Evaluation of Small-Sample Compromised Randomization: Long-Term Effects of Early Childhood Intervention on Health and on Addictive Behavior*

Rodrigo Pinto**

Abstract

Barros's contribution to the literature of policy evaluation stems from advances in statistical methods applied to rigorous empirical analysis. Another central question of Barros's research is the analysis of efficient policies for reducing economic inequality. By pursuing this question, Barros's line of work has recently shifted towards the study of early childhood investment as a tool for promoting economic growth and improving the odds of children born in disadvantaged families (Barros and Olinto, 2008, Barros et al., 2011). We follow Barros's steps by developing a formal statistical evaluation of the Perry preschool program, the oldest and most cited early childhood experiment in the U.S. The evaluation of the Perry program poses three statistical challenges. First, the small sample size casts doubts on standard inference methods that are based on asymptotic assumptions. Second, compromised randomization calls into question the validity of the simple statistical procedures often applied to social experiments. Third, the large number of outcome variables gives rise to the danger of selectively reporting significant estimates. We develop a statistical method that accounts for all these problems and that is tailored to the problems we face in the Perry intervention. We focus on their long-term impact on health and addictive behavior variables. We use new data through age 40, which had never been analyzed before. We find that treated females have fewer negative effects of drug/alcohol use on a range of later-life activities. For males, we find that treated participants use fewer hard drugs, such as heroin and hashish.

Keywords: Early Childhood Intervention; Preschool Randomization; Social Experiment; Multiple Hypothesis Testing.

JEL Codes: I21, C93, J15, V16.

*Submitted in October 2011. Revised in November 2011. Rodrigo Pinto is a MacArthur fellow at the economics department of the University of Chicago. This paper is based on a previous work of Heckman et al. (2010b). I thank James Heckman for invaluable collaboration which greatly strengthened the analysis and ideas of this paper. In addition, I thank Seong Moon and Azeem Shaikh for productive discussions on the topic of this research. This research was supported in part by the American Bar Foundation, 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, Robert Dugger, and NICHD R01HD043411. The views expressed in this presentation are those of the author and not necessarily those of the funders listed here.

**The University of Chicago. E-mail: rodrig@uchicago.edu

Brazilian Review of Econometrics
v. 30, n° 2, pp. 451–485 November 2010

1. Introduction

Barros's contribution to the literature of policy evaluation relies on two major lines of work: (1) the study of policy efficacy in reducing economic inequality and (2) rigorous methodological advances in policy evaluation.

Recently, Barros's line of work has shifted towards the analysis of the early years of education (Barros and Olinto, 2008, Barros et al., 2011). Barros's papers examine the impact of early childhood investments on child development. His research adds to a growing literature that unravels the importance of early years of life on the evolution of human skills. Examples of such literature are Heckman et al. (2010a) and Heckman et al. (2010b). We follow Barros's steps by evaluating an important early childhood intervention. Specifically, we target the Perry preschool program, the oldest and most cited early childhood experiment in the U.S. We follow Barros's line of work by providing a statistical analysis that is both formal and relevant to the statistical questions we face.

The Perry Preschool program is an early childhood intervention that provided preschool education to low-IQ, disadvantaged African-American children living in Ypsilanti, Michigan. It was conducted in the 1960s and participants were followed from ages 3 to 40. Perry long-term effects constitute a cornerstone for the ones that advocate the efficacy of investing in early childhood. In this paper, we focus on the long-term impact of early childhood intervention on health and on addictive behavior. To this end, we use new data through age 40, which had never been analyzed before.

Perry is a social experiment in which participants are randomly assigned to treatment and control groups. Even though social experiments that use the randomization method can generate valuable information about the effectiveness of interventions, they are usually plagued by three statistical problems. First, the randomization of social experiments is often compromised. This calls into question the validity of the simple statistical procedures often applied to analyze the Perry study. In addition, the sample size is small. This feature casts some doubts on the validity of inference methods that are based on asymptotic assumptions, which are accurate only for large datasets. Finally, Perry has an overwhelming number of outcomes. This creates the danger of selectively reporting "significant" effects from a large pool of possible effects, putting a downward bias on the reported p-values. This paper develops tools to overcome these three statistical problems.

This paper develops and applies small-sample permutation procedures aimed at testing hypotheses on samples obtained from a less-than-ideal randomization. We correct estimated treatment effects for imbalances produced by compromises of the initially intended randomization protocol, such as post-randomization reassignment. We further address the potential problem with the arbitrary selection of statistically significant outcomes by using a multiple-hypothesis testing based on the stepdown procedure (Romano and Wolf, 2005). In summary, we find that treated females have fewer negative effects of drug/alcohol use on their life activ-

ities and used fewer hard drugs. Our results show that the treatment had fewer pronounced impacts on males than on females. Still, treated males use fewer hard drugs than control ones.

This paper is organized as follows. Section 2 describes the Perry experiment. Section 3 discusses the statistical challenges posed by the analysis of the Perry experiment. Section 4 introduces the standard evaluation model. Section 5 presents our methodology for single hypothesis testing. Section 6 presents our methodology for multiple hypothesis testing. Our empirical analysis is presented in Section 7. Section 8 concludes.

2. The Perry Preschool Program Experiment and Curriculum

As mentioned in Heckman et al. (2010b), the HighScope Perry Program was a small sample experiment: 123 children allocated over five entry cohorts. The experiment was conducted from the early to mid-1960s in the district of the Perry Elementary School, a public school in Ypsilanti, Michigan. The data were surveyed at the onset of the intervention, at age 3, and annually until age 15. Additional follow-ups were conducted at ages 19, 27, and 40. The set of information comprises numerous measures such as economic, criminal, and educational outcomes and psychological measures on cognition and personality.

Preschool Overview The Perry experiment consisted of four waves. The first wave admitted 4-year-olds who only received one year of treatment. The last wave was taught alongside a group of 3-year-olds who are not included in the Perry data but were used to provide a uniform treatment. The preschool class consisted of 20–25 children aged 3 to 4 years. Classes were 2.5 hours every weekday during the regular school year (mid-October through May). The experiment was based on a low children to teacher ratio ranging from 5 to 6.25 over the course of the program. The staff consisted of former public school teachers especially trained to assist disadvantaged children. They were certified in elementary, early childhood, and special education (Schweinhart et al., 1993, p. 32).

Home Visits The Perry experiment also had a home visit component. The intervention was based on weekly home visits lasting 1-1/2 hours conducted by the preschool teachers. Home visits intended to involve the mother in the educational process and advance the implementation of the Perry curriculum in their home, (Schweinhart et al., 1993, p. 32). Teachers also helped with problems that arose in the family setting of the participant's home. Field trips to stimulating environments, such as a zoo, were also conducted.

Curriculum The Perry experiment differs from standard early childhood interventions by focusing on the enhancements of noncognitive abilities instead of

cognitive ones (or language). The Perry Preschool curriculum was based on the concept of *active learning*. Activities are based on problem-solving and guided by open-ended questions. In other words, the topics in the curriculum were not based on specific facts or topics, but rather on *key developmental factors* related to planning, expression, and understanding. (Schweinhart et al., 1993) explain that children were encouraged to plan, carry out, and then reflect on their own activities. These factors were then organized into 10 topical categories, such as "creative representation," "classification" (recognizing similarities and differences), "number," and "time." These educational principles were reflected in the types of open-ended questions asked by teachers: for example,"What happened? How did you do that? Can you show me? Can you help another child?" (Schweinhart et al., 1993, p. 33).

As the curriculum was developed over the course of the program, its details and application varied. While the first year involved "thoughtful experimentation" on the part of the teachers, experience with the program and a series of seminars during subsequent years led to the development and systematic application of teaching principles with "an essentially Piagetian theory-base." During the later years of the program, all activities took place within a structured daily routine intended to help children "to develop a sense of responsibility and to enjoy opportunities for independence" (Schweinhart et al., 1993, p. 32–33).

Eligibility Criteria The target population of the intervention consisted of families living in the surrounding area of the Perry Elementary School. The families considered for the intervention were poorer than the average disadvantaged African-American families in the U.S. at that time. Nevertheless, the target population was representative of a large segment of the disadvantaged African-American population. In terms of external validity analysis, Heckman et al. (2010b) explains that if the Perry program were applied nationwide at the time of the onset, 17% of the male cohort and 15% of the female cohort would be eligible for the Perry program.

Randomization Protocol The Perry randomization protocol was not simple. Weikart et al. (1978) have a detailed description of the randomization protocol. Families were assigned to treatment and control groups on the basis of their pre-program variables for each designated eligible entry cohort. The pre-program variables used in the randomization protocol were: wave cohort, IQ at entry, socioeconomic status and maternal employability at the onset of the program. We refer to Heckman et al. (2010b) for a detailed explanation of the randomization protocol.

3. Statistical Challenges in Analyzing the Perry Program

Our aim is to draw valid inference from the Perry study. To do that, we have to solve three statistical challenges:

(1) Small sample size;

(2) Compromise in the randomization protocol; and

(3) the large number of outcomes.

The multiplicity of outcomes creates the danger of selectively reporting significant treatment effects out of a large pool of outcomes, which generates downward biased p-values.

The small sample size of the Perry study casts some doubts on the validity of classical tests. These tests are based on central limit theorems that are only valid when the number of participants tends to infinity. Thus, they produce inferences based on p-values that are only asymptotically valid. A huge strand of the literature demonstrates that classical testing procedures can be unreliable when sample sizes are small and the data are non-normal.

The second problem refers to the compromising aspects of the randomization protocol. The randomization protocol implemented in the Perry study differs from the original one. Specifically, treatment and control statuses were reassigned to a subset of persons after being initially randomly allocated. Statistically speaking, reassignments can induce correlation between treatment assignment and baseline characteristics of participants, which, in turn, can violate the assumption of independence between treatment assignment and outcomes in the absence of treatment effects.

It is important to distinguish the correlation induced by compromises from traditional sampling variation. While sampling variation is important for increasing the inference power, sampling variation does not impair the validity of tests that do not correct for it. Compromised randomization, on the other hand, can generate biased inferences. We control for this problem by conditioning on the variables used in the randomization protocol.

The existence of numerous outcomes surveyed in the Perry experiment can lead to the selective reporting of statistically significant outcomes, as determined by usual single hypothesis tests, without correcting for the effects of such preliminary screening on actual p-values. To make this statement more precise, suppose that a single hypothesis test rejects a true null hypothesis at significance level α. Thus, the probability of rejecting a single hypothesis out of K true hypothesis is given by $1-(1-\alpha)^K$. As the number of outcomes K increases, the likelihood of rejecting a true null hypothesis departs from α. This practice is sometimes termed "cherry picking". Our solution to this potential problem is based on a multiple hypothesis testing called stepdown procedure described in Section 6.

