Y) = \delta X^{[L]} + h(X^{[P]}, \epsilon_Y)$, where ϵ_Y is independent of X. Define $\tilde{Y} \equiv Y - \delta X^{[L]} = h(X^{[P]}, \epsilon_Y)$. Assuming that $(Y - \delta X^{[L]}) \perp\!\!\!\perp X^{[L]} \mid X^{[P]}$, and denoting the adjusted Y by $\tilde{Y} = Y - \delta X^{[L]}$, we obtain the following equalities:

$$\begin{aligned} F_{Y|X=a_j}(y) = &\quad F_{Y|X^{[L]}=a_j^{[L]},X^{[P]}=a_j^{[P]}}(y) \\ = &\quad F_{\tilde{Y}|X^{[P]}=a_j^{[P]}}(y - \delta X^{[L]}). \end{aligned}$$

By virtue of this assumption, we can purge the influence of $X^{[L]}$ on Y by subtracting $\delta X^{[L]}$ and can construct valid permutation tests of the null hypothesis of no treatment effect conditioning on $X^{[P]}$. Conditioning nonparametrically, using a smaller set of measures, we are able to create restricted permutation orbits that contain substantially larger numbers of participants than if we condition more finely. In an extreme case, one can assume that all conditioning variables enter linearly.

Conditional Permutation and Linearity Assumptions If δ were known, we could control for the effect of $X^{[L]}$ by permuting $\tilde{Y} = Y - \delta X^{[L]}$ within the groups of participants that share same pre-program variables $X^{[P]}$. However, δ is rarely known. We surmount this problem by using a regression procedure due to Freedman and Lane (1983). Under the null hypothesis, D is not in the model and our permutation approach solves the problem raised by estimating δ by permuting the residuals from the regression of Y on $X^{[L]}$ in orbits that share the same values of $X^{[P]}$, leaving D fixed. The test statistic recorded for each permutation is the t-statistic corresponding to the coefficient representing treatment assignment.[25]

In a series of Monte Carlo studies, Anderson and Legendre (1999) show that the Freedman-Lane procedure generally gives the best results in terms of Type-I error and power among a number of similar permutation-based approximation methods. In another paper, Anderson and Robinson (2001) compare an exact permutation method (where δ is known) with a variety of permutation-based methods. They find that the Freedman-Lane procedure generates test statistics that are distributed most like those generated by the exact method.

4.5 Multiple-Hypothesis Testing: The Stepdown Algorithm

There are many measures in the Perry follow-up study. Some of them are measures of the same variable at different stages of the life cycle of participants. To generate inference using evidence from the study in a robust and defensible way, we use a stepdown algorithm for multiple-hypothesis testing. The procedure

[25] The procedure is described in greater detail in Web Appendix E.

begins with the null hypothesis associated with the most statistically significant statistics and then "steps down" to null hypotheses associated with less significant statistics. The validity of this procedure follows from the analysis of Romano and Wolf (2005), who provide general results on the use of stepdown multiple-hypothesis testing procedures.

We test the hypothesis of no treatment effect for each outcome. We test the null hypothesis of no treatment effect for all K outcomes jointly. The complement of the joint null hypothesis is the hypothesis that there exists at least one hypothesis, out of K, for which there is a treatment effect. After testing for the joint null for all K hypotheses, a stepdown algorithm is performed for the $K - 1$ remaining outcomes targeting the most statistically significant one among the reduced set. The process continues for K cycles. At the end of the procedure, the stepdown method provides K new p-values associated with each original single p-value that correct for the effect of multiple-hypothesis testing on p-values.

The stepdown multiple-hypothesis algorithm of Romano and Wolf (2005) is less conservative than traditional procedures, such as the Bonferroni or Holm procedures, by accounting for relationships among the outcomes. Lehmann and Romano (2005) and Romano and Wolf (2005) discuss the stepdown procedure in depth. We summarize their analysis in Web Appendix F.

We note that there is considerable arbitrariness in defining the blocks of hypotheses that are jointly tested in a multiple hypothesis testing procedure. The Perry study collects information on 715 measures on a variety of diverse outcomes. Associated with each measure is a single null hypothesis. One could test all hypotheses in a single block. However, a test that groups very diverse measures into a single block lacks interpretability. To avoid arbitrariness in selecting blocks of hypotheses, we group hypotheses into economically and substantively meaningful groups, e.g., income, education, health, test scores, and behavioral indices are treated as separate blocks. Each block is of independent interest and would be selected by economists on *a priori* grounds, drawing on information from previous studies on the aspect of participant behavior represented by that block. We test outcomes by age and detect pronounced life cycle effects by gender.

5 Empirical Results

Our empirical findings are consistent with those reported in most of the previous literature on the Perry Preschool program. We find large gender differences in treatment effects for different outcomes at different ages (Heckman, 2005; Schweinhart et al., 2005). However, in contrast to the recent analysis of Anderson (2008), we find statistically significant treatment effects for males on many outcomes. These effects persist after controlling for corrupted randomization and multiple-hypothesis testing. Anderson conducts tests on linear age-specific indices that aggregate treatment effects across conceptually very different outcomes. In contrast, we avoid indices and analyze economically interpretable blocks of outcomes by age. Another difference between our analyses is that his analysis does not correct for the compromised nature of the randomization in the Perry study while ours does. These differences in analytical approaches lead to substantially different conclusions about the effect of the Perry program on males. We discuss other differences between our analysis and his in Section 7.

Tables 3–6 summarize the estimated effects of the Perry program on outcomes grouped by type and age of measurement.[26] Tables 3 and 4 report results for females. Tables 5 and 6 are for males. The first column of each table is the control mean for the indicated outcome. The next two columns are the treatment effect sizes, where the "unconditional" effect is the difference in means between the treatment and control group, and the "conditional" effect is the coefficient on the treatment assignment variable in a linear regression of the outcome with four covariates: maternal employment, paternal presence, socio-economic status (SES) index, and Stanford-Binet IQ, all measured at the age of study entry. The next column gives the estimated effect from the partially linear Freedman-Lane procedure that conditions on socio-economic status. The next four columns are p-values testing the null hypothesis of no treatment effect for the indicated outcome. The second-to-last column, "gender difference-in-difference", tests the null hypothesis of no difference in mean treatment effects between males and females. The final column gives the count of non-missing observations for the indicated outcome.

Outcomes are placed in ascending order of the "partially linear" Freedman-Lane p-value that is described below. This is the order in which the outcomes would be discarded from the joint null hypothesis in the stepdown multiple-hypothesis testing algorithm.[27] The ordering of outcomes differs in the tables for males and females. Additionally, some outcomes are reported for only one gender when insufficient observations were available for reliable testing of the hypothesis for the other gender.[28]

Single p-Values Tables 3–6 show four varieties of p-values for testing the null hypothesis of no treatment effect. The first such value, labeled "naïve", is based on a simple permutation test of the hypothesis of no difference in means between treatment and control groups. This test uses no conditioning, imposes no restrictions on the permutation group, and does not account for imbalances or the compromised Perry randomization. These naïve p-values are very close to their asymptotic equivalents. For evidence on this point, see Web Appendix G.[29]

[26] Perry follow-ups were at ages 19, 27, and 40. We group the outcomes by age whenever they have strong age patterns, for example, in the case of employment or income.

[27] For more on the stepdown algorithm, see Web Appendix F.

[28] Observations are missing to different degrees for different variables.

[29] Anderson (2008) constructs his p values in a similar fashion drawing without replacement and notes that the permutation-based and asymptotic results are in close agreement.

Table 3: Main Outcomes, Females: Part 1

	Outcome	**Age**	**Ctl. Mean**	**Effect**		**p-values**					*N*
				Uncond.[a]	Cond.[b]	Naïve[c]	Full Lin.[d]	Partial Lin.[e]	Part. Lin. (adj.)[f]	Gender D-in-D[g]	
Education	Mentally Impaired?	≤19	0.36	-0.28	-0.29	**.008**	**.009**	**.005**	**.017**	.337	46
	Learning Disabled?	≤19	0.14	-0.14	-0.15	**.009**	**.016**	**.009**	**.025**	**.029**	46
	Yrs. of Special Services	≤14	0.46	-0.26	-0.29	**.036**	**.013**	**.013**	**.025**	.153	51
	Yrs. in Disciplinary Program	≤19	0.36	-0.24	-0.19	**.089**	.127	**.074**	**.074**	.945	46
	HS Graduation	19	0.23	0.61	0.49	**.000**	**.000**	**.000**	**.000**	**.003**	51
	GPA	19	1.53	0.89	0.88	**.000**	**.001**	**.000**	**.001**	**.009**	30
	Highest Grade Completed	19	10.75	1.01	0.94	**.007**	**.008**	**.002**	**.006**	**.052**	49
	# Years Held Back	≤19	0.41	-0.20	-0.14	**.067**	.135	**.097**	.178	.106	46
	Vocational Training Certificate	≤40	0.08	0.16	0.13	**.070**	.106	.107	.107	.500	51
Health	No Health Problems	19	0.83	0.05	0.12	.265	.107	.137	.576	.308	49
	Alive	40	0.92	0.04	0.04	.273	.249	.197	.675	.909	51
	No Treat. for Illness, Past 5 Yrs.	27	0.59	0.05	0.14	.369	.188	.241	.690	.806	47
	No Non-Routine Care, Past Yr.	27	0.00	0.04	0.02	.484	.439	.488	.896	.549	44
	No Sick Days in Bed, Past Yr.	27	0.45	-0.05	-0.04	.623	.597	.529	.781	.412	47
	No Doctors for Illness, Past Yr.	19	0.54	-0.02	-0.01	.559	.539	.549	.549	.609	49
	No Tobacco Use	27	0.41	0.11	0.08	.208	.348	.298	.598	.965	47
	Infrequent Alcohol Use	27	0.67	0.17	0.07	.103	.336	.374	.587	.924	45
	Routine Annual Health Exam	27	0.86	-0.06	-0.09	.684	.751	.727	.727	.867	47
Fam.	Has Any Children	≤ 19	0.52	-0.12	-0.05	.218	.419	.328	.601	—	48
	# Out-of-Wedlock Births	≤ 40	2.52	0.29	-0.51	.652	.257	.402	.402	—	42
Crime	# Non-Juv. Arrests	≤27	1.88	-1.60	-2.22	**.016**	**.003**	**.003**	**.005**	.571	51
	Any Non-Juv. Arrests	≤27	0.35	-0.15	-0.18	.148	.122	.125	.125	.440	51
	# Total Arrests	≤40	4.85	-2.65	-2.88	**.028**	**.037**	**.041**	.128	.566	51
	# Total Charges	≤40	4.92	-2.68	-2.81	**.030**	**.037**	**.042**	.128	.637	51
	# Non-Juv. Arrests	≤40	4.42	-2.26	-2.62	**.044**	**.046**	**.051**	.150	.458	51
	# Misd. Arrests	≤40	4.00	-1.88	-2.19	**.078**	**.078**	**.085**	.232	.549	51
	Total Crime Cost[h]	≤40	293.50	-271.33	-381.03	**.013**	.108	**.090**	.197	.858	51
	Any Arrests	≤40	0.65	-0.09	-0.11	.181	.280	.239	.310	.824	51
	Any Charges	≤40	0.65	-0.09	-0.13	.181	.280	.239	.310	.799	51
	Any Non-Juv. Arrests	≤40	0.54	-0.02	0.02	.351	.541	.520	.520	.463	51
	Any Misd. Arrests	≤40	0.54	-0.02	0.02	.351	.541	.520	.520	.519	51