4. The Basic Evaluation Model

A standard model of program evaluation describes the observed outcome Y of participant i by $Y_i = D_i Y_{i,1} + (1 - D)Y_{i,0}$, where $(Y_{i,1}, Y_{i,0})$ are potential outcomes corresponding to the outcome Y for agent i when treatment is *fixed* at treatment and control statuses, respectively. D_i denotes the assignment indicator: $D_i = 1$ if treatment occurs, $D_i = 0$ otherwise. The focus of this paper is on testing the null hypothesis of no treatment effect or, equivalently, that treatment and control outcome distributions are the same: $Y_{i,1} \stackrel{d}{=} Y_{i,0}$, where "$\stackrel{d}{=}$" denotes equality in distribution.

An evaluation problem arises in standard observational studies because either $Y_{i,1}$ or $Y_{i,0}$ is observed, but not both. As a result, in nonexperimental samples, the simple difference-in-means between treatment and control groups, $E\left(Y_{i,1}|D_i = 1\right) - E\left(Y_{i,0}|D_i = 0\right)$, is not generally equal to the average treatment effect, $E(Y_{i,1} - Y_{i,0})$, or to the treatment effect conditional on participation, $E\left(Y_{i,1} - Y_{i,0} \mid D_i = 1\right)$. Bias can arise from participant self-selection into the treatment group. Rigorous analysis of treatment effects distinguishes impacts due to participant characteristics from impacts due to the program itself.

Randomized experiments solve the *selection bias* problem by inducing independence between (Y_0, Y_1) and D, interpreted as a treatment assignment indicator, $(Y_0, Y_1) \perp\!\!\!\perp D$, where Y_1, Y_0, and D are vectors of pooled variables across participants and $\perp\!\!\!\perp$ denotes independence. Selection bias can be induced by *randomization compromises*, which occur when the implemented randomization differs from an ideal randomization protocol in a way that threatens the statistical independence of treatment assignments D and the joint distribution of counterfactual outcomes (Y_0, Y_1).

Perry randomization was compromised by the reassignment of treatment and control labels after initial draws produced an imbalanced distribution of pre-program variables. This creates a potential for biased inference, as described in the previous subsection. It is important to distinguish sampling variation from compromises due to reassignment. While sampling variation denotes the imbalance of pre-program variables that can randomly occur, compromises due to reassignments generate a skewed distribution of pre-program variables that is not random. While controlling for sampling variation increases the power of the statistical inference, not controlling for it does not invalidate the inference. On the other hand, not controlling for reassignments generates biased inference.

5. Testing Methodology

This paper develops a framework for small-sample inference based on permutation testing conditional on a given sample. This section specifies our notation and the theoretical framework for our testing procedures. Our aim is to provide a statistical solution to the challenges described in Section 3.

5.1 Setup and Notation

General We use calligraphic capital letters to denote sets. Capital letters denote two different entities: either the maximum index of a set of natural numbers or random variables. The usage should be clear from context. We use lower case letters to index elements of sets. We represent a vector of pooled elements of a set with parentheses followed by its respective indexing. As an example let $[V_1, \ldots, V_N]$ be the N-dimensional vector V indexed by the set $\mathcal{V} = \{1, \ldots, N\}$, and let it be represented by $V \equiv (V_v; v \in \mathcal{V})$.

Treatment Assignment The set of indices of Perry participants is $\mathcal{I}$, where $\mathcal{I} = \{1, \ldots, I\}$ and $I = 123$. Let D_i be the treatment assignment for participant $i \in \mathcal{I}$, where $D_i = 1$ if i is treated and $D_i = 0$ otherwise. Let $D = (D_i; i \in \mathcal{I})$ be the vector of random assignments.

Outcomes and Hypotheses We represent outcome k by the random vector Y^k, which represents an I-dimensional vector of values of variables Y_i^k for participants i, $Y^k = (Y_i^k; i \in \mathcal{I})$. The index set of outcomes from 1 to K is represented by $\mathcal{K} = \{1, \ldots, K\}$. Our aim is to test the null hypothesis of no treatment effect for outcome Y^k. This hypothesis is written as $H_k : Y^k \perp\!\!\!\perp D$, that is, Y^k is independent of D. The joint null hypothesis of no treatment effect for outcomes $Y^k; \forall\, k \in \mathcal{K}$ is represented by $H_{\mathcal{K}} \equiv \cap_{k \in \mathcal{K}} H_k$.

Permutation A transformation of the vector of treatment assignments D that permutes the position of its elements is represented by gD and is defined as follows:

$gD = (\widetilde{D}_i; i \in \mathcal{I} | \widetilde{D}_i = D_{\pi_g(i)}$, where π_g is a permutation function (i.e., $\pi_g : \mathcal{I} \to \mathcal{I}$ is a bijection)).

The permutation function π_g is indexed by g. To simplify notation, we represent the permutation that acts on the data by g. This transformation can be applied to any data that is indexed by $\mathcal{I}$. We use the permutation over the treatment assignment D, where gD is the vector of permuted assignments. Equivalently, a permutation can be written as a linear transformation $gD \equiv B_g D$, where B_g is a permutation matrix[1] that swaps the elements of any variable D according to the permutation g.

The Randomization Hypothesis Permutation-based inference seeks to test the Randomization Hypothesis, which states that the joint distribution of some

[1] A permutation matrix A of dimension L is a square matrix $A \equiv (a_{ij}) : i, j = 1, \ldots, L$ where each row and each column has a single element equal to one and all other elements equal to zero within the same row or column. Formally, $a_{ij} \in \{0, 1\}$; $\sum_{j=1}^{L} a_{ij} = 1$, $\sum_{i=1}^{L} a_{ij} = 1$, for all i, j.

outcome Y is invariant under permutations $g \in \mathscr{G}$, i.e., that outcome distributions are invariant to the swap of its elements according to g. We represent the set of valid permutations for which the Randomization Hypothesis holds by $\mathscr{G}$, so $\forall g \in \mathscr{G}, (Y, gD) \stackrel{d}{=} (Y, D)$, where, as in the text, "$\stackrel{d}{=}$" means equality in distribution. A consequence of the randomization hypothesis is that for any statistic T of outcome Y and treatment status D, we have $T(Y, gD) \stackrel{d}{=} T(Y, D)$ whenever the hypothesis holds. Moreover, if T is invariant to the relative ordering of the pair (Y_i, D_i) in the vector (Y, D), then permuting Y instead of D generates the same distribution of the test statistic $T(Y, D)$. Stated differently, the distribution of the test statistic $T(Y, D)$ will not change if the outcome positions of some treated and control participants are swapped in accordance with permutations $g \in \mathscr{G}$. Equivalently, we can write $T(Y, D) \stackrel{d}{=} T(gY, D)$.

5.2 Conditional Exchangeability and Independence under the Randomization Hypothesis

An idealized randomization based on a fair coin is known to generate treatment assignments D that are unconditionally independent of outcomes Y and pre-program variables $X = (X_i, i \in \mathcal{I})$. When randomization is compromised, the randomization hypothesis must be altered to account for the failure of the unconditional independence between treatment assignments D and outcomes Y.

The randomization procedure in the Perry experiment is compromised by reassignment of treatment labels to balance pre-program variables across treatments and controls. The randomization protocol ranked children by IQ score and then allocated treatment status to either all odd-ranked or all even-ranked children and control status to the rest. Alterations to this basic assignment rule occurred from two types of treatment-assignment swaps between individuals. The first type of swap was intended to balance observable pre-program variables (namely, SES index and gender). The second type of swap was made after the designation of treatment status, and was intended to remove children with working mothers from the treatment group due to logistical problems associated with their participation in the treatment program. Compromises of the Perry randomization protocol embody both types of swaps. The latter compromises the independence between D and X, and may also create a potential dependence between treatment status D and some unobserved variables $V = (V_i; i \in \mathcal{I})$ as well.

Formally, treatment assignments can be said to have been generated by a randomization mechanism described by a deterministic function $\boldsymbol{M}$. The arguments of $\boldsymbol{M}$ are the variables that can affect treatment assignment. Define R as a random variable that describes the outcome of a randomization device (in the Perry study, the flip of a coin). Prior to determining the realization of R, two groups were formed on the basis of observed variables X (e.g., on IQ). Then R was realized by a randomization device. By construction, the distribution of R does not depend

on the composition of the two groups. After the realization of R, some individuals were swapped across initially assigned treatment groups based on some X values (e.g., mother's working status) and possibly on some unobserved variables V as well. By assumption, R is independent of (X, V), that is, $R \perp\!\!\!\perp (X, V)$. $\boldsymbol{M}$ captures all aspects of the treatment assignment mechanism. Perry randomization protocol can be represented by the following Assumptions **A-1** and **A-2**.

Assumption A-1 $D \sim \boldsymbol{M}(R, X) \ : \ \text{supp}(R) \times \text{supp}(X) \to \mathcal{D}\,;\ R \perp\!\!\!\perp X$, *where* $\text{supp}(D) = \mathcal{D}$, *and "*supp*" denotes support.*

We represent the unobserved variables that affect outcomes for participant i by V_i and the vector of unobserved variables by $V = (V_i\ ;\ i \in \mathcal{I})$. The independence between unobserved variables and treatment assignments is obtained by the act of randomization and can be stated as follows:

Assumption A-2 $R \perp\!\!\!\perp V$.

In other words, the random variables R are a device used to generate the independence of treatment status and unobserved variables V. When the intended randomization is compromised by reassignment based on observed pre-program variables X, the assignment mechanism depends on X. In this case, substantial correlation between the final compromised treatment assignments D and unobserved variables V can exist through the existing dependence between pre-program variables X and unobserved variables V.

The available information on the randomization protocol is prone to state that some participants had their initial treatment status reassigned in an effort to lower program costs. It is also known that only participants whose mothers were employed were swapped. We interpret the reassignments of the randomization protocol as being random *conditional on maternal working status*. We represent the mechanism of treatment assignment by $\boldsymbol{M}$ as defined in assumption **A-1**, whose arguments are known and observed.

As a concrete example, suppose that there was only one child per family in Perry and there were no swaps after initial ranking by IQ score. Denote $\widetilde{IQ}$ as vector of indicator variables equal to 1 for odd-ranked IQs within each wave. The Perry treatment assignment mechanism is characterized as

$$D = \sum_{w=1}^{5} \mathbf{1}[W = w] \odot \left(\mathbf{1}[\widetilde{IQ} = 1] \cdot b_w + \mathbf{1}[\widetilde{IQ} = 0] \cdot (1 - b_w) \right),$$

where $(b_1, \ldots, b_5)$ are independent Bernoulli random variables representing the outcomes of the coin toss used to assign treatment status after the initial IQ-score ranking and $\odot$ is a Hadamard product.[2]

[2]This is an element-wise product.