Notes: Monetary values adjusted to thousands of year-2006 dollars using annual national CPI. (a) Unconditional difference in means between the treatment and control groups; (b) Conditional treatment effect with linear covariates Stanford-Binet IQ, Socio-economic Status index (SES), maternal employment, father's presence at study entry — this is also the effect for Freedman-Lane under a full linearity assumption, whose respective p-value is computed in column "Full Lin."; (c) One-sided p-values for the hypothesis of no treatment effect based on conditional permutation inference, without orbit restrictions or linear covariates — estimated effect size in the "unconditional effect" column; (d) One-sided p-values for the hypothesis of no treatment effect based on the Freedman-Lane procedure, without restricting permutation orbits and assuming linearity in all covariates (maternal employment, paternal presence, Socio-economic Status index (SES), and Stanford-Binet IQ) — estimated effect size in the "conditional effect" column; (e) One-sided p-values for the hypothesis of no treatment effect based on the Freedman-Lane procedure, using the linear covariates maternal employment, paternal presence, and Stanford-Binet IQ, and restricting permutation orbits within strata formed by Socio-economic Status index (SES) being above or below the sample median and permuting siblings as a block; (f) p-values from the previous column, adjusted for multiple inference using stepdown procedure; (g) Two-sided p-value for the null hypothesis of no gender difference in mean treatment effects, tested using mean differences between treatments and controls using the conditioning and orbit restriction setup described in (e); (h) Total crime costs include victimization, police, justice, and incarceration costs, where victimizations are estimated from arrest records for each type of crime using data from urban areas of the Midwest, police and court costs are based on historical Michigan unit costs, and the victimization cost of fatal crime takes into account the statistical value of life (see Heckman, Moon, Pinto, Savelyev, and Yavitz (2009) for details).

Table 4: Main Outcomes, Females: Part 2

	Outcome	**Age**	**Ctl. Mean**	**Effect**		**p-values**					*N*
				Uncond.[a]	Cond.[b]	Naïve[c]	Full Lin.[d]	Partial Lin.[e]	Part. Lin. (adj.)[f]	Gender D-in-D[g]	
Employment	No Job in Past Year	19	0.58	-0.34	-0.37	**.006**	**.007**	**.003**	**.007**	**.009**	51
	Jobless Months in Past 2 Yrs.	19	10.42	-5.20	-5.47	**.054**	**.099**	**.020**	**.036**	.102	42
	Current Employment	19	0.15	0.29	0.23	**.023**	**.045**	**.032**	**.032**	.373	51
	No Job in Past Year	27	0.54	-0.29	-0.25	**.017**	**.058**	**.037**	**.071**	.157	48
	Current Employment	27	0.55	0.25	0.18	**.036**	**.096**	**.042**	**.063**	.220	47
	Jobless Months in Past 2 Yrs.	27	10.45	-4.21	-2.14	**.077**	.285	.165	.165	.908	47
	No Job in Past Year	40	0.41	-0.25	-0.22	**.032**	**.092**	**.056**	.111	.464	47
	Jobless Months in Past 2 Yrs.	40	5.05	-1.05	1.05	.343	.654	.528	.627	.573	46
	Current Employment	40	0.82	0.02	-0.08	.419	.727	.615	.615	.395	46
Earnings[h]	Monthly Earn., Current Job	19	2.08	-0.61	-0.47	.750	.701	.725	—	.677	15
	Monthly Earn., Current Job	27	1.13	0.69	0.48	**.050**	.144	.109	.139	.752	47
	Yearly Earn., Current Job	27	15.45	4.60	2.18	.169	.339	.277	.277	.873	47
	Yearly Earn., Current Job	40	19.85	4.35	4.46	.251	.272	.224	.274	.755	46
	Monthly Earn., Current Job	40	1.85	0.21	0.27	.328	.316	.261	.261	.708	46
Earnings & Emp.[h]	No Job in Past Year	19	0.58	-0.34	-0.37	**.006**	**.007**	**.003**	**.010**	**.009**	51
	Jobless Months in Past 2 Yrs.	19	10.42	-5.20	-5.47	**.054**	**.099**	**.020**	**.056**	.102	42
	Current Employment	19	0.15	0.29	0.23	**.023**	**.045**	**.032**	**.064**	.373	51
	Monthly Earn., Current Job	19	2.08	-0.61	-0.47	.750	.701	.725	.725	.677	15
	No Job in Past Year	27	0.54	-0.29	-0.25	**.017**	**.058**	**.037**	**.094**	.157	48
	Current Employment	27	0.55	0.25	0.18	**.036**	**.096**	**.042**	**.094**	.220	47
	Monthly Earn., Current Job	27	1.13	0.69	0.48	**.050**	.144	.109	.188	.752	47
	Jobless Months in Past 2 Yrs.	27	10.45	-4.21	-2.14	**.077**	.285	.165	.241	.908	47
	Yearly Earn., Current Job	27	15.45	4.60	2.18	.169	.339	.277	.277	.873	47
	No Job in Past Year	40	0.41	-0.25	-0.22	**.032**	**.092**	**.056**	.156	.464	47
	Yearly Earn., Current Job	40	19.85	4.35	4.46	.251	.272	.224	.423	.755	46
	Monthly Earn., Current Job	40	1.85	0.21	0.27	.328	.316	.261	.440	.708	46
	Jobless Months in Past 2 Yrs.	40	5.05	-1.05	1.05	.343	.654	.528	.627	.573	46
	Current Employment	40	0.82	0.02	-0.08	.419	.727	.615	.615	.395	46
Economic	Savings Account	27	0.45	0.27	0.23	**.036**	**.087**	**.051**	.132	.128	47
	Car Ownership	27	0.59	0.13	0.12	.164	.221	.147	.250	.887	47
	Checking Account	27	0.27	0.01	-0.03	.472	.586	.472	.472	.777	47
	Credit Card	40	0.50	0.04	0.06	.425	.355	.233	.483	.737	46
	Checking Account	40	0.50	0.08	0.04	.321	.413	.237	.450	.675	46
	Car Ownership	40	0.77	0.06	0.03	.280	.409	.257	.394	.157	46
	Savings Account	40	0.73	0.06	-0.08	.309	.722	.516	.516	**.071**	46
	Ever on Welfare	18–27	0.82	-0.34	-0.21	**.009**	**.084**	**.049**	.154	**.074**	47
	> 30 Mos. on Welfare	18–27	0.55	-0.27	-0.18	**.036**	.152	**.072**	.187	**.087**	47
	# Months on Welfare	18–27	51.23	-21.51	-11.39	**.060**	.241	.120	.265	.122	47
	Never on Welfare	16–40	0.92	0.16	0.13	.110	.129	.132	.221	.970	51
	Never on Welfare (Self Rep.)	26–40	0.41	-0.09	-0.14	.759	.787	.664	.664	.118	46

Notes: Monetary values adjusted to thousands of year-2006 dollars using annual national CPI. (a) Unconditional difference in means between the treatment and control groups; (b) Conditional treatment effect with linear covariates Stanford-Binet IQ, Socio-economic Status index (SES), maternal employment, father's presence at study entry — this is also the effect for Freedman-Lane under a full linearity assumption, whose respective *p*-value is computed in column "Full Lin."; (c) One-sided *p*-values for the hypothesis of no treatment effect based on conditional permutation inference, without orbit restrictions or linear covariates — estimated effect size in the "unconditional effect" column; (d) One-sided *p*-values for the hypothesis of no treatment effect based on the Freedman-Lane procedure, without restricting permutation orbits and assuming linearity in all covariates (maternal employment, paternal presence, Socio-economic Status index (SES), and Stanford-Binet IQ) — estimated effect size in the "conditional effect" column; (e) One-sided *p*-values for the hypothesis of no treatment effect based on the Freedman-Lane procedure, using the linear covariates maternal employment, paternal presence, and Stanford-Binet IQ, and restricting permutation orbits within strata formed by Socio-economic Status index (SES) being above or below the sample median and permuting siblings as a block; (f) *p*-values from the previous column, adjusted for multiple inference using stepdown procedure; (g) Two-sided *p*-value for the null hypothesis of no gender difference in mean treatment effects, tested using mean differences between treatments and controls using the conditioning and orbit restriction setup described in (e); (h) Age-19 measures are conditional on at least some earnings during the period specified — observations with zero earnings are omitted in computing means and regressions.