Assumption **A-2** simply states that the randomization procedure is not based on unobserved variables V. If unobserved variables V were not used to assign treatment status, then the relevant information on (X, V) can be represented by the observed characteristics X. Program participants are characterized by (X, V). X, V, and R generate D. Any permutation g of the elements in (X, V), conditioned on R, generates the same permutation of D:

$$(\boldsymbol{M}(g(X, V), R) = gD)|R. \tag{1}$$

This logic leads to the following proof of the exchangeability of treatment assignments, conditional on X.

Theorem 5.1 *Treatment assignments D are exchangeable for participants with the same X if the randomization does not rely on the unobserved variable V of the participants.*

Proof Let $\mathscr{G}_X$ be the set of permutations among participants with the same X. In this case, $gX = X \ \forall \ g \in \mathscr{G}_X$. By assumption, $D = \boldsymbol{M}(R, X)$, so for $\forall \ g \in \mathscr{G}_X$,

$$\begin{aligned} \mathrm{P}(D \in A) &= \mathrm{E}\left(\mathrm{E}\left(\mathbf{1}[\boldsymbol{M}(R, X) \in A]|R\right)\right) \\ &= \mathrm{E}\left(\mathrm{E}\left(\mathbf{1}[\boldsymbol{M}(R, gX) \in A]|R\right)\right) \\ &= \mathrm{E}\left(\mathrm{E}\left(\mathbf{1}[gD \in A]|R\right)\right) \\ &= \mathrm{P}(gD \in A), \end{aligned}$$

where $\mathscr{G}_X$ is defined by:

$$\mathscr{G}_X = \left\{g, \pi_g : \mathcal{I} \to \mathcal{I}; X_i = X_{\pi_g(i)}, \forall i \in \mathcal{I}\right\}.$$

An important feature of the exchangeability property of treatment status is that it relies on limited information on the randomization protocol. Exchangeability 5.1 does not require a full specification of the distribution D nor the assignment mechanism $\boldsymbol{M}$. Instead, it only requires information on which variables were used in the randomization protocol. A useful consequence of this weak requirement is that D-exchangeability remains valid under any randomization compromises that are based only on the information contained in X. More importantly, if the randomization protocol is known to be compromised, then valid exchangeable properties exist if compromises are based on observed variables.

Conditional Independence The goal of a randomization trial is to solve the problem of selection bias by inducing independence between unobserved variables and treatment status. As a consequence, we can show that treatment assignment D is independent of counterfactual outcomes (Y_0, Y_1), conditional on X. This statistical property comes from the observation that R is independent of

(Y_0, Y_1) by construction. The following theorem proves the conditional independence $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$, assuming that D is generated by (R, X) via $\boldsymbol{M}$ and that the X are observed:

Lemma L-1. Under assumptions **A-1** and **A-2**, $(Y_1, Y_0) \perp\!\!\!\perp D \mid X$.

Proof

$$
\begin{aligned}
& (Y_1, Y_0) \perp\!\!\!\perp R \mid X && \text{(by Assumption } \mathbf{A\text{-}2}\text{)} \\
\Rightarrow\ & (Y_1, Y_0) \perp\!\!\!\perp \phi(R) \mid X && \text{(for any particular function } \phi\text{)} \\
\Rightarrow\ & (Y_1, Y_0) \perp\!\!\!\perp \boldsymbol{M}(R, X) \mid X && \text{(by Assumption } \mathbf{A\text{-}1}\text{)} \\
\therefore\ & (Y_1, Y_0) \perp\!\!\!\perp D \mid X.
\end{aligned}
$$

Lemma L-1 is a consequence of assumptions **A-1**, **A-2**. Conditional on X, the arguments that determine $Y_{i,d}$ for $d \in \{0, 1\}$ are the unobserved variables V, which is independent of R by assumption **A-2**. Moreover, R is independent of (Y_0, Y_1). Thereby, any function of R and X is also independent of (Y_0, Y_1) conditional on X. Thus, assumption **A-1** states that conditional on X, treatment assignments depend only on R, so the counterfactual outcomes $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$.[3]

In the same fashion of the exchangeability property, Lemma L-1 is valid whenever the randomization protocol is based on observed variables X, but not on V.

Biased selection can occur in the context of a randomized experiment if treatment assignment is generated on the basis of variables that are not observed and are correlated with outcomes. If protocol $\boldsymbol{M}$ is based in part on an unobserved variable V that impacts Y:

Assumption A-1′ $\boldsymbol{M}(R, X, V) : \text{supp}(R) \times \text{supp}(X) \times \text{supp}(V) \to \mathcal{D}$.

5.3 Testing the Null Hypothesis of No Treatment Effect

Our goal is to test the conditional null hypothesis of no treatment effect, that is, if control and treated outcome vectors share the same distribution conditioned on pre-program variables X. Formally, we write:

Hypothesis H-1 $Y_1 \stackrel{d}{=} Y_0 \mid X$,

where $\stackrel{d}{=}$ denotes equality in distribution.

The hypothesis of no treatment effect can be restated in an equivalent form. Under Lemma L-1, Hypothesis **H-1** is equivalent to

Hypothesis H-1′ $Y \perp\!\!\!\perp D \mid X$.

[3] By the same reasoning, if we make Z explicit, we can also use **A-2** to show that $Z \perp\!\!\!\perp D \mid X$.

The equivalence is demonstrated in Heckman et al. (2010b). We restate their arguments for the sake of completeness. Let A_J denote a set in the support of a random variable J. Then

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

We refer to hypotheses **H-1** and **H-1′** interchangeably throughout this paper.

The compromising aspects of the randomization protocol can be accounted for by conditioning on the known set of variables that were used in the actually implemented randomization protocol. Conditioning on X corrects for the randomization compromises because it corrects for the induced correlation of D and Y via X. In other words, compromising aspects of the randomization imply that treatment assignments D are not independent of pre-program variables X. A statistical dependence between Y and D may be induced via X if variables X impact outcomes. In this case, the impact of D on Y may reflect the induced correlation of D and X instead of treatment effects themselves. Conditioned on X, treatment status D and outcomes Y must be independent under the hypothesis of no treatment effects.

In summary, under our maintained assumptions and compromised randomization, $(Y_1,Y_0)\perp\!\!\!\perp D\mid X$ holds, but $(Y_1,Y_0)\perp\!\!\!\perp D$ may not. Thus, a natural way to test the null hypothesis is to condition on X as stated in Hypotheses **H-1** and **H-1′**.

5.4 Useful Exchangeability Properties for Testing Procedures

The mechanics of testing the hypothesis $Y\perp\!\!\!\perp D\mid X$ rely on the exchangeability properties of the joint distribution (Y,D). The following theorem shows that the joint distribution of (Y,D) is invariant across the set of permutations $\mathscr{G}_X$ that swap treatment assignments D within the same strata of X values, $(Y,D)\overset{d}{=}(Y,gD)$.

Theorem 5.2 *Suppose that the randomization is as described in Theorem 5.1. Under Hypothesis* ***H-1****, the joint distribution of outcomes Y and treatment assignments D are invariant under permutations $\mathscr{G}_X$ of treatment assignments within strata formed by values of X:* $(Y,D)\overset{d}{=}(Y,gD)\ \forall\ g\in\mathscr{G}_X$.

Proof Let $\mathscr{G}_X$ be the set of permutations within participants that share the same data on X. Then, by Theorem 5.1, $D\overset{d}{=}gD$. Moreover, Lemma L-1 shows that

$(Y_1, Y_0) \perp\!\!\!\perp D \mid X$. Thus, for all $g \in \mathscr{G}_X$ we can write

$$\begin{aligned}
\mathrm{P}(\,(Y, gD) \in (A_Y, A_D)|X) &= \mathrm{E}\left(\mathbf{1}[Y \in A_Y] \odot \mathbf{1}[gD \in A_D]|X\right) \\
&= \mathrm{E}\left(\mathbf{1}[D \odot Y_1 + (1-D) \odot Y_0 \in A_Y]\right. \\
&\left.\odot\, \mathbf{1}[gD \in A_D]|X\right) \\
&= \mathrm{E}\left(\mathbf{1}[Y_0 \in A_Y] \odot \mathbf{1}[gD \in A_D]|X\right) \\
&\qquad \text{by } Y_{i,1} \stackrel{d}{=} Y_{i,0} \quad \forall i \in \mathcal{I}, \\
&\text{due to Hypothesis } \textbf{H-1} \\
&= \mathrm{E}\left(\mathbf{1}[Y_0 \in A_Y]|X\right) \odot \mathrm{E}\left(\mathbf{1}[gD \in A_D]|X\right) \\
&\qquad \text{by } (Y_1, Y_0) \perp\!\!\!\perp D \mid X \\
&= \mathrm{E}\left(\mathbf{1}[Y_0 \in A_Y]|X\right) \odot \mathrm{E}\left(\mathbf{1}[D \in A_D]|X\right) \\
&\qquad \text{by Theorem 5.1, } D \stackrel{d}{=} gD \\
&= \mathrm{P}(Y \in A_Y|X)\,\mathrm{P}(D \in A_D|X) \\
&= \mathrm{P}((Y, D) \in (A_Y, A_D)|X) \\
&\qquad \text{by } Y \perp\!\!\!\perp D|X.
\end{aligned}$$

One particular consequence of $(Y, D) \stackrel{d}{=} (Y, gD)$ affects the use of test statistics. As mentioned, if a test statistic relies only on the relationship between D and Y (that is, (Y_i, D_i), regardless of its rank position in the matrix (Y, D)), then permuting D is equivalent to permuting Y for testing purposes. For example, suppose we test using Student's t. Then the value of the t-statistics computed after a permutation of two elements of D is the same as if we had permuted the associated elements of Y instead. Put another way, using (gY, D) instead of (Y, gD) would provide equivalent inference results in this setting.

5.5 Restricted Permutation Groups and Sampling

Under the Randomization Hypothesis of no treatment effect, outcomes for treatments and controls are exchangeable within each stratum $X = x$. This section formally defines the procedure.