Table 5: Main Outcomes, Males: Part 1

	Outcome	Age	Ctl. Mean	Effect		p-values					N
				Uncond.[a]	Cond.[b]	Naïve[c]	Full Lin.[d]	Partial Lin.[e]	Part. Lin. (adj.)[f]	Gender D-in-D[g]	
Education	Mentally Impaired?	≤19	0.33	-0.13	-0.19	.106	**.072**	**.057**	.190	.337	66
	Yrs. in Disciplinary Program	≤19	0.42	-0.12	-0.26	.313	.153	.134	.334	.945	66
	Yrs. of Special Services	≤14	0.46	-0.04	-0.10	.458	.256	.205	.349	.153	72
	Learning Disabled?	≤19	0.08	0.08	0.08	.840	.841	.766	.766	**.029**	66
	Highest Grade Completed	19	11.28	0.08	0.01	.429	.383	.312	.718	**.052**	72
	GPA	19	1.79	0.02	-0.01	.464	.517	.333	.716	**.009**	47
	Vocational Training Certificate	≤40	0.33	0.06	0.06	.231	.304	.406	.729	.500	72
	HS Graduation	19	0.51	-0.03	0.00	.633	.510	.416	.583	**.003**	72
	# Years Held Back	≤19	0.39	0.08	0.12	.740	.852	.745	.745	.106	66
Health	Alive	40	0.92	0.05	0.05	.160	.174	.146	.604	.909	72
	No Sick Days in Bed, Past Yr.	27	0.38	0.10	0.14	.208	.135	.162	.582	.412	70
	No Treat. for Illness, Past 5 Yrs.	27	0.64	0.00	0.01	.465	.417	.375	.826	.806	70
	No Doctors for Illness, Past Yr.	19	0.56	0.07	0.02	.210	.435	.453	.835	.609	72
	No Non-Routine Care, Past Yr.	27	0.17	-0.03	-0.02	.600	.548	.548	.823	.549	63
	No Health Problems	19	0.95	-0.07	-0.08	.849	.843	.862	.862	.308	72
	Infrequent Alcohol Use	27	0.58	0.18	0.21	**.072**	**.024**	**.052**	.139	.924	66
	No Tobacco Use	27	0.46	0.12	0.10	.143	.220	.260	.436	.965	70
	Routine Annual Health Exam	27	0.74	-0.04	0.01	.622	.397	.451	.451	.867	68
Crime	# Non-Juv. Arrests	≤27	5.36	-2.33	-2.64	**.029**	**.028**	**.017**	**.047**	.571	72
	# Fel. Arrests	≤27	2.33	-1.12	-1.07	**.046**	**.081**	**.043**	.101	—	72
	Any Non-Juv. Arrests	≤27	0.72	0.02	-0.05	.501	.422	.291	.418	.440	72
	Any Fel. Arrests	≤27	0.49	0.00	-0.01	.494	.575	.442	.442	—	72
	Any Non-Juv. Arrests	≤40	0.92	-0.14	-0.12	**.090**	.124	**.078**	.192	.463	72
	Any Fel. Arrests	≤40	0.44	-0.16	-0.15	**.047**	.133	**.083**	.191	—	72
	Any Arrests	≤40	0.95	-0.13	-0.11	**.072**	.142	.123	.181	.824	72
	Any Misd. Arrests	≤40	0.87	-0.11	-0.08	.166	.281	.191	.191	.519	72
	# Misd. Arrests	≤40	8.46	-3.13	-3.42	**.037**	**.043**	**.021**	**.039**	.549	72
	# Non-Juv. Arrests	≤40	11.72	-4.26	-4.45	**.039**	**.053**	**.025**	**.041**	.458	72
	# Total Arrests	≤40	12.41	-4.20	-4.44	**.056**	**.073**	**.036**	**.053**	.566	72
	# Fel. Arrests	≤40	3.26	-1.14	-1.03	.112	.173	**.092**	**.092**	—	72
	# Non-Victimless Charges[i]	≤40	3.08	-1.59	-1.65	**.029**	**.048**	**.027**	.113	.175	72
	# Total Charges	≤40	13.38	-4.38	-5.08	**.063**	**.081**	**.041**	.152	.637	72
	Total Crime Cost[h]	≤40	775.90	-351.22	-515.10	.153	.108	**.070**	.209	.858	72
	Any Non-Victimless Charges[i]	≤40	0.62	-0.16	-0.15	.105	.179	.112	.263	.957	72
	Ever Incarcerated	≤40	0.23	-0.08	-0.11	.260	.159	.114	.206	.563	72
	Any Charges	≤40	0.95	-0.13	-0.09	**.072**	.142	.123	.123	.799	72

Notes: Monetary values adjusted to thousands of year-2006 dollars using annual national CPI. (a) Unconditional difference in means between the treatment and control groups; (b) Conditional treatment effect with linear covariates Stanford-Binet IQ, Socio-economic Status index (SES), maternal employment, father's presence at study entry — this is also the effect for Freedman-Lane under a full linearity assumption, whose respective p-value is computed in column "Full Lin."; (c) One-sided p-values for the hypothesis of no treatment effect based on conditional permutation inference, without orbit restrictions or linear covariates — estimated effect size in the "unconditional effect" column; (d) One-sided p-values for the hypothesis of no treatment effect based on the Freedman-Lane procedure, without restricting permutation orbits and assuming linearity in all covariates (maternal employment, paternal presence, Socio-economic Status index (SES), and Stanford-Binet IQ) — estimated effect size in the "conditional effect" column; (e) One-sided p-values for the hypothesis of no treatment effect based on the Freedman-Lane procedure, using the linear covariates maternal employment, paternal presence, and Stanford-Binet IQ, and restricting permutation orbits within strata formed by Socio-economic Status index (SES) being above or below the sample median and permuting siblings as a block; (f) p-values from the previous column, adjusted for multiple inference using stepdown procedure; (g) Two-sided p-value for the null hypothesis of no gender difference in mean treatment effects, tested using mean differences between treatments and controls using the conditioning and orbit restriction setup described in (e); (h) Total crime costs include victimization, police, justice, and incarceration costs, where victimizations are estimated from arrest records for each type of crime using data from urban areas of the Midwest, police and court costs are based on historical Michigan unit costs, and the victimization cost of fatal crime takes into account the statistical value of life (see Heckman, Moon, Pinto, Savelyev, and Yavitz (2009) for details); (i) Non-victimless crimes are those associated with victimization costs: murder, rape, robbery, assault, burglary, larceny, and motor vehicle theft (see Heckman, Moon, Pinto, Savelyev, and Yavitz (2009) for details).

Table 6: Main Outcomes, Males: Part 2

	Outcome	Age	Ctl. Mean	Effect		p-values					*N*
				Uncond.[a]	Cond.[b]	Naïve[c]	Full Lin.[d]	Partial Lin.[e]	Part. Lin. (adj.)[f]	Gender D-in-D[g]	
Employment	Current Employment	19	0.41	0.14	0.13	.101	.144	.103	.196	.373	72
	Jobless Months in Past 2 Yrs.	19	3.82	1.47	1.31	.784	.763	.781	.841	.102	70
	No Job in Past Year	19	0.13	0.11	0.09	.924	.827	.857	.857	**.009**	72
	Jobless Months in Past 2 Yrs.	27	8.79	-3.66	-4.09	**.059**	**.057**	**.033**	**.065**	.908	69
	No Job in Past Year	27	0.31	-0.07	-0.07	.260	.295	.192	.294	.157	72
	Current Employment	27	0.56	0.04	0.09	.367	.251	.219	.219	.220	69
	Current Employment	40	0.50	0.20	0.29	**.059**	**.011**	**.011**	**.024**	.395	66
	Jobless Months in Past 2 Yrs.	40	10.75	-3.52	-4.59	**.082**	**.040**	**.018**	**.026**	.573	66
	No Job in Past Year	40	0.46	-0.10	-0.15	.249	.123	**.068**	**.068**	.464	72
Earnings[h]	Monthly Earn., Current Job	19	2.74	-0.16	0.09	.591	.408	.442	—	.677	30
	Monthly Earn., Current Job	27	1.43	0.88	0.99	**.017**	**.014**	**.011**	**.018**	.752	68
	Yearly Earn., Current Job	27	21.51	3.50	3.67	.227	.248	.186	.186	.873	66
	Yearly Earn., Current Job	40	24.23	7.17	4.62	.147	.270	.150	.203	.755	66
	Monthly Earn., Current Job	40	2.11	0.50	0.44	.224	.277	.195	.195	.708	66
Earnings & Emp.[h]	Current Employment	19	0.41	0.14	0.13	.101	.144	.103	.279	.373	72
	Monthly Earn., Current Job	19	2.74	-0.16	0.09	.591	.408	.442	.736	.677	30
	Jobless Months in Past 2 Yrs.	19	3.82	1.47	1.31	.784	.763	.781	.841	.102	70
	No Job in Past Year	19	0.13	0.11	0.09	.924	.827	.857	.857	**.009**	72
	Monthly Earn., Current Job	27	1.43	0.88	0.99	**.017**	**.014**	**.011**	**.037**	.752	68
	Jobless Months in Past 2 Yrs.	27	8.79	-3.66	-4.09	**.059**	**.057**	**.033**	**.084**	.908	69
	Yearly Earn., Current Job	27	21.51	3.50	3.67	.227	.248	.186	.360	.873	66
	No Job in Past Year	27	0.31	-0.07	-0.07	.260	.295	.192	.294	.157	72
	Current Employment	27	0.56	0.04	0.09	.367	.251	.219	.219	.220	69
	Current Employment	40	0.50	0.20	0.29	**.059**	**.011**	**.011**	**.035**	.395	66
	Jobless Months in Past 2 Yrs.	40	10.75	-3.52	-4.59	**.082**	**.040**	**.018**	**.045**	.573	66
	No Job in Past Year	40	0.46	-0.10	-0.15	.249	.123	**.068**	.137	.464	72
	Yearly Earn., Current Job	40	24.23	7.17	4.62	.147	.270	.150	.203	.755	66
	Monthly Earn., Current Job	40	2.11	0.50	0.44	.224	.277	.195	.195	.708	66
Economic	Car Ownership	27	0.59	0.15	0.18	**.089**	**.072**	**.059**	.152	.887	70
	Savings Account	27	0.46	-0.01	0.03	.555	.425	.397	.610	.128	70
	Checking Account	27	0.23	-0.04	-0.02	.591	.610	.575	.575	.777	70
	Savings Account	40	0.36	0.37	0.36	**.002**	**.002**	**.001**	**.003**	**.071**	66
	Car Ownership	40	0.50	0.30	0.32	**.004**	**.003**	**.002**	**.004**	.157	66
	Credit Card	40	0.36	0.11	0.08	.180	.279	.206	.327	.737	66
	Checking Account	40	0.39	0.01	-0.01	.463	.558	.491	.491	.675	66
	Never on Welfare	16–40	0.82	0.15	0.17	.101	**.086**	**.028**	.104	.970	72
	Never on Welfare (Self Rep.)	26–40	0.38	0.18	0.18	**.058**	**.075**	**.051**	.147	.118	64
	> 30 Mos. on Welfare	18–27	0.08	-0.01	0.02	.571	.482	.430	.619	**.087**	66
	# Months on Welfare	18–27	6.84	0.59	0.14	.563	.566	.517	.646	.122	66
	Ever on Welfare	18–27	0.26	0.06	0.02	.697	.635	.590	.590	**.074**	66