Partitioning the Data Suppose without loss of generality that the data on the pre-program variables X take on J distinct values, say $\{a_1, a_2, \ldots, a_J\}$. Let the index set $\mathcal{I}$ for participants be partitioned into J disjoint sets $\mathcal{I}_j$ and let $j \in \mathcal{J} \equiv \{1, \ldots, J\}$ where each set $\mathcal{I}_j$ is defined by the set of participants that share the same value a_j for pre-program variables X. Recall that x_i is the value of the pre-program variable X for participant i. We can define $\mathcal{I}_j$ by:

$$\mathcal{I}_j \equiv \{i \in \mathcal{I};\, x_i = a_j\}.$$

By definition, the union of the disjoint sets $\mathcal{I}_j$ over $j \in \mathcal{J}$ is equal to the full set of participants $\mathcal{I}$, which is the definition of a partition. Alternatively, we can define the partition of the participants by:

$$\mathcal{I} = \bigcup_{j=1}^{J} \mathcal{I}_j, \text{ where } x_i = x_{i'} \Leftrightarrow i, i' \in \mathcal{I}_j, \text{ for some } j.$$

Definition of a Restricted Permutation Group Under our assumptions, the set of admissible permutations g comprises those permutations that only permute indices of participants who share the same values on the pre-program variables. Notationally, permutations can only occur within each set $\mathcal{I}_j$, that is, among participants whose values of pre-program variables are equal to a_j. We call these "restricted permutations". A formal definition of the restricted permutation set $\mathcal{G}_X$ can be written as:

$$g \in \mathcal{G}_X \Leftrightarrow \pi_g : \mathcal{I} \to \mathcal{I} \text{ is such that } \forall\, i \in \mathcal{I}_j, \pi_g(i) \in \mathcal{I}_j \text{ for all } j \in \mathcal{J}.$$

This definition states that if a permutation g operates on the participant index i, which belongs to some partition set $\mathcal{I}_j$, then the permutation image $\pi_g(i)$ of that participant index also belongs to the same partition set $\mathcal{I}_j$. The definition allows for multiple swaps in different partition sets, but all swaps are restricted to occur only *within* each partition set. For example, suppose that $\mathcal{I}_1 = \{1, 2\}$ and $\mathcal{I}_2 = \{3, 4\}$. Then a permutation g for the set $\mathcal{I}_1$ and $\mathcal{I}_2$ that does not permute the elements in other sets can be defined by

$$\pi_g : \mathcal{I} \to \mathcal{I}; \qquad \pi_g \equiv \begin{cases} \pi_g(i) = i\, \forall\, i \in \mathcal{I} \setminus \left(\mathcal{I}_1 \cup \mathcal{I}_2\right) \\ \pi_g(1) = 2; \pi_g(2) = 1; \\ \pi_g(3) = 4; \pi_g(4) = 3. \end{cases}$$

Alternatively, the permutation g' defined by

$$\pi_{g'} : \mathcal{I} \to \mathcal{I}; \qquad \pi_{g'} \equiv \begin{cases} \pi_{g'}(i) = i\, \forall\, i \in \mathcal{I} \setminus \left(\mathcal{I}_1 \cup \mathcal{I}_2\right) \\ \pi_{g'}(1) = 1; \pi_{g'}(2) = 3; \\ \pi_{g'}(3) = 2; \pi_{g'}(4) = 4, \end{cases}$$

permutes index across partition sets and thus it does not satisfy the conditions required for inclusion in $\mathcal{G}$. Recall that we can also write the restricted permutation in terms of a linear transformation B_g such that $B_g D \equiv gD$, where B_g is the permutation matrix that imposes the restricted permutation g.

Sampling Procedure Among all possible restricted permutations $\mathcal{G}_X$ defined in the previous subsection, we select as valid permutations only the ones that result in equal label assignments for siblings. In other words, gD assigns the

same treatment labels to all members of the same family. A sampling procedure randomly selects J draws of permutations $g \in \mathscr{G}_X$ with replacement. Consequently, we have J permutation matrices B_g that correspond to each of the draws of the permutation g. We index these J permutations as g_j, where $j = 1, \ldots, J$. The sample data are described by the identity permutation, which we define as the $(J+1)^{\text{st}}$ permutation (notationally, g_{J+1}).

1. To respect the non-random assignment of siblings, we permute only the eldest siblings (who were those actually randomized). After each permutation we assign the younger siblings to the same group to which the elder siblings were assigned. In this step, we follow the randomization protocol exactly. Further steps of the randomization protocol are approximated, as described below.

2. The IQ pairing and pre-randomization swaps are directed at balancing IQ, gender, and SES index. We forbid permutations between genders as well as between the top and bottom half of the SES index. Sensitivity analysis reveals that inference is robust to this choice of percentiles.

3. The post-randomization swaps led to unbalanced working status of mothers. However, we are unable to restrict permutations based on mother's working status due to data limitations, although we use it as a linear covariate.

Simple Permutation Test Procedure Our permutation test is based on the following algorithm:

1. Sample a permutation $g \in \mathscr{G}_X$ with replacement.

2. Compute a test statistic for the permutation draw, based on data modified by the permutation matrix B_g and observed data.

3. Repeat steps 1 and 2 to simulate the permutation distribution of the test statistic.

After a "reasonable" number of draws, we compute a test statistic (e.g., Student's t for difference in means between the treatment and control groups) using simulated permutation distribution. An example of a permutation-based p-value is the fraction of the computed permutation distribution that is bigger than the statistic computed using the original unpermuted data. We use the mid-p-value described in Section 5.7. The next section describes the construction of our test statistic in greater detail.

5.6 Conditional Inference in Small Samples and the Test Statistic

The small sample of Perry experiment (123 participants) poses practical difficulties in partitioning participants into detailed categories based on the five pre-program variables. Restricted permutation orbits would have so few observations

as to preclude reliable inference. We obtain "reasonably-sized" restricted permutation orbits by imposing the additional assumption of a linear relationship between certain pre-program variables and outcomes. To this end, we divide the vector X into two parts: variables $X^{[L]}$, which are assumed to have a linear relation with Y, and the remaining variables $X^{[N]}$, whose relationship with Y is unconstrained. Using this partition, write $X = [X^{[L]}, X^{[N]}]$. Our linear assumption can be written as:

Assumption A-3 $Y_{i,d} \equiv \delta_d X_i^{[L]} + f(d, X_i^{[N]}, V_i);\ d \in \{0,1\},\ i \in \mathcal{I}.$

Under hypothesis **H-1**, $\delta_1 = \delta_0 = \delta$ and $\tilde{Y} \equiv Y - \delta X^{[L]} = f(X^{[N]}, V)$. Using assumption **A-3**, we can rework the arguments of Section 5.3 to prove that, under the null hypothesis of no treatment effect, the exchangeability of $\tilde{Y}$ holds among participants who share the same value of $X^{[N]}$ even if they have different values of $X^{[L]}$. Formally, we have that

$$(\tilde{Y}, D) \stackrel{d}{=} (\tilde{Y}, gD)\ ;\ g \in \mathcal{G}_{X^{[N]}},$$

where $\mathcal{G}_{X^{[N]}}$ is the set of permutations that swap the participants who share the same values of covariates $X^{[N]}$.

Under Assumption **A-3**, we do not have to partition the data for all possible combinations of $X^{[L]}$ and $X^{[N]}$ — we only partition based on values of $X^{[N]}$, the variables not assumed to have a linear relationship with the outcomes Y. If δ were known, permuting $\tilde{Y} = Y - \delta X^{[L]}$ (instead of Y) within the groups of participants that share the same pre-program variables $X^{[N]}$ would solve the problem of linear conditioning on $X^{[L]}$. However, δ is unknown. We address this problem by using an approach based on Freedman and Lane (1983), which entails permuting the residuals from the regression of Y on $X^{[L]}$ in orbits that share the same values of $X^{[N]}$, leaving D fixed. Specifically, Freedman and Lane (1983) use a conditional exchangeability principle and assume a fully linear model:

Under Assumptions **A-1**, **A-2**, and **A-3**, we can write our outcome Y as:

$$Y = f((X^{[L]}), D(X), \varepsilon(X^{[N]})) = \delta X^{[L]} + \Delta D + \varepsilon(X^{[N]}),$$

where $\Delta = 0$. As mentioned, if δ is known, we can use the residuals $\widetilde{Y} = Y - \delta X$ in a permutation test of the null $\Delta = 0$. However, δ is generally not known and must be estimated. The Freedman-Lane procedure assumes exchangeability of errors under the null, i.e., that the errors ε of the regression $Y = \delta X + \varepsilon$ are exchangeable under the null of no treatment effect: $(H_0 : \Delta = 0)$. We capture the concept of exchangeable errors in the Freedman-Lane procedure by permuting the residuals from the linear regression of Y on $X^{[L]}$ that excludes D. Permuting D and comparing test statistics for the different permutations assumes no statistical relationship between $X^{[L]}$ and D. Namely, it assumes no correlation between $X^{[L]}$ and D, which is unreasonable. We account for the nonlinear relationship

between Y and $X^{[N]}$ by using the permutation matrix B_g associated with restricted permutations $\mathscr{G}_{X^{[N]}}$, which only permutes participants who share the same values of pre-program variables $X^{[N]}$. Notationally, define the residuals from permutation g as $\tilde{\varepsilon}_g$ such that

$$\begin{aligned}\tilde{\varepsilon}_g &\equiv B_g Q_X Y \\ &= B_g(Y - \widehat{Y}),\end{aligned}$$

where $\widehat{Y}$ is the estimated Y and the matrix Q_X is defined as

$$Q_X \equiv (I - P_X),$$

where I is the identity matrix and

$$P_X \equiv X^{[L]}((X^{[L]})'X^{[L]})^{-1}(X^{[L]})'.$$

Matrices P_X and Q_X are well-known linear transformations. P_X is a linear projection in the space generated by the columns of $X^{[L]}$. Q_X is the projection into the orthogonal space generated by $X^{[L]}$. We can write the $\tilde{Y}_g = P_X Y + \tilde{\varepsilon}_g$ for a new outcome that preserves the linear relationship between X and Y, but permutes the errors. Use $\tilde{Y}_g$ as the permuted outcome data for permutation g and compute the new linear coefficient estimated for the dummy variable of treatment assignment D. This coefficient is the Freedman-Lane coefficient for permutation g and is given by:

$$\begin{aligned}\Delta^g &= (D'Q_X D)^{-1} D'Q_X \tilde{Y}_g \\ &= (D'Q_X D)^{-1} D'Q_X \left[X^{[L]}\left((X^{[L]})'X^{[L]}\right)^{-1} X^{[L]}Y + B_g Q_X Y\right] \\ &= (D'Q_X D)^{-1} D'Q_X (B_g Q_X Y).\end{aligned}$$

For notational purposes, we use Δ^j for the Freedman-Lane coefficient associated with outcome Y and permutation g_j (indexed by j), that is,

$$\Delta^j \equiv (D'Q_X D)^{-1} D'Q_X B'_{g_j} Q_X Y.$$

In a series of Monte Carlo studies, Anderson and Robinson (2001) compare the distributions of the test statistics under various approximate permutation methods with the distribution from a conceptually exact permutation method. All approximate methods produce permutation distributions under H_0 that converge to the same distribution. However, only the Freedman-Lane procedure has an expected correlation of 1 with the exact test, while the other methods are found to have smaller correlations. Thus, the Freedman-Lane procedure comes closest

to attaining the results of an exact test (where δ is known). In a series of Monte Carlo experiments, they show, for samples of the size used in this paper, that the Freedman-Lane size is very close to the exact size where δ is known. Another paper, by Anderson and Legendre (1999), conducts extensive Monte Carlo simulations and shows that the Freedman-Lane procedure generally gives the best results in terms of Type-I error and power. On the basis of these studies, we use the Freedman-Lane coefficient as the primary test statistic. Section 5.7 explains how to use statistic Δ^j to make inferences based on mid-p-values.