Notes: Monetary values adjusted to thousands of year-2006 dollars using annual national CPI. (a) Unconditional difference in means between the treatment and control groups; (b) Conditional treatment effect with linear covariates Stanford-Binet IQ, Socio-economic Status index (SES), maternal employment, father's presence at study entry — this is also the effect for Freedman-Lane under a full linearity assumption, whose respective *p*-value is computed in column "Full Lin."; (c) One-sided *p*-values for the hypothesis of no treatment effect based on conditional permutation inference, without orbit restrictions or linear covariates — estimated effect size in the "unconditional effect" column; (d) One-sided *p*-values for the hypothesis of no treatment effect based on the Freedman-Lane procedure, without restricting permutation orbits and assuming linearity in all covariates (maternal employment, paternal presence, Socio-economic Status index (SES), and Stanford-Binet IQ) — estimated effect size in the "conditional effect" column; (e) One-sided *p*-values for the hypothesis of no treatment effect based on the Freedman-Lane procedure, using the linear covariates maternal employment, paternal presence, and Stanford-Binet IQ, and restricting permutation orbits within strata formed by Socio-economic Status index (SES) being above or below the sample median and permuting siblings as a block; (f) *p*-values from the previous column, adjusted for multiple inference using stepdown procedure; (g) Two-sided *p*-value for the null hypothesis of no gender difference in mean treatment effects, tested using mean differences between treatments and controls using the conditioning and orbit restriction setup described in (e); (h) Age-19 measures are conditional on at least some earnings during the period specified — observations with zero earnings are omitted in computing means and regressions.

The next three p-values are based on variants of a procedure due to Freedman and Lane (1983) for combining regression with permutation testing for admissible permutation groups. The first Freedman-Lane p-value, labeled "full linearity", tests the significance of the treatment effect adjusting outcomes using linear regression with four covariates: maternal employment, paternal presence, socio-economic status (SES) index, and Stanford-Binet IQ, all measured at study entry.[30] The second Freedman-Lane type p-value, labeled "partial linearity", allows for a nonparametric relationship between the SES index and outcomes, assuming a linear relationship for the other three covariates. This nonparametric conditioning on SES is achieved by restricting the orbits of the permutations used in the test: the exchangeability of treatment assignments between observations is assumed only on subsamples with similar values of the SES index. In addition, the permutation distribution for the partially linear p-values permute siblings as a block. Admissible permutations do not assign siblings to different treatment and control statuses. These two modifications account for the compromised randomization of the Perry study.[31] The third p-value for the Freedman-Lane procedure incorporates an adjustment for multiple hypothesis testing using the stepdown algorithm described below.

Stepdown p-Values and Multiple-Hypothesis Testing We divide outcomes into blocks for multiple-hypothesis testing by type of outcome, similarities on the type of measure, and age if there is an obvious age pattern.[32] In Tables 3–6, these blocks are delineated by horizontal lines.[33] In our analysis, within each block, the "partially linear" (adjusted) p-value is the set of p-values obtained from the partially linear model adjusted for multiple-hypothesis testing using the stepdown algorithm. The adjusted p-value in each row corresponds to a joint hypothesis test of the indicated outcome and the outcomes listed below within that block. Specifically, the joint null hypothesis is that there is no treatment effect for the remaining outcomes. The alternative is that there is a treatment effect for at least one of the remaining outcomes. This stepwise ordering is the reason why we report outcomes placed in ascending order of their p-values. The stepdown-adjusted p-values are based on these values, and the most individually-significant remaining outcome — the one most likely to contribute to the significance of the joint null hypothesis — is removed from the joint null hypothesis at each successive step.

The first stepdown p-value within a block is especially important because it tests the overall joint null hypothesis of no treatment effect for all outcomes in the block. The inference obtained from this procedure

[30] Note that these are the same four used to produce the conditional effect size previously described.

[31] Partial linearity is a valid assumption if full linearity is a valid assumption, although the converse need not necessarily hold since a nonparametric approach is less restrictive than a linear parametric approach.

[32] Education, health, family composition, criminal behavior, employment status, earnings, and general economic activities are the categories of variables on which blocks are selected on *a priori* grounds.

[33] This approach differs from that taken by Anderson (2008), who aggregates conceptually distinct outcomes into estimated linear indices. His tests are conducted on the constructed indices.

is analogous to that obtained from the classical asymptotic F-test for the difference in means for the set of outcomes in question. The effect of the adjustment that stepdown introduces is that the probability of rejecting any true null hypothesis at *any* step of the stepwise joint hypothesis testing procedure is kept below a certain threshold.

In summary, the stepdown algorithm proceeds as follows. For each joint hypothesis, and for each set of permutations, the stepdown procedure records the maximum p-value across those generated by tests of the null hypothesis of no treatment effect for each outcome, separately. The stepdown-adjusted p-value is the proportion of permutations which have a stepdown statistic larger than the statistic for the non-permuted data (the sample data).[34]

Statistics For most outcomes, we use the t-statistic from the difference in means or the coefficient on D in a Freedman-Lane procedure as test statistics.[35] All p-values are computed using 30,000 draws under the relevant permutation procedure. All inference is based on one-sided p-values with the assumption that treatment is not harmful. An exception is the test for differences in treatment effects by gender, which are based on two-sided p-values.

Main Results Tables 3–6 show many statistically significant treatment effects and gender differences that survive multiple hypothesis testing. In summary, females show strong effects for educational outcomes, early employment and other early economic outcomes, as well as reduced numbers of arrests. Males, on the other hand, show strong effects on a number of outcomes, demonstrating a substantially reduced number of arrests and lower probability of imprisonment, as well as strong effects on earnings at age 27, employment at age 40, and other economic outcomes recorded at age 40.

A principal contribution of this paper is to tackle the statistical challenges posed by the problems of small sample size, imbalance in the covariates, and compromised randomization. In doing so, we find substantial differences in inference between the testing procedures that use naïve p-values versus the Freedman-Lane p-values. The latter correct for the compromised nature of the randomization protocol. The rejection rate when correcting for these problems is often *higher*, sharpening the evidence for treatment effects from the Perry program. This is evidenced by a general fall in p-values when moving from “naïve” to “full linearity” to “partial linearity”. Using a procedure that corrects for imperfections in the randomization protocol often strengthens the evidence for a program effect. In several cases, outcomes that are statistically insignificant at a ten percent level using naïve p-values are shown to be statistically significant using p-values derived from

[34] See Web Appendix F for details on how we implement stepdown as well as a more general theoretical description of the procedure.

[35] For full-scale IQ, we use the Mann-Whitney U-test statistic, which uses ranks of IQ distributions instead of IQ scores.

the partially linear Freedman-Lane model. For instance, consider the p values for "lifetime crime costs" and "ever receiving welfare at ages 16–40" for males.

Schooling Within the group of hypotheses for education, the only statistically-significant treatment effect for males is the effect associated with being classified as mentally impaired through age 19 (Table 5). However, as Table 3 shows, there are strong treatment effects for females on high school GPA, graduation, highest grade completed, mental impairment, learning disabilities, etc. Additionally, we fail to reject the overall joint null hypotheses for both school achievement and for lifetime educational outcomes. The hypothesis of no difference between sexes in schooling outcomes is rejected for the outcomes of highest grade completed, GPA, high school graduation, and the presence of a learning disability. The unimpressive education results for males, however, do not necessarily mean that the pattern would be reproduced if the program were replicated today. We briefly discuss this point in Section 6.[36] We discuss the effects of the intervention on cognitive test scores in Web Appendix I. Heckman, Malofeeva, Pinto, and Savelyev (2009) discuss the impact of the Perry program on noncognitive skills. They decompose treatments effects into effects due to cognitive and noncognitive enhancements of the program.