5.7 Formal Permutation Testing with Mid-p-Values

In this section, we formally define a mid-p-value under permutation testing and prove that it constitutes a valid level-α test.[4]

Let the set of all relevant permutations of the treatment assignment D be represented by $\{g_1, \ldots, g_J\}$. By relevant permutations we mean permutations that change the vector of treatment assignment, that is $gD \neq D$. Following the notation of Section 5.6, let the test statistics be Δ^j, where j accounts for a permutation index.

For each permutation g, we computed the pre-pivoted statistic T^j associated with Δ^j for the permuted treatment assignment $g_j D$. Specifically, we have:

$$T^j = \sum_{l=1}^{J+1} \mathbf{1}[\Delta^j \geqslant \Delta^l]/(J+1).$$

Δ^j may be substituted for T^j without affecting single-hypothesis-testing results, but Romano and Wolf (2005) recommend rank statistics to increase comparability for multiple hypothesis testing. See Beran (1988) for the statistical benefits of pre-pivoting statistic.

The mid-p-values may be defined as

$$p \equiv \frac{1}{2(J+1)} \left(\sum_{l=1}^{J+1} \mathbf{1}[T^l \geqslant T^{J+1}] + \sum_{l=1}^{J+1} \mathbf{1}[T^l > T^{J+1}] \right).$$

To accurately describe our testing procedure, we need a few more definitions. Fix a nominal level for the testing procedure at α and define

$$a = (J+1) - \lceil \alpha \cdot (J+1) \rceil,$$

where $\lceil \alpha \cdot (J+1) \rceil$ denotes the largest integer less than or equal to $\alpha \cdot (J+1)$. Let the ordered values of $T^j; j = 1, \ldots, J+1$ be represented by $T^{(1)}, \ldots, T^{(J+1)}$.

[4]Note that in this section, we use the fact that, under the randomization hypothesis, any real-valued statistic of the permuted data (i.e. p^j, T^j ; $j = 1, \ldots, J+1$) that provides $J+1$ distinct values as g varies in $\mathscr{G}$ is uniformly distributed across these $J+1$ values. For more details, see Lehmann and Romano (2005, Chapter 15).

Define α_0 as the percentage of test statistics T^j that are strictly greater than $T^{(a)}$:

$$\alpha_0 \equiv \frac{1}{(J+1)} \sum_{j=1}^{J+1} \mathbf{1}[T^j > T^{(a)}].$$

Define α_1 by the percentage of the test statistics T^j that are bigger than or equal to $T^{(a)}$:

$$\alpha_1 \equiv \frac{1}{(J+1)} \sum_{j=1}^{J+1} \mathbf{1}[T^j \geqslant T^{(a)}].$$

Observe that $\alpha \in [\alpha_0, \alpha_1]$. Let the interval $[0, 1]$ be partitioned into the three intervals $[0, \alpha_0), [\alpha_0, \alpha_1]$ and $(\alpha_1, 1]$. Our testing procedure assigns different *rejection probabilities* whenever p lies in each one of these intervals. Namely, we reject the null hypothesis if $p \in [0, \alpha_0)$, we do not reject it if $p \in (\alpha_1, 1]$, and we reject it with probability $\frac{\alpha - \alpha_0}{\alpha_1 - \alpha_0}$, if $p \in [\alpha_0, \alpha_1]$. This procedure can be summarized by the use of a test function ϕ. We reject the null hypothesis with probability τ, where τ is given by:

$$\tau \equiv \mathbf{1}[p < \alpha_0] \cdot 1 + \mathbf{1}[p > \alpha_1] \cdot 0 + \mathbf{1}[p \in [\alpha_0, \alpha_1]] \cdot \frac{\alpha - \alpha_0}{\alpha_1 - \alpha_0}.$$

The following theorem shows that this testing procedure yields a level α test.

Theorem 5.3 *Suppose that the randomization hypothesis holds. Let $J > 0$ and $0 < \alpha < 1$ be given. Then, the test that rejects $H_0 : Y \perp\!\!\!\perp D|X$ with probability τ defined above satisfies $\mathrm{P}\{\text{reject } H_0 \mid X\} = \alpha$ whenever H_0 is true.*

Proof

$$
\begin{aligned}
\mathrm{P}\{\text{reject } H_0 \mid X\} &= \mathrm{P}\{\tau = 1\} \\
&= \mathrm{E}[\tau] \\
&= \mathrm{E}\left[\mathbf{1}[p < \alpha_0] + \mathbf{1}[p \in [\alpha_0, \alpha_1]] \cdot \frac{\alpha - \alpha_0}{\alpha_1 - \alpha_0}\right] \\
&= \mathrm{E}\left[\mathbf{1}[p^{J+1} < \alpha_0] + \mathbf{1}[p^{J+1} \in [\alpha_0, \alpha_1]] \cdot \frac{\alpha - \alpha_0}{\alpha_1 - \alpha_0}\right] \\
&\quad \text{(because } p = p^{J+1}\text{)} \\
&= \left[\frac{1}{J+1}\sum_{j=1}^{J+1} \mathbf{1}[p^j < \alpha_0] + \mathbf{1}[p^j \in [\alpha_0, \alpha_1]] \cdot \frac{\alpha - \alpha_0}{\alpha_1 - \alpha_0}\right] \\
&\quad \text{(because } p^j \text{ is uniformly distributed across} \\
&\quad J+1 \text{ permutation values)} \\
&= \left[\frac{1}{J+1}\left(\sum_{j=1}^{J+1} \mathbf{1}[T^j > T^{(a)}]\right.\right. \\
&+ \left.\left.\sum_{j=1}^{J+1} \mathbf{1}[T^j = T^{(a)}] \cdot \frac{\alpha - \alpha_0}{\alpha_1 - \alpha_0}\right)\right] \\
&= \left[\frac{\sum_{j=1}^{J+1} \mathbf{1}[T^j > T^{(a)}]}{J+1}\right. \\
&+ \frac{\left(\sum_{j=1}^{J+1} \mathbf{1}[T^j \geqslant T^{(a)}] - \sum_{j=1}^{J+1} \mathbf{1}[T^j > T^{(a)}]\right)}{J+1} \\
&\cdot \left.\frac{\alpha - \alpha_0}{\alpha_1 - \alpha_0}\right] \\
&= \left[\alpha_0 + (\alpha_1 - \alpha_0) \cdot \frac{\alpha - \alpha_0}{\alpha_1 - \alpha_0}\right] \\
&= \alpha.
\end{aligned}
$$

The cardinality of the set $\mathscr{G}$ can be so large that computing p-values over all elements becomes infeasible. In this case, we use a test that uses random samples of J permutations $g \in \mathscr{G}$ plus the identity permutation as the $J+1$ draw.[5] By construction, a test that uses random sampling of elements in the permutation set has the same expectation as a test that uses all elements in the permutation set.

[5] Recall that draw $J+1$ is the sample data.

6. Multiple Hypothesis Testing with Stepdown

We correct for the possibility of arbitrary selection of statistically significant outcomes by performing a multiple hypothesis testing. A major question in multiple inference is the definition of a generalized Type-I error. Two criteria are often used: (1) the *familywise error rate* (FWER), which is the probability of rejecting any true null hypothesis, and (2) the *false discovery proportion* (FDP), which is the proportion of true null hypotheses rejected. The stepdown algorithm described below exhibits *strong FWER control*: FWER is held at or below a specified level regardless of the true configuration of the full set of hypotheses Lehmann and Romano, 2005.[6] We test a number of hypotheses simultaneously, mandating the choice of FWER as a criterion. FDP is more appropriate in the context of a very high number of hypotheses, such as dozens or hundreds of hypotheses, a common occurrence in fields such as genomics.

6.1 Overview of Multiple Hypothesis Testing

Two traditional, but conservative, methods for multiple hypothesis testing are the Bonferroni and Holm procedures (see (Lehmann and Romano, 2005), for a description of these tests). Their goal is to test K joint hypotheses. Each single hypothesis is represented by H_k where $k \in \mathcal{K} \equiv \{1, \ldots, K\}$, for which we have individual-hypothesis p-values $p_1, \ldots, p_K$. The joint hypothesis is given by $H_{\mathcal{K}}$ defined by:

$$H_{\mathcal{K}} = \bigcap_{k \in \mathcal{K}} H_k,$$

To control for FWER $< \alpha$, the traditional procedures use the following rejection rules:

Bonferroni:

Reject each H_k with $p_k \leq \alpha/K$.

Holm:

(1) Order the original p-values, with the notation $p_{(1)}, \ldots, p_{(K)}$;

(2) Find the highest k with $p_{(k)} \leq \alpha/(K - k + 1)$;

(3) Reject the hypotheses $H_{(1)}, \ldots, H_{(k)}$.

These two methods are computationally simple to implement, but they do not account for dependence between outcomes, while less conservative methods described below do.

[6] For further discussion of stepdown and its alternatives, see Westfall and Young (1993), Benjamini and Hochberg (1995), Romano and Shaikh (2004, 2006), Romano and Wolf (2005), Benjamini et al. (2006).

Modern work is based on the classical procedure of "closure methods."[7] General closure methods belong to a testing tradition called multiple comparison procedures (MCP). These constitute a more flexible and comprehensive framework for multiple hypothesis testing on the power set $\wp(H_{\mathcal{K}})$ of hypotheses $H_{\mathcal{K}}$. However, closure methods have two disadvantages: they are computationally impractical for large numbers of hypotheses, and computing the test statistics dictated by some joint hypotheses may be infeasible. Closure methods, such as those developed in Einot and Gabriel (1975) and Begun and Gabriel (1981), are based on a stepwise MCP. They start with the biggest set $\mathcal{K}$ of joint hypotheses and proceed through smaller sets of joint hypotheses.