Employment and Earnings Results for employment and earnings are displayed in Table 4 for females and Table 6 for males. The treatment effects in these outcomes exhibit gender differences and a distinctive age pattern. For females, we observe statistically significant employment effects in the overall joint null hypotheses at ages 19 and 27. Only one outcome does not survive stepdown adjustment-jobless months in past two years at age 27. At age 40, however, there are no statistically significant earnings effects for females considered as individual outcomes, and hence, in sets of joint null hypotheses by age. For males, we observe no significant employment effects at age 19. We reject the overall joint null hypotheses of no difference in employment outcomes at ages 27 and 40. We also reject the null hypotheses of no treatment effect on age-40 employment outcomes individually. When male earnings outcomes are considered alone, we reject only the overall joint null hypothesis at age 27. However, when earnings are considered together with employment, we reject both the overall age-27 and age-40 joint null hypotheses. As is the case for females, earnings outcomes do not survive the stepdown adjustment for combined earnings and employment outcomes at age 40.

Economic Activity Tests for other economic outcomes, shown in Tables 4 and 6, reinforce the conclusions drawn from the analysis of employment outcomes above. Treated males and females are generally more likely to have savings accounts and own cars at the same ages that they are more likely to be employed. The effects

[36] We present a more extensive discussion of this point in Web Appendix K.

on welfare dependence are strong for males when considered through age 40, but weak when considered only through age 27; the converse is true for females.

Criminal Activity Tables 3 and 5 show strong treatment effects on criminal activity for both genders. Males are arrested far more frequently than females, and on average male crimes tend to be more serious, but there are *no* statistically significant gender differences for comparable outcomes. By age 27, control females had been arrested 1.88 times on average during adulthood, including 0.27 felony arrests, while the comparable figures for control males are 5.36 and 2.33.[37] Also, treated males are statistically significantly less likely to be in prison at age 40 than their control counterparts.[38] Figure 4 shows cumulative distribution functions for charges cited at all arrests through age 40 for the male subsample. Figure 4a includes all types of charges, while Figure 4b includes only charges with nonzero victim costs. The latter category of charges is relevant because the costs of criminal victimization resulting from crimes committed by the Perry sample play a key role in determining the economic return to the Perry Preschool program. This is reflected in the statistical significance of estimated differences in total crime costs between treated and untreated groups at the 10% level based on the Freedman-Lane procedure using the partially linear model for both males and females. Total crime costs include victimization, police, justice, and incarceration costs, where victimizations are estimated from arrest records for each type of crime using data from urban areas of the Midwest, police and court costs are based on historical Michigan unit costs, and the victimization cost of fatal crime takes into account the statistical value of life.[39] In terms of the overall joint null hypotheses for the number of arrests, for males we reject at age 27 and for age-40 count measures but not for indicator measures for whether there were any arrests in those same categories. For females, we reject the joint null hypothesis at age 27 and fail to reject at age 40. However, these tests are based on a smaller set of outcomes due to limitations in the data for female crime outcomes.

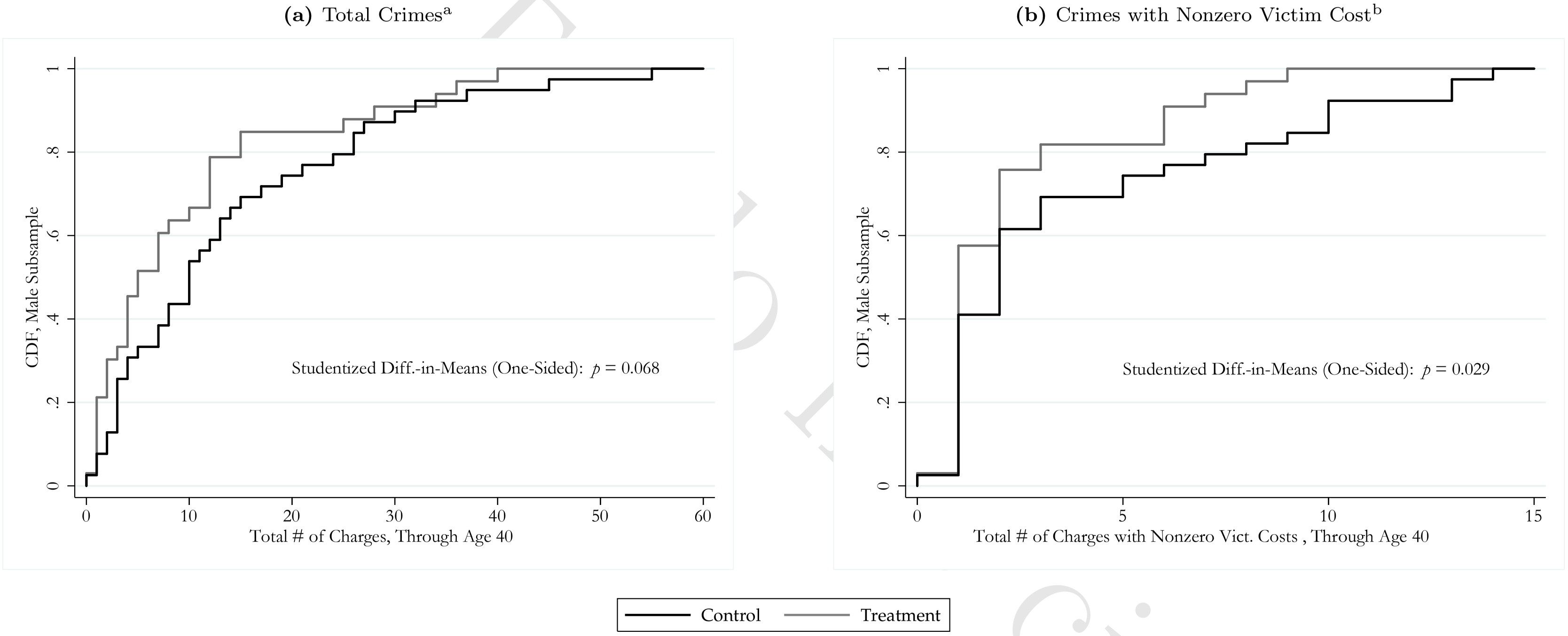
Sensitivity Analysis Our calculations, based on the Freedman-Lane procedure under the assumption of partial linearity, rely on linear parametric approximations and on a particular choice of SES index percentiles to define permutation orbits. Other choices are possible. Any or all of the four covariates that we use in the Freedman-Lane procedure under full linearity could have been used as conditioning variables to define restricted permutation orbits under a partial linearity assumption. We choose SES to condition on because it is a composite of many of the socio-economic characteristics of study participants, and likely has a complex

[37] Statistics for female felony arrests are not shown in the table due to their low reliability: small sample is combined with low incidence of felony arrests.

[38] The set of crime hypotheses is different for males and females due to small sample sizes: we cannot reliably measure the probability of incarceration for females for Perry sample.

[39] Heckman, Moon, Pinto, Savelyev, and Yavitz (2009) present a detailed analysis of total crime cost and its contributions to the economic return to the Perry program.

Figure 4: CDF of Lifetime Charges: Males



(a) Total Crimes[a]

(b) Crimes with Nonzero Victim Cost[b]

Notes: (a) Includes all charges cited at arrests through age 40; (b) Includes all charges with nonzero victim costs cited at arrests through age 40.

interaction with the outcomes.

It is informative to conduct a sensitivity analysis on the effects of choice of conditioning strata, which correspond to the covariates whose relationship with the outcome is assumed to be nonlinear rather than linear. To test the sensitivity of our results to the choice of stratum, we run a series of partially linear Freedman-Lane procedures varying assumptions regarding the set of which covariates enter linearly.

As previously noted, the four pre-program covariates in question can be used either as a Freedman-Lane regressor, assuming a linear relationship with outcomes, or as conditioning variables that limit the orbits of permutations to their selected quantiles which allows for a nonlinear relationship. In Web Appendix H, we perform two types of sensitivity analysis. The first shows that the results reported in Tables 3–6 are robust to variations in the way that percentiles of the SES index are used to generate the strata on which permutations are restricted. The second shows that our results are robust to choices of which covariates enter the outcome model linearly.

Benefit-Cost and Rate of Return Analyses Heckman, Moon, Pinto, Savelyev, and Yavitz (2009) calculate rates of return and compute benefit-cost ratios to determine the private and public returns to the Perry Preschool program. Their analysis includes costs and benefits due to earnings, education, welfare and government assistance, and crime. They adjust estimates for compromised randomization by conditioning lifetime net benefit streams on imbalanced pre-program variables. They also develop standard errors for their estimates. No previous estimates of the rate of return to the Perry program report standard errors. Retrospective earnings data are augmented with data generated from various imputation and extrapolation schemes to construct full earnings profiles through age 65. Sensitivity analysis is conducted to examine the effects of alternative earnings interpolation/extrapolation methods and assumptions used in computing crime costs on the estimated rate of return. In addition, calculations are performed under different assumptions about the deadweight loss of taxation.

Table 7 summarizes their estimates of the Perry program's internal rate of return — the annualized effective compounded return rate that can be earned on capital invested in it. We report estimates that are corrected for imbalance in covariates and compromised randomization and those that are not. Standard errors are generated by a bootstrapping procedure described in Heckman, Moon, Pinto, Savelyev, and Yavitz (2009).