Let $\mathcal{K}' \subseteq \mathcal{K}$. The test of the joint hypothesis $H_{\mathcal{K}'} = \bigcap_{k \in \mathcal{K}'} H_k$ at a significance level α uses a statistic $T_{\mathcal{K}'}$ with a critical value $c_{\mathcal{K}'}(\alpha_{\mathcal{K}'})$ at level $\alpha_{\mathcal{K}'}$. Higher values of $T_{\mathcal{K}'}$ provide evidence against hypothesis $H_{\mathcal{K}}$, and under $H_{\mathcal{K}'}$, $c_{\mathcal{K}'}(\alpha_{\mathcal{K}'})$ can be defined as:

$$\alpha_{\mathcal{K}'} \equiv \mathrm{P}(T_{\mathcal{K}'} > c_{\mathcal{K}'}(\alpha_{\mathcal{K}'})).$$

that is, $c_{\mathcal{K}'}(\alpha_{\mathcal{K}'})$ is the α-highest quantile of the distribution of the test statistic $T_{\mathcal{K}'}$.

For the Newman (1939) and Keuls (1952) procedure, $\alpha_{\mathcal{K}'} = \alpha$. For the Ryan (1959) procedure,

$$\alpha_{\mathcal{K}'} = 1 - (1 - \alpha)^{\frac{|\mathcal{K}'|}{|\mathcal{K}|}}.$$

The test of $H_{\mathcal{K}'}$ is called $\alpha_{\mathcal{K}'}$-critical if the computed test statistic $T_{\mathcal{K}'}$ for the sample is bigger than its critical value $c_{\mathcal{K}'}(\alpha_{\mathcal{K}'})$. An MCP rejects $H_{\mathcal{K}'}$ if all sets $\mathcal{K}'' \supseteq \mathcal{K}'$ are $\alpha_{\mathcal{K}''}$-critical, where $\mathcal{K}$ is the biggest set of joint hypotheses to be tested, in particular, $\mathcal{K}' \subseteq \mathcal{K}$. In other words, hypothesis $H_{\mathcal{K}'}$ is only rejected if all combinations of the joint hypothesis in $\mathcal{K}$ that include the hypothesis in $\mathcal{K}'$ are also rejected.

Observe that if a set of hypotheses $\mathcal{K}'$ is not $\alpha_{\mathcal{K}'}$-critical, that is, it is not rejected, then all combination sets of $\mathcal{K}'$ are also not rejected. This rule is called *acceptance by implication* Begun and Gabriel (1981) and it ensures logical coherence. If one joint hypothesis is not rejected, all subsets of the hypotheses will also fail to be rejected.

Traditional MCP algorithms start by targeting the larger set of joint hypotheses $H_{\mathcal{K}}$. If not rejected, all remaining combinations of hypotheses are not rejected either. If $H_{\mathcal{K}}$ is rejected, the procedure computes the critical value for all combinations of $K-1$ hypotheses in the set $\mathcal{K}$ without the most statistically significant hypothesis. A new round of rejections requires the computation of the critical values of all combinations of $K-2$ hypotheses in $\mathcal{K}$ without the two most statistically significant hypotheses, and so forth.

One computational problem arising from the method is the exponential increase of intersection hypotheses as $\mathcal{K}$ increases. In the worst case, this could require as

[7] See Lehmann and Romano (2005).

many as $2^K - 1$ tests. Another drawback is the computation of the critical values, which may be difficult for some of the intersection hypotheses. Closure methods strongly control for FWER, as shown in Marcus et al. (1976).

6.2 Subset Pivotality and the Free Stepdown Procedure

Data and Hypotheses Assume that the data Y have the true generating distribution $P \in \Omega$. The objective is to test the joint hypotheses $H_{\mathcal{K}} = \cap_{k \in \mathcal{K}} H_k$, where each H_k corresponds to a family of distributions $\omega_k \subseteq \Omega$, which may contain the true data generating distribution P:

$$H_k : P \in \omega_k.$$

Assume that the evidence against hypothesis H_k has been summarized using a p-value $p_k; k \in \mathcal{K}$. Let $p_{\mathcal{K}} = (p_k \,; k \in \mathcal{K})$ be the vector of random p-values generated from P. Let $\mathcal{K}(P)$ be the set of indices of the true hypothesis.

Subset Pivotality The distribution of $p_{\mathcal{K}}$ has the *subset pivotality* property if the joint distribution of any subvector $p_{\mathcal{L}} = (p_l \; ; \; l \in \mathcal{L})$; for an $\mathcal{L} \subset \mathcal{K}$ it would be identical if either $\mathcal{K}(P) = \mathcal{K}$ or $\mathcal{K}(P) = \mathcal{L}$. Westfall and Young (1993) clarify further by stating that the subset pivotality condition requires that the multivariate distribution of any subvector of p-values is unaffected by the truth or falsehood of hypotheses corresponding to the p-values that are not included in the subvector.

Westfall and Young (1993) argue that the subset pivotality condition is important for two reasons. First, resampling is particularly convenient under this condition: resampling is done under the assumption that all null hypotheses are true, rather than a subset of the hypotheses. Second, when subset pivotality holds, resampling-based methods provide strong control for FWER. At the time Westfall and Young (1993) paper was published, it was believed that subset pivotality was a necessary condition for FWER strong control. However, Romano and Wolf (2005) provide an algorithm that strongly controls for FWER under weaker conditions.

Cases of Failure Westfall and Young (1993) consider the problem of testing whether the correlations of a vector of N normally distributed random variables are all zero. Notationally, $H_{(i,j)} : \rho_{i,j} = 0$ and $\mathcal{K} = \{(i,j); i,j \in \{1,\ldots,N\}\}$. In large samples, a traditional test statistic is $T_{(i,j)} = \sqrt{n} \cdot r_{(i,j)}$, where n is the sample size and $r_{(i,j)}$ is the sample correlation between variables i and j. Suppose that hypotheses $H_{(1,2)}$ and $H_{(1,3)}$ are true, and that all others are false. Previous analyses by Aitkin (1969, 1971) show that the joint distribution of $[T_{(1,2)}, T_{(1,3)}]$ is approximately normal, with zero means, unit variances, and correlation $\rho_{2,3}$. The key observation is that the joint distribution of the test statistics for hypotheses $H_{(1,2)}$ and $H_{(1,3)}$ has different statistical properties depending on whether $\rho_{2,3} = 0$

or $\rho_{2,3} \neq 0$. Consider the hypothesis $H_{(2,3)} : \rho_{2,3} = 0$ as part of a set of hypotheses $\{H_{(1,2)}, H_{(1,3)}, H_{(2,3)}\}$. In this case, inference on the joint set of hypotheses $H_{(1,2)}$ and $H_{(1,3)}$ changes depending on whether hypothesis $H_{(2,3)}$ is true or not. The subset pivotality condition fails here because the distribution of $[T_{(1,2)}, T_{(1,3)}]$ depends upon the value of $\rho_{2,3}$, which is associated with another hypothesis not directly tested by $T_{(1,2)}$ or $T_{(1,3)}$. Observe that subset pivotality would hold if the hypotheses of interest involved only the means of the normal random variables.

6.2.1 The Free Stepdown Procedure

Westfall and Young (1993) use the assumption of subset pivotality to develop a stepdown procedure that exhibits strong controls over FWER. As mentioned above, p_k denotes the p-value associated with hypothesis k and the set of hypotheses is indexed by $\mathcal{K} = \{1, \ldots, K\}$. Without loss of generality, let the computed p-value statistic be sorted in increasing order; that is, $\hat{p}_1 \leq \hat{p}_2 \leq \cdots \leq \hat{p}_K$. Using some resampling method, let $(p_1^j, \ldots, p_K^j)$ be the j^{th} draw of the vector of p-values. These draws generate the joint testing distribution of $(p_1, \ldots, p_K)$ under $H_{\mathcal{K}}$. Let J be the total number of draws, that is, $j \in \{1, \ldots, J\}$.

Using this notation, the Westfall and Young (1993) algorithm is defined by:

1. For each draw j, compute the successive minima $q_k^j = \min\{p_k^j, \ldots, p_K^j\}$. This step enforces the original monotonicity of observed p-values. Note that k denotes the original rank of the outcome by significance, with $k = 1$ being the most significant and $k = K$ the least significant.

2. For each $k \in \mathcal{K}$, compute $\bar{p}_k = (\sum_{j=1}^{J} \mathbf{1}[q_k^j \leq \hat{p}_k])/J$. This step gives the percentage of times that the adjusted draws $(q_k^j \,;\, j = 1, \ldots, J)$ are equal to or smaller than p_k.

3. For each hypothesis $k \in \mathcal{K}$, enforce the successive maxima $\tilde{p}_k = \max\{\bar{p}_1, \ldots, \bar{p}_k\}$. This final enforcement of monotonicity ensures that larger unadjusted p-values correspond to larger adjusted ones.

The final $\tilde{p}_k$ are the adjusted p-values proposed by Westfall and Young (1993). Anderson (2008) claims to use this algorithm in performing multiple-hypothesis inference. However, the description of his algorithm does not comply with the one proposed in Westfall and Young (1993). Specifically, his algorithm is described as:

1. For each draw j, compute the successive minima $q_k^j = \min\{p_k^j, \ldots, p_K^j\}$. This step enforces the original monotonicity of experimentally observed p-values.

2. For each $k \in \mathcal{K}$, compute $\bar{p}_k = (\sum_{j=1}^{J} \mathbf{1}[q_k^j < \hat{p}_k])/J$. This step gives the percentage of times that the adjusted draws $(q_k^j \,;\, j = 1, \ldots, J)$ are strictly smaller than p_k.

3. For each hypothesis $k \in \mathcal{K}$, enforce the successive minima $\tilde{p}_k = \min \{\bar{p}_k, \ldots, \bar{p}_K\}$.

His procedure is different from the one proposed by Westfall and Young (1993) in the last step. Observe that while Westfall and Young (1993) use successive maxima on adjusted p-values, Anderson (2008) uses successive minima. Anderson (2008) does not provide any proof that the method he uses strongly controls for FWER.

6.3 Stepdown Multiple Hypothesis Testing

Stepdown methods improve upon general closure methods in two ways. First, they require only K separate tests. Second, the method tests joint hypotheses using only the test statistics for individual hypotheses, sidestepping the need to construct and compute specific test statistics for a large number of intersection hypotheses. Westfall and Young (1993) describe various methods of resampling outcomes Y for stepdown procedures, but those methods rely on the assumption of subset pivotality.