Since reduced crime is a major benefit of the Perry program and estimating the costs of crime entails some element of judgement, we analyze the sensitivity of our results to alternative assumptions. "High" assigns a high value of life ($4.1 million in 2006 dollars) to evaluate murders. "Low" assigns the same cost as that of assault ($13 thousand). We also distinguish estimates that break out very detailed components of crimes ("Separate") from those that aggregate crimes into two categories ("Property/Violent"). Alternative conventions regarding costs are used in the literature.[40]

We adjust upward the costs of government services to account for the deadweight costs of taxation. The estimated rates of return reflect different assumptions about deadweight costs in the literature. The estimated benefit-cost ratios are computed under alternative assumptions on the appropriate social discount rate. It is common in the literature to use a 3% value.[41]

A general pattern emerges from Table 7. Rates of return survive adjustment for compromised randomization. If anything, adjusted rates of return are more precisely estimated than unadjusted rates of return. For benefit-cost ratios, adjustment tends to make estimates less precise. The evidence supports a high rate of return to the Perry program on par with or above the estimated rate of return to World War II equity of 5.8% (DeLong and Magin, 2009). However, the estimated rates of return are well below the 16% rate of return reported by Rolnick and Grunewald (2003) and the 17% rate of return reported by Belfield, Nores, Barnett, and Schweinhart (2006).

Understanding Treatment Effects Heckman, Malofeeva, Pinto, and Savelyev (2009) go beyond treatment effects by explaining the channels through which treatment operates. Their paper uses factor analysis to estimate a model of latent cognitive and noncognitive traits. The model motivating their analysis is one in which treatment effects operate by enhancing cognitive and non-cognitive abilities which determine, in part, program outcomes. Treatment effects can be decomposed in terms of shifts in the distributions of these abilities and the effects of the abilities on outcomes. Their model allows for a third component in the treatment effect decomposition, which represents the effect not explained by their measures of cognitive and noncognitive abilities. Estimates based on this model reveal that abilities for the treated and for the controls are statistically different in terms of variance and mean. Further, early childhood investment embodied in the Perry program has a substantial impact on non-cognitive abilities.

Measures of IQ—purely cognitive measures—exhibit a surge for the treatment group at ages 3 and 4. This difference fades into insignificance by age 10. Yet, despite a lack of statistically significant differences in IQ levels, strong treatment effects remain for both genders at later ages. This suggests that enhancements of non-cognitive skills are a main channel through which Perry treatment effects are produced.

[40] See Heckman, Moon, Pinto, Savelyev, and Yavitz (2009).

[41] The appropriate social discount rate is a hotly debated topic. Some have argued for a zero or negative social discount rate (Dasgupta, Mäler, and Barrett, 2000).

Table 7: IRRs(%) and Benefit-to-Cost Ratios, Adjusted and Unadjusted for Imbalance in Covariates and Compromise in the Randomization (Standard errors in parenthesis)

		Return To:	Individual			Society[d]			Society[d]			Society[d]		
		Arrest Ratio[a] **Murder Cost**[b]				Separate High ($4.1M)			Separate Low ($13K)			Prop. / Violent Low ($13K)		
	Deadweight Loss[c]		All[e]	Male	Fem.	All[e]	Male	Fem.	All[e]	Male	Fem.	All[e]	Male	Fem.
Internal Rates of Return	0%	Adjusted[f]	7.6 *(1.8)*	8.4 *(1.7)*	7.8 *(1.1)*	9.9 *(4.1)*	11.4 *(3.4)*	17.1 *(4.9)*	9.0 *(3.5)*	12.2 *(3.1)*	9.8 *(1.8)*	8.9 *(3.8)*	12.5 *(2.8)*	10.7 *(2.2)*
		Unadjusted	7.4 *(1.2)*	8.0 *(1.2)*	7.9 *(1.6)*	8.0 *(4.2)*	9.6 *(4.8)*	16.3 *(4.3)*	9.4 *(2.5)*	12.4 *(3.0)*	10.4 *(3.3)*	9.2 *(3.4)*	12.6 *(4.1)*	11.1 *(4.3)*
	50%	Adjusted[f]	6.2 *(1.2)*	6.8 *(1.1)*	6.8 *(1.0)*	9.2 *(2.9)*	10.7 *(3.2)*	14.9 *(4.8)*	8.1 *(2.6)*	11.1 *(3.1)*	8.1 *(1.7)*	8.1 *(2.9)*	11.4 *(3.0)*	9.0 *(2.0)*
		Unadjusted	6.0 *(1.4)*	6.5 *(1.4)*	6.8 *(0.8)*	7.6 *(5.0)*	9.2 *(5.2)*	14.4 *(3.9)*	8.6 *(2.8)*	11.3 *(3.1)*	9.2 *(2.9)*	8.4 *(4.0)*	11.5 *(4.7)*	9.8 *(3.9)*
	100%	Adjusted[f]	5.3 *(1.1)*	5.9 *(1.1)*	5.7 *(0.9)*	8.7 *(2.5)*	10.2 *(3.1)*	13.6 *(4.9)*	7.6 *(2.4)*	10.4 *(2.9)*	7.5 *(1.8)*	7.6 *(2.6)*	10.7 *(3.1)*	8.3 *(2.1)*
		Unadjusted	5.1 *(1.1)*	5.6 *(1.3)*	5.7 *(1.3)*	7.4 *(3.6)*	8.9 *(3.6)*	13.2 *(4.3)*	8.1 *(2.1)*	10.6 *(2.5)*	8.6 *(3.1)*	8.0 *(2.6)*	10.8 *(3.6)*	9.1 *(3.2)*
	Discount Rate					All	Male	Fem.	All	Male	Fem.	All	Male	Fem.
Benefit-Cost Ratios	0%	Adjusted[f]	—	—	—	31.5 *(11.3)*	33.7 *(17.3)*	27.0 *(14.4)*	19.1 *(5.4)*	22.8 *(8.3)*	12.7 *(3.8)*	21.4 *(6.1)*	25.6 *(9.6)*	14.0 *(4.3)*
		Unadjusted	—	—	—	29.1 *(10.7)*	30.8 *(17.3)*	25.1 *(12.1)*	21.0 *(5.5)*	25.4 *(8.5)*	13.7 *(3.5)*	23.4 *(6.2)*	28.5 *(10.0)*	14.7 *(3.8)*
	3%	Adjusted[f]	—	—	—	12.2 *(5.3)*	12.1 *(8.0)*	11.6 *(7.1)*	7.1 *(2.3)*	8.6 *(3.7)*	4.5 *(1.4)*	7.9 *(2.7)*	9.5 *(4.4)*	5.1 *(1.7)*
		Unadjusted	—	—	—	11.2 *(5.0)*	11.0 *(8.1)*	10.7 *(5.9)*	8.2 *(2.3)*	10.1 *(3.6)*	5.1 *(1.3)*	9.1 *(2.6)*	11.2 *(4.3)*	5.6 *(1.5)*
	5%	Adjusted[f]	—	—	—	6.8 *(3.4)*	6.2 *(5.1)*	7.1 *(4.6)*	3.9 *(1.5)*	4.7 *(2.3)*	2.4 *(0.8)*	4.3 *(1.7)*	5.1 *(2.8)*	2.8 *(1.1)*
		Unadjusted	—	—	—	6.2 *(3.3)*	5.5 *(5.2)*	6.6 *(3.9)*	4.6 *(1.4)*	5.7 *(2.2)*	2.8 *(0.8)*	5.1 *(1.6)*	6.3 *(2.7)*	3.1 *(0.9)*
	7%	Adjusted[f]	—	—	—	3.9 *(2.3)*	3.2 *(3.4)*	4.6 *(3.1)*	2.2 *(0.9)*	2.7 *(1.5)*	1.4 *(0.5)*	2.5 *(1.1)*	2.9 *(1.8)*	1.7 *(0.7)*
		Unadjusted	—	—	—	3.5 *(2.2)*	2.8 *(3.5)*	4.2 *(2.6)*	2.7 *(0.9)*	3.4 *(1.4)*	1.6 *(0.5)*	3.0 *(1.0)*	3.7 *(1.7)*	1.9 *(0.6)*

Source: Heckman, Moon, Pinto, Savelyev, and Yavitz (2009).

Notes: In this table, kernel matching is used to impute missing values in earnings before age 40, and PSID projection for extrapolation of later earnings using a dynamic regression model. In calculating benefit-to-cost ratios, deadweight loss of taxation is assumed at 50%. Standard errors in parentheses are calculated by Monte Carlo resampling of prediction errors and bootstrapping. Heckman, Moon, Pinto, Savelyev, and Yavitz (2009) produce a range of estimates under alternative assumptions that are consistent with the estimates reported in Table 7. (a) A ratio of victimization rate (from the National Criminal Victimization Study) to arrest rate (from the Uniform Crime Report), where "Prop. /Violent" uses common ratios based on a crime being either violent or property and "Separate" does not; (b) "high" murder cost accounts for statistical value of life, while "low" does not; (c) Deadweight cost is dollars of welfare loss per tax dollar; (d) The sum of returns to program participants and the general public; (e) "All" is computed from an average of the profiles of the pooled sample, and may be lower or higher than the profiles for each gender group; (f) Lifetime net benefit streams are adjusted for corrupted randomization by being conditioned on unbalanced pre-program variables.

6 External Validity

This section evaluates the representativeness of the Perry sample. We construct a comparison group using the 1979 National Longitudinal Survey of Youth (NLSY79), a widely used nationally representative longitudinal dataset. The NLSY79 has panel data on wages, schooling, and employment for a cohort of young adults, ages 14-22 at their first interview in 1979. This cohort has been followed ever since. For our purposes, an important feature is that the NLSY79 contains information on cognitive test scores as well as non-cognitive measures, and has rich information on family background. This survey is a particularly good choice for comparison as the birth years of its subjects (1957–1964) include those of the Perry sample (1957–1962). The NLSY79 also oversamples African-Americans.

The Matching Procedure We use a matching procedure to create NLSY79 comparison groups for Perry controls by simulating the application of the Perry eligibility criteria to the full NLSY79 sample. Specifically we use the Perry eligibility criteria to construct samples in the NLSY79. Thus, the comparison group corresponds to the subset of NLSY79 participants that would likely be eligible for the Perry program if it were a nationwide intervention.

We do not have identical information on the NLSY79 respondents and the Perry entry cohorts, so we approximate a Perry-eligible NLSY79 comparison sample. In the absence of IQ scores in the NLSY79, we use AFQT scores as a proxy for IQ. We also construct a pseudo-SES index for each NLSY79 respondent using the available information.[42]

We use two different subsets of the NLSY79 sample to draw inferences about the representativeness of the Perry sample. For an initial comparison group, we use the full African-American subsample in NLSY79. We then apply the approximate Perry eligibility criteria to create a second comparison group based on a restricted sub-sample of the NLSY79 data. Comparability in later life outcomes between the restricted group and the Perry control group suggests that the Perry sample, while not necessarily representative of the African-American population as a whole, is representative of a particular subsample of that population. Specifically, this subsample reflects the eligibility requirements of the Perry program, such as low IQ of the child and a low parental SES index.

The US population in 1960 was 180 million people, of which 10.6% (19 million) were black.[43] We use the NLSY79, a representative sample of the total population that was born between 1957 and 1964, to estimate the number of persons in the US that resemble the Perry population at entry (age 3). According to the NLSY79, the black cohort born in 1957–1964 is composed of 2.2 million males and 2.3 million females. We

[42] For details, see the Web Appendix `http://jenni.uchicago.edu/Perry/cost-benefit/reanalysis`

[43] Visit: `http://www.census.gov/population/www/documentation/twps0056/twps0056.html` for more details.

estimate that 17% of the male cohort and 15% of the female cohort would be eligible for the Perry program if it were applied nationwide. This translates into a population estimate of 712,000 persons out of this 4.5 million black cohort resemble the Perry population.[44] For further information on the comparison groups and their construction, see Web Appendix J and Tables J.1 and J.2 for details.

How Representative is the Perry Sample of the Overall African-American Population of the US? Compared to the unrestricted African-American NLSY79 subsample, Perry program participants are more disadvantaged in their family backgrounds. This is not surprising given that the Perry program was targeted toward disadvantaged children. Further, Perry participants experience less favorable outcomes later in life, including lower high school graduation rates, employment rates, and earnings. However, if we impose restrictions on the NLSY79 subsample that mimic the sample selection criteria of the Perry program, we obtain a roughly comparable group. Figure 5 demonstrates this comparability for parental highest grade completed at the time children are enrolled in the program. Web Appendix Figures J.1-J.5 report similar plots for other outcomes, including mother's age at birth, earnings at age 27 and earnings at 40.[45] Tables J.1–J.2 present additional detail. The Perry sample is representative of disadvantaged African-American populations.

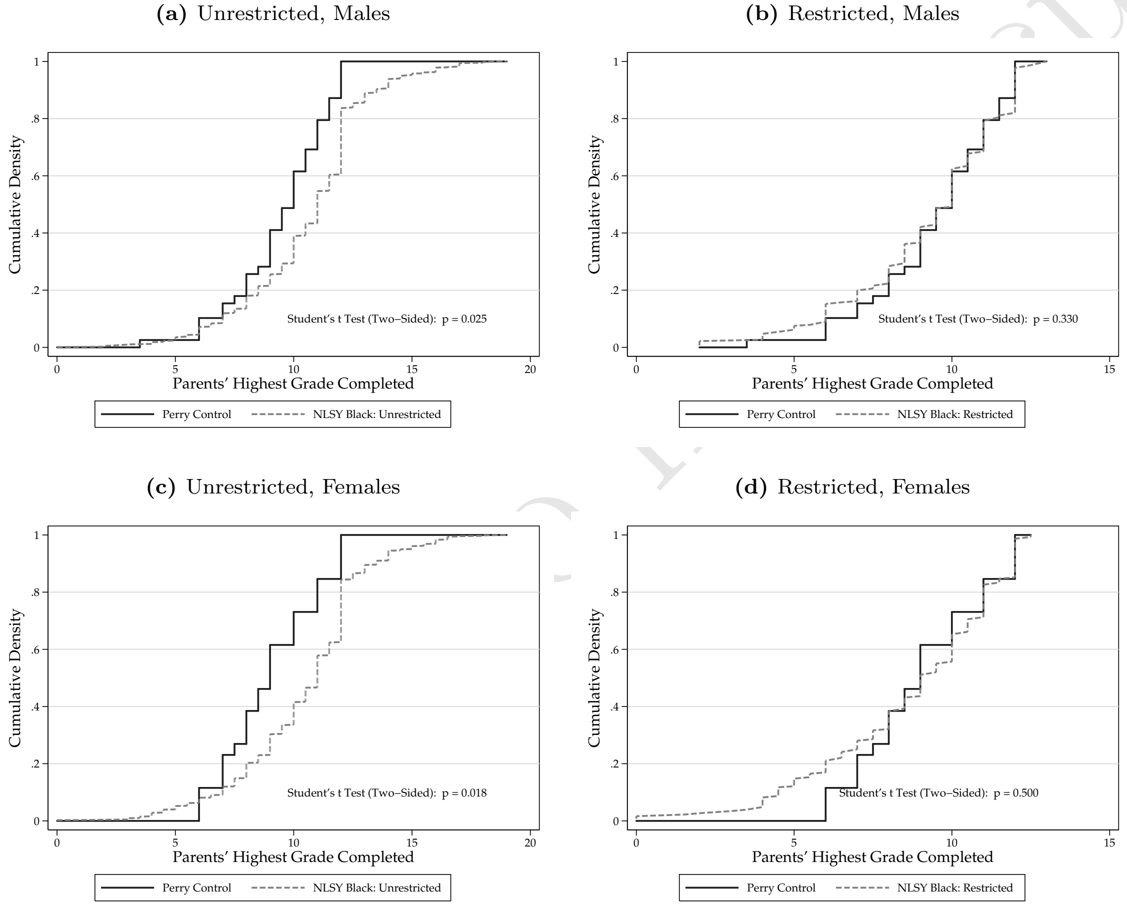
In Web Appendix K, we consider another aspect of the external validity of the Perry experiment. Perry participants were caught up in the boom and bust of the Michigan auto industry and its effects on related industries. In the 1970s, as Perry participants entered the workforce, the male-friendly manufacturing sector was booming. Employees did not need high school diplomas to get good entry-level jobs in manufacturing. The industry began to decline as Perry participants entered their late 20s and men were much more likely than women to be employed in the manufacturing sector.

This pattern may explain the gender patterns for treatment effects found in the Perry experiment. Neither treatments nor controls needed high school diplomas to get good jobs. As the manufacturing sector collapsed, neither group fared well. However, as noted in Web Appendix K, male treatment group members were more likely to adjust to economic adversity by migrating than were controls, which may account for their greater economic success at age 40. The economic history of the Michigan economy may play an important role in explaining the age pattern of observed treatment effects for males, thereby diminishing the external validity of the study.

[44] When a subsample of the NLSY79 is formed using three criteria that characterize the Perry sample — low values of a proxy for the Perry socio-economic status (SES) index, low achievement test (AFQT) score, and non-firstborn status — this subsample represents 713,725 people in the U.S. See Web Appendix J and Tables J.1 and J.2 for details.

[45] One exception to this pattern is that Perry treatment and control earnings are worse off than their matched sample counterparts.

Figure 5: Perry vs. NLSY79: Mean Parental Highest Grade Completed



(a) Unrestricted, Males

(b) Restricted, Males

(c) Unrestricted, Females

(d) Restricted, Females

Notes: Unrestricted NLSY79 is the full black subsample. Restricted NLSY79 is the black subsample limited to those satisfying the approximate Perry eligibility criteria: at least one elder sibling, Socio-economic Status (SES) index at most 11, and 1979 AFQT score less than the black median.

7 Comparison to Other Analyses

We compare the approach used in this paper to that used in two other studies. Schweinhart et al. (2005) analyze the Perry data through age 40 using large sample statistical tests. They do not account for the compromised randomization of the experiment, or the multiplicity of hypotheses tested. Heckman (2005) sounds a warning note about the potential problem of selectively reporting "significant" effects from a large collection of possible effects without adjusting the p-values for the multiplicity of hypotheses selected.

Anderson (2008) applies a multiple-inference procedure due to Westfall and Young (1993) to three early intervention experiments: the well-known Abecedarian Project (Campbell and Ramey, 1994), the Perry Preschool program, and the Early Training Project (Gray and Klaus, 1970). However, he ignores the problem of compromised randomization and does not correct for covariate imbalances.[46,47]

To reduce the dimensionality of the testing problem, Anderson creates linear indices of outcomes at three stages of the life cycle for treated and controlled persons. For each study, the outcomes used to construct the index are the same for both gender groups but the weights depend on gender.[48] Different outcomes are used at different stages of the life cycle. Across studies, an attempt is made to use "comparable" outcome measures but no evidence on the comparability of the measures is presented in his paper. The outcomes used to construct each index are quite diverse and group a variety of very different outcomes (e.g., crime, employment, education). The populations treated are also diverse in terms of the background of participants and controls. In addition, the treatments given are very different across studies. No adjustment is made for differences in populations served or services offered across programs. Anderson uses his constructed indices to test for gender differences within and across programs and reports evidence that the Perry program does not "work" for boys. Since the programs compared are very different in ways he does not adjust for, it is difficult to interpret his cross-program comparisons.

His indices also lack interpretability. He does not use a monetary metric like the rate of return or the benefit cost ratio as do Heckman, Moon, Pinto, Savelyev, and Yavitz (2009).[49] An alternative interpretable metric—the effect of programs on cognitive and noncognitive skills—is studied in Heckman, Pinto, and Savelyev (2009). All of our papers differ from Anderson (2008) in finding that Perry improved the status of both genders on a variety of measures.