A recent result by Romano and Wolf (2005) shows that strong FWER control can be obtained by ensuring a certain monotonicity condition on the test statistics for the joint hypothesis that is weaker than subset pivotality. This monotonicity condition states that the critical value for a joint hypothesis that contains the subset of true hypotheses must be at least as large as the critical value for the joint hypothesis formed only by true hypotheses. Notationally, let $\mathcal{K}(P)$ be the set of indices of the true hypothesis, such that, $\mathcal{K}(P) \subseteq \mathcal{K}$. So that under probability law P, the monotonicity condition is defined by:

$$c_{\mathcal{K}}(\alpha) \geq c_{\mathcal{K}(P)}(\alpha).$$

In other words, the critical value for the full set of joint hypotheses indexed by $\mathcal{K}$, which contains the true hypothesis indices $\mathcal{K}(P)$, is greater than or equal to the critical value for the hypothesis that comprises only true hypothesis $H_{\mathcal{K}(P)}$.

In this framework, a set of sufficient conditions for strong FWER control can be stated as follows:

1. The joint-hypothesis test statistic at each stepdown stage is chosen to be the maximum of the individual-hypothesis test statistics.

2. If a permutation-based inference is adopted, then the same draw of permutation is used to compute all test statistics at each stage.

3. The permutation set from which permutations are drawn is chosen such that, under the null hypotheses, the distribution of the data is invariant for each permutation.

Below, we discuss how to construct tests that satisfy the first two conditions. The third condition applies to permutation testing of randomization hypotheses in general and requires constructing the permutation groups using knowledge of the experimental design that generated the data.

6.4 The Stepdown Algorithm

Data and Hypotheses Assume that we start with outcomes $Y^k \ ; \ k \in \mathcal{K} \equiv \{1, \ldots, K\}$, which have a true generating distribution $P \in \Omega$. The objective is to test a set of null hypotheses $H_{\mathcal{K}} = \bigcap_{k \in \mathcal{K}} H_k$ jointly, where each H_k corresponds to a family of distributions $\omega_k \subset \Omega$ which may contain the true data generating distribution P:

$$H_k : P \in \omega_k.$$

Permutation Testing In randomized experiments, the goal is to test the joint hypothesis of no treatment effect across outcomes $Y^k \ ; \ k \in \mathcal{K}$. The general representation of this hypothesis is given by $H_k \ : \ Y^k \perp\!\!\!\perp D$, where D is the treatment status. Thus H_k corresponds to a family of distributions ω_k in which the treatment status D is independent of outcome Y^k. Let $\mathscr{G}$ be a set of permutations, such that the randomization hypothesis holds, that is, the joint distribution of (Y^k, D), such that $k \in \mathcal{K}$, is invariant under permutations g in $\mathscr{G}$ whenever the true generating distribution P belongs to the family of distributions specified by $H_{\mathcal{K}}$. Formally,

$$P \in \bigcap_{k \in \mathcal{K}} \omega_k \ \Rightarrow \ \left[(Y^k, D) \stackrel{d}{=} (Y^k, gD) \ \forall \ g \in \mathscr{G}, \ \forall \ k \in \mathcal{K} \right].$$

Let $T_k \equiv T(Y^k, D)$ be the test statistic computed using the sample data, for which greater values provide evidence against the null hypothesis H_k. Let $T_k^g \equiv T(Y^k, gD)$ be the test statistic computed using the permuted data according to $g \in \mathscr{G}$. The distribution of T_k can be generated by varying g across $\mathscr{G}$.

Sets of Joint Hypotheses The stepdown method starts by testing the full set of joint null hypotheses $H_{\mathcal{K}}$. For notation purposes, define the set of hypotheses in this first step by $\mathcal{K}_1$, such that $\mathcal{K}_1 \equiv \mathcal{K}$. In each $K - 1$ successive step, the most individually significant hypothesis — the one most likely to contribute to the significance of the joint null hypothesis — is dropped from the set of null hypotheses, and the joint test is performed on the reduced set of hypotheses. Thus, the set of hypotheses for the second step is given by

$$\mathcal{K}_2 = \mathcal{K}_1 \setminus \{k^*\} \ ; \ k^* = \ \text{argmax}(T_k; k \in \mathcal{K}_1).$$

Likewise, the set of hypotheses for the step s is given by:

$$\mathcal{K}_s = \mathcal{K}_{s-1} \setminus \{k^*\} \ ; \ k^* = \ \text{argmax}(T_k; k \in \mathcal{K}_{s-1})$$

Finally, the final step targets the least significant hypothesis: $\mathcal{K}_K = \{\text{argmin}(T_k; k \in \mathcal{K})\}$.

Joint Test Statistics and Critical Values The test statistic for any step s that tests the joint hypothesis $H_{\mathcal{K}_s}$, with $\mathcal{K}_s$ as defined above, is given by:

$$T_{\mathcal{K}_s} = \max\left(T_k; k \in \mathcal{K}_s\right).$$

Let $T^g_{\mathcal{K}_s} \equiv \max\left(T^g_k; k \in \mathcal{K}_s\right)$, which is the maximum of the test statistics T^g_k such that $k \in \mathcal{K}_s$ and $g \in \mathscr{G}$. The distribution of $T_{\mathcal{K}_s}$ can be generated by varying g across $\mathscr{G}$. The critical value for each hypothesis $H_{\mathcal{K}_s}\ ;\ s \in \{1, \ldots, K\}$ at level α is defined as the value of the α-highest quantile of the distribution of $T_{\mathcal{K}_s}$. Namely, if we relabel the statistics $T^g_{\mathcal{K}_s}\ ;\ g \in \mathscr{G}$ by arranging them in increasing order:

$$T^{(1)}_{\mathcal{K}_s} \leq \cdots \leq T^{(|\mathscr{G}|)}_{\mathcal{K}_s}.$$

then the critical value for $T_{\mathcal{K}_s}$ is given by:

$$c_{\mathcal{K}_s}(\alpha) = T^{(a)}_{\mathcal{K}_s},$$

where $a = \lceil (1-\alpha)|\mathscr{G}| \rceil$, that is, the biggest integer smaller than or equal to $(1-\alpha)|\mathscr{G}|$. According to Romano and Wolf (2005), the use of the maximum operator in the definition of the joint statistic ensures the required monotonicity property of the critical values.

We assume full enumeration of the permutation set $\mathscr{G}$ for generating the distribution of the test statistics and for computing the critical values described in this section. However, for implementing the method, it is common to randomly sample permutations $g \in \mathscr{G}$ and use the sampled permutations for computing the statistics. Romano and Wolf (2005, p. 99, Corollary 3) show that FWER control of the stepdown procedure persists when using randomly sampled permutations in $\mathscr{G}$ instead of their full enumeration.

The Stepdown Algorithm The stepdown algorithm described in Romano and Wolf (2005) is defined by: Beginning with $\mathcal{K}_1 = \mathcal{K}$,

$[s = 1]$	If $T_{\mathcal{K}_1} \leq c_{\mathcal{K}_1}(\alpha)$, accept all $H_k; k \in \mathcal{K}_1$ and stop. Otherwise, let $\mathcal{K}_2 = \mathcal{K}_1 \setminus \{k^*\}; k^* = \text{ argmax}(T_k; k \in \mathcal{K}_1)$.
$\vdots$	
$[1 < s < K]$	If $T_{\mathcal{K}_s} \leq c_{\mathcal{K}_s}(\alpha)$, accept all $H_k; k \in \mathcal{K}_s$ and stop. Otherwise, $\mathcal{K}_{s+1} = \mathcal{K}_s \setminus \{k^*\}; k^* = \text{ argmax}(T_k; k \in \mathcal{K}_s)$.
$\vdots$	
$[s = K]$	If $T_{\mathcal{K}_K} \leq c_{\mathcal{K}_K}(\alpha)$, accept $H_{\mathcal{K}_K}\ ; \mathcal{K}_K = \{\text{argmin}(T_k; k \in \mathcal{K}).\}$ Otherwise, reject all $H_k\ ;\ k \in \mathcal{K}$.

(Romano and Wolf, 2005, p. 99, Corollary 2) demonstrate strong FWER control on a test of multiple-hypotheses $H_{\mathcal{K}}$ at level α if one performs this stepdown algorithm using the joint test statistics and the critical values defined above.

In contrast with the classical tests, the stepdown procedure strongly controls for FWER, while classical tests do not. Moreover, the procedure can generate as many adjusted p-values as there are hypotheses. Thus, it provides a way to determine which hypotheses are rejected.

There is some arbitrariness in defining the blocks of hypotheses that are jointly tested in a multiple-hypothesis testing procedure. The Perry study collects information on a variety of diverse outcomes. A single null hypothesis is associated with each outcome to avoid arbitrariness in selecting blocks of hypotheses associated with a single health topic. Each block is of independent interest and would be selected 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.

7. Empirical Results

We now apply our tools to analyze the Perry data. We find large gender differences in treatment effects for addictive behavior outcomes at age 40. We find statistically significant treatment effects for males on the use of hard drugs. For females, we find significant effects on the use of hallucinogenic substances and the impact of alcohol on many aspects of their lives. These effects persist after controlling for compromised randomization and multiple-hypothesis testing.

Tables 1–2 summarize the estimated effects of the Perry program on addictive behavior outcomes grouped by meaning at the age of 40. Table 1 reports results for females while Table 2 shows the results for males.

The tables present a set of results that examine the impact of Perry early childhood on addictive behavior outcomes. Each table comprises eight columns. The first column provides the name of the variable. We switch the sign of the outcome in order that increasing values of outcomes are beneficial for participants.

The variables are grouped into two blocks separated by empty lines. We divide outcomes into blocks for multiple-hypothesis testing by type of outcome and similarities as to the type of measure. The categories of variables were selected on *a priori* grounds. The first block examines the impact of drug and alcohol use on a range of life aspects. The last block examines the impact of Perry intervention on hard drug use. Due to the lack of heroin addiction among females and low variance in marijuana usage, we adopt a different selection of variables for females. In the last block of female outcomes, we replace heroin and marijuana use with the overall impact of addictive substance on life and the particular impact of addictive substance upon their jobs. The last column of each table performs a multiple hypothesis testing on each block of variables.

The second column of Tables 1–2 provides the control mean. Positive values of

the difference in means provide evidence that the treatment works in the desired direction. The third column presents the difference in means between treatment and control groups. The fourth one shows the t-statistic for the difference in means. The fifth column shows the asymptotic p-value for the t-statistic associated with the difference in means using the t-distribution. This p-value does not control for small sample size of Perry data and neither controls for the compromising aspects of the randomization protocol.