[46] The Westfall and Young procedure he uses assumes subset pivotality (see Appendix F for a definition). This is a strong assumption that is not required in the Romano Wolf (2005) procedure that we employ. Subset pivotality assumes that the distribution of test statistics in a subset of hypotheses is invariant to the truth or falsity of hypotheses in a larger set of hypotheses that contains the set of hypotheses being tested. Appendix F.3 presents an example for a commonly encountered testing problem where the condition is violated. Romano and Wolf (2005) provide other examples.

[47] Anderson makes a mistake in applying the Westfall-Young procedure. The mistake leads him to *understate* true p-values. See Appendix F.3.

[48] Following O'Brien (1984), weights are constructed to minimize the variance of the created index.

[49] A leading economist in the field of child development has recently urged developmental psychologists to move beyond "effect" sizes to consider rates of return and benefit-cost ratios (Duncan and Magnuson, 2007).

8 The Matching Assumption

In this paper, we account for imbalance in the covariates and compromised randomization by assuming conditional (on X) exchangeability and the partial linearity of each outcome within sub-samples defined by values of baseline measures. This is a matching assumption.

Matching is often criticized when used in non-randomized evaluations because the proper conditioning set is not in general known. Augmenting or decreasing the conditioning information is not guaranteed to produce conditional independence between treatment assignment D and outcomes (Y_1, Y_0). Without invoking further assumptions, there is no objective principle for determining which set of measures X will satisfy the assumption of conditional independence, $(Y_1, Y_0) \perp\!\!\!\perp D \mid X$, used in matching.[50] For Perry, the X that we use are known to be ones that affected assignment to treatment, even though the exact treatment assignment rule is unknown (see Subsection 4.2).

In related work, Heckman, Pinto, Shaikh, and Yavitz (2009) take a more conservative approach to the problem of compromised randomization using weaker assumptions. Their inference is based on a partially identified model in which the distribution of D conditional on X is not fully known because $\mathbf{M}$ is not fully determined. Unmeasured variables determining assignment may also affect outcomes. Their inference procedure uses a worst-case scenario for rejecting the null hypothesis whenever there is uncertainty about the distribution of D conditional on X. In doing so, they estimate conservative bounds for inference on treatment effects that are consistent with the available documentation of the protocol.[51]

The current paper is less conservative because it adopts stronger assumptions: conditional exchangeability of treatment assignments within coarse strata of pre-program X and assumes a linear relationship between some pre-program measures and outcomes. As expected, this less conservative approach results in sharper conclusions, although there is still surprisingly broad agreement in the inference generated from these two approaches.

9 Conclusion

Proper analysis of the Perry experiment presents many statistical challenges. These challenges include small-sample inference, accounting for imperfections in randomization, and accounting for large numbers of outcomes. The last of these refers to the risk of selecting statistically significant outcomes that are "cherry picked" from a larger set of unreported results.

We propose and implement a combination of methods to account for these problems. We control for the

[50] See the discussion of these aspects of matching in Heckman and Navarro (2004). See also Heckman and Vytlacil (2007).

[51] We discuss their approach formally in Web Appendix L.

violations of the initial randomization protocol and imbalanced background variables. We estimate family-wise error rates that account for the multiplicity of the outcomes. We consider the external validity of the program. The methods developed and applied here have applications to many social experiments with small samples when there is imbalance in covariates between treatments and controls, reassignment after randomization, and numerous multiple hypotheses.

Our analysis is the first to study the criteria used in the Perry randomization protocol and to control for the compromise in the randomization as implemented. We devise and implement a resampling method that mimics the treatment assignment distribution actually used.

The pattern of treatment response by gender varies with age. Males exhibit statistically significant treatment effects for criminal activity, later life income, and employment (ages 27 and 40), whereas, female treatment effects are strongest for education and early employment (ages 19 and 27). The general pattern is one of strong early results for females, with males catching up later in life.

Our analysis of external validity shows that Perry families are disadvantaged compared to the general US black population. However, the application of the Perry eligibility rules to the NLSY79 yields a substantial population of comparable individuals. Based on the NLSY79 data, we estimate that 712,000 persons in the US resemble the Perry population—about 16% of the black population born in 1957–1964, the birth years of the Perry participants.

The estimated rate of return to the Perry program is in the range of 6–10% for both boys and girls. This is on par with the historical rate of return to equity. Our estimates are, however, well below the estimates of 16-17% reported in the literature.

In summary, our analysis shows that accounting for corrupted randomization, multiple-hypothesis testing and small sample sizes, there are strong effects of the Perry Preschool program on the outcomes of boys and girls. However, there are important differences by age in the strengths of treatment effects by gender.

References

Anderson, M. (2008, December). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool and early training projects. *Journal of the American Statistical Association 103*(484), 1481–1495.

Anderson, M. J. and P. Legendre (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation 62*, 271–303.

Anderson, M. J. and J. Robinson (2001, March). Permutation tests for linear models. *The Australian and New Zealand Journal of Statistics 43*(1), 75–88.

Belfield, C. R., M. Nores, W. S. Barnett, and L. Schweinhart (2006). The High/Scope Perry Preschool program: Cost-benefit analysis using data from the age-40 followup. *Journal of Human Resources 41*(1), 162–190.

Campbell, F. A. and C. T. Ramey (1994, April). Effects of early intervention on intellectual and academic achievement: A follow-up study of children from low-income families. *Child Development 65*(2), 684–698. Children and Poverty.

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.

Cunha, F., J. J. Heckman, L. J. Lochner, and D. V. Masterov (2006). Interpreting the evidence on life cycle skill formation. In E. A. Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education*, Chapter 12, pp. 697–812. Amsterdam: North-Holland.

Dasgupta, P., K.-G. Mäler, and S. Barrett (2000). Intergenerational equity, social discount rates and global warming. Unpublished manuscript, Department of Economics, University of Cambridge. Revised version of the paper with the same title that was published in *Discounting and Intergenerational Equity*, (Washington, DC: Resources for the Future, 1999).

DeLong, J. and K. Magin (2009, Winter). The U.S. equity return premium: Past, present and future. *Journal of Economic Perspectives 23*(1), 193208.

Duncan, G. J. and K. Magnuson (2007). Penny wise and effect size foolish. *Child Development Perspectives 1*(1), 46–51.

Freedman, D. and D. Lane (1983, October). A nonstochastic interpretation of reported significance levels. *Journal of Business and Economic Statistics 1*(4), 292–298.

Gray, S. W. and R. A. Klaus (1970). The early training project: A seventh-year report. *Child Development 41*(4), 909–924.

Hanushek, E. and A. A. Lindseth (2009). *Schoolhouses, Courthouses, and Statehouses: Solving the Funding-Achievement Puzzle in America's Public Schools.* Princeton, NJ: Princeton University Press.

Hayes, A. (1996, June). Permutation test is not distribution-free: Testing $h_0 : \rho = 0$. *Psychological Methods 1*(2), 184–198.

Heckman, J. J. (2005). Invited comments. In L. J. Schweinhart, J. Montie, Z. Xiang, W. S. Barnett, C. R. Belfield, and M. Nores (Eds.), *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40*, pp. 229–233. Ypsilanti, MI: High/Scope Press. Monographs of the High/Scope Educational Research Foundation, 14.

Heckman, J. J., L. Malofeeva, R. Pinto, and P. A. Savelyev (2009). The effect of the Perry Preschool Program on the cognitive and non-cognitive skills of its participants. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Q. Yavitz (2009). The rate of return to the Perry Preschool program. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J. and S. Navarro (2004, February). Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics 86*(1), 30–57.

Heckman, J. J., R. Pinto, and P. A. Savelyev (2009). The noncognitive determinants of achievement test scores. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J., R. Pinto, A. M. Shaikh, and A. Yavitz (2009). Compromised randomization and uncertainty of treatment assignments in social experiments: The case of Perry Preschool Program. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J. and J. A. Smith (1995, Spring). Assessing the case for social experiments. *Journal of Economic Perspectives 9*(2), 85–110.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, pp. 4875–5144. Amsterdam: Elsevier.

Herrnstein, R. J. and C. A. Murray (1994). *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.

Kurz, M. and R. G. Spiegelman (1972). *The Design of the Seattle and Denver Income Maintenance Experiments*. Menlo Park, CA: Stanford Research Institute.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses* (Third ed.). New York: Springer Science and Business Media.

Micceri, T. (1989, January). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin 105*(1), 156–166.

O'Brien, P. C. (1984, December). Procedures for comparing samples with multiple endpoints. *Biometrics 40*(4), 1079–1087.

Reynolds, A. J. and J. A. Temple (2008). Cost-effective early childhood development programs from preschool to third grade. *Annual Review of Clinical Psychology 4*(1), 109–139.

Rolnick, A. and R. Grunewald (2003). Early childhood development: Economic development with a high public return. Technical report, Federal Reserve Bank of Minneapolis, Minneapolis, MN.

Romano, J. P. and M. Wolf (2005, March). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association 100*(469), 94–108.

Schweinhart, L. J., H. V. Barnes, and D. Weikart (1993). *Significant Benefits: The High-Scope Perry Preschool Study Through Age 27.* Ypsilanti, MI: High/Scope Press.

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.* Ypsilanti, MI: High/Scope Press.

The Pew Center on the States (2009, March). The facts. Response to ABC News Segements on Pre-Kindergarten. Available online at: http://preknow.org/documents/the_facts.pdf. Last accessed March 24, 2009.

Weikart, D. P., J. T. Bond, and J. T. McNeil (1978). *The Ypsilanti Perry Preschool Project: Preschool Years and Longitudinal Results Through Fourth Grade.* Ypsilanti, MI: Monographs of the High/Scope Educational Research Foundation.

Westfall, P. H. and S. S. Young (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* John Wiley and Sons.