The sixth column shows a small-sample one-sided p-value computed using a naive permutation method, where the the p-value is computed based on an unrestricted permutation of the vector of treatment status. We use the mid-p-value of the pre-pivoted statistic as described in Section 5.7. The p-value shown in the sixth column is a statistic that tests the hypothesis of no treatment effect based on conditional permutation inference, without orbit restrictions or linear covariates. This p-value can be understood as a naive attempt to account for the small sample size of Perry intervention. These naive p-values are very close to their asymptotic equivalents.

The seventh column shows the one-sided p-values for the hypothesis of no treatment effect based on the pre-pivoted Freedman-Lane statistic as described in Section 5.6. We assume a linear relation for the covariates maternal employment, paternal presence, and Stanford-Binet IQ, and we use restricted permutation orbits within strata formed by the socioeconomic status index (SES) being above or below the sample median and permuting siblings as a block. This p-value accounts for both, the small sample size of Perry intervention and for the compromising aspects of its randomization.

The last column shows the p-values from the previous column adjusted for multiple inference using the stepdown algorithm of Romano and Wolf (2005). Our type-I error for the multiple hypothesis testing is the *familywise error rate* (FWER), which is the probability of rejecting any true null hypothesis. Romano and Wolf (2005) show that strong FWER control is obtained with the stepdown algorithm, that is, FWER is held at or below a specified level regardless of the true configuration of the full set of hypotheses.

We use the mid-p-value statistics for all p-values based on permutation tests. All p-values are computed using 5,000 draws under the relevant permutation procedure. All inference is based on one-sided p-values under the assumption that treatment is not harmful. We use pre-pivoted statistics.

Inference for female data is presented in Table 1. In summary, females show strong effects for negative impacts of drug/alcohol use on life activities and of hard drug usage. The first block of variables shows that the intervention has strong effect on the reduction of the negative impact of drug and alcohol use on many aspects of female treated participants. In particular, the aspects surveyed by Perry intervention are: child care, living standards at home, and physical and emotional health. The results survive stepdown adjustment. The last block of

variables shows that the treated females also benefit from a decrease in both the usage of hard drugs and in the overall effect of drug use on their lives and jobs. Treated females use less cocaine and have fewer problems with drug use in their jobs. These results also survive stepdown adjustment.

Inference for male data is presented in Table 2. In summary, Perry treatment had less pronounced impacts on males than on females. Still, treated males use fewer hard drugs than control ones. The first block of variables shows that the intervention did not consistently decrease the negative impact of drug and alcohol usage on the lifetime activities of treated males. The results are not statistically significant for any inference method. The last block of variables shows that treated males benefit from a decrease in the usage of hard drugs. Treated females use less heroin and marijuana or hashish. These results survive stepdown adjustment.

8. Summary and Conclusions

Barros's contribution to the literature of policy evaluation stems from relevance and sophisticated methodology. His research adds to this literature by advancing in statistical methods applied to rigorous empirical analysis. On the other hand, a central question of Barros's research is the analysis of efficient policies for reducing economic inequality. By pursuing this question, Barros's line of work has recently shifted towards the study of early childhood investment (Barros and Olinto, 2008). Barros examines whether early investment can be used as a powerful tool for promoting economic growth and increasing the likelihood of economic success of children born in disadvantaged families (Barros et al., 2011). We follow Barros's steps by providing a formal statistical evaluation of an important early childhood intervention. Specifically, we target the Perry preschool program, the oldest and most cited early childhood experiment in the U.S. We follow Barros's line of work by providing a rigorous statistical analysis that is both formal and relevant to the question of reducing economic inequality.

The Perry preschool program is an early childhood intervention that provided preschool education to low-IQ, disadvantaged African-American children living in Ypsilanti, Michigan. The Perry program is a major cornerstone for the ones that advocate the efficacy of investing in early childhood. In this paper, we focus on the long-term impact of this early childhood intervention on health variables and on addictive behavior. We use new data through age 40, which had never been analyzed before.

Few social experiments perfectly implement planned treatment assignment protocols. A proper analysis of such experiments requires recognizing the sampling plan as implemented and its sampling characteristics. Accounting for the implementation of social experiments produces more reliable results and unbiased inferences. Perry is no exception. The evaluation of the Perry program faces three statistical problems that are pervasive in social experiments:

Table 1
Main addictive behavior outcomes, females

	Females						
Variable	Contr. Mean	Diff Mean	t-stat	Ass Pval	Per. Pval.	Single	Stepdown
Drug/Alcohol Had Negative Effects on Your Marriage or Love Life	−1,136	0,136	1,904	0,032	0,021	0,162	0,162
Drug/Alcohol Had Negative Effects on Caring for Your Children	−1,273	0,273	2,934	0,003	0,001	0,003	0,010
Drug/Alcohol Had Negative Effects on Caring for Your Home	−1,227	0,227	2,598	0,006	0,003	0,009	0,026
Drug/Alcohol Had Negative Effects on Your Physical Health	−1,273	0,273	2,934	0,003	0,002	0,003	0,012
Drug/Alcohol Had Negative Effects on Your Mental or Emotional Health	−1,227	0,227	2,598	0,006	0,003	0,027	0,051
Cocaine or Crack or Free Base	−1,182	0,140	1,531	0,066	0,065	0,038	0,076
LSD or Other Hallucinogens	−1,046	0,046	1,046	0,151	0,092	0,244	0,244
Your Use of Drugs or Alcohol Had any Negative Effects	−1,409	0,326	2,738	0,004	0,007	0,016	0,046
Drug/Alcohol Had Negative Effects on your Job	−1,273	0,273	2,934	0,003	0,003	0,008	0,030

The above table presents a set of results that examine the impact of Perry early childhood on addictive behavior outcomes. The table contains eight columns. The first column provides the name of the variable. We switch the sign of the outcome so that increasing values of outcomes are beneficial for participants. The second column provides the control mean. Positive values of the difference in means provide evidence that the treatment works in the desired direction. The third column presents the difference in means between treatment and control groups. The fourth one shows the t-statistic for the difference in means. The fifth column shows the asymptotic p-value for the t-statistic associated with the difference in means using the t-distribution. The sixth column shows a small-sample one-sided p-value computed using a naive permutation method, where the the p-value is computed based on an unrestricted permutation of the vector of treatment status. We use the mid-p-value of the pre-pivoted statistic as described in subection 5.7. The p-value shown in the sixth column is a statistic that tests the hypothesis of no treatment effect based on conditional permutation inference, without orbit restrictions or linear covariates. The seventh column shows the one-sided p-values for the hypothesis of no treatment effect based on the pre-pivoted Freedman-Lane statistic as described in Section 5.6. We assume a linear relation for maternal employment, paternal presence, and Stanford-Binet IQ, and we use restricted permutation orbits within strata formed by the socioeconomic status index (SES) being above or below the sample median and permuting siblings as a block. The last column shows the p-values from the previous column adjusted for multiple inference using the stepdown algorithm of Romano and Wolf (2005). Our type-I error for the multiple hypothesis testing is the *familywise error rate* (FWER), which is the probability of rejecting any true null hypothesis. Romano and Wolf (2005) show that strong FWER control is obtained with the stepdown algorithm, that is, FWER is held at or below a specified level regardless of the true configuration of the full set of hypotheses.

Table 2
Main addictive behavior outcomes, males

	Males						
Variable	Contr. Mean	Diff Mean	t-stat	Ass Pval	Per. Pval.	Single	Stepdown
Drug/Alcohol Had Negative Effects on Your Marriage or Love Life	−1,229	−0,038	−0,350	0,636	0,625	0,451	0,583
Drug/Alcohol Had Negative Effects on Caring for Your Children	−1,171	0,005	0,050	0,480	0,480	0,360	0,577
Drug/Alcohol Had Negative Effects on Caring for Your Home	−1,229	−0,005	−0,045	0,518	0,514	0,347	0,576
Drug/Alcohol Had Negative Effects on Your Physical Health	−1,229	−0,105	−0,933	0,823	0,825	0,696	0,696
Drug/Alcohol Had Negative Effects on Your Mental or Emotional Health	−1,257	−0,043	−0,379	0,647	0,644	0,393	0,570
Marijuana or Hashish	−1,714	0,232	1,914	0,030	0,031	0,002	0,007
Cocaine or Crack or Free Base	−1,353	−0,026	−0,213	0,584	0,586	0,291	0,291
LSD or Other Hallucinogens	−1,086	0,017	0,245	0,404	0,439	0,192	0,317
Heroin	−1,143	0,143	2,164	0,017	0,022	0,014	0,039

The above table presents a set of results that examine the impact of Perry early childhood on addictive behavior outcomes. The table contains eight columns. The first column provides the name of the variable. We switch the sign of the outcome so that increasing values of outcomes are beneficial for participants. The second column provides the control mean. Positive values of the difference in means provide evidence that the treatment works in the desired direction. The third column presents the difference in means between treatment and control groups. The fourth one shows the t-statistic for the difference in means. The fifth column shows the asymptotic p-value for the t-statistic associated with the difference in means using the t-distribution. The sixth column shows a small-sample one-sided p-value computed using a naive permutation method, where the the p-value is computed based on an unrestricted permutation of the vector of treatment status. We use the mid-p-value of the pre-pivoted statistic as described in subection 5.7. The p-value shown in the sixth column is a statistic that tests the hypothesis of no treatment effect based on conditional permutation inference, without orbit restrictions or linear covariates. The seventh column shows the one-sided p-values for the hypothesis of no treatment effect based on the pre-pivoted Freedman-Lane statistic as described in Section 5.6. We assume a linear relation for maternal employment, paternal presence, and Stanford-Binet IQ, and we use restricted permutation orbits within strata formed by the socioeconomic status index (SES) being above or below the sample median and permuting siblings as a block. The last column shows the p-values from the previous column adjusted for multiple inference using the stepdown algorithm of Romano and Wolf (2005). Our type-I error for the multiple hypothesis testing is the *familywise error rate* (FWER), which is the probability of rejecting any true null hypothesis. Romano and Wolf (2005) show that strong FWER control is obtained with the stepdown algorithm, that is, FWER is held at or below a specified level regardless of the true configuration of the full set of hypotheses.

(1) the randomization was compromised;

(2) the sample size is small; and

(3) an overwhelming number of outcomes creates the danger of selectively reporting "significant" effects from a large pool of possible effects, biasing downward the reported p-values.

This paper develops tools to solve these three statistical problems.

We solve the problem of compromised randomization and small sample size by developing a small-sample permutation test tailored to the features of the less-than-ideal randomization of Perry intervention. We account for compromises in the randomization protocol by conditioning on background variables to control for the violations of the initial randomization protocol and imbalanced background variables. We address the potential problem of the arbitrary selection of statistically significant outcomes by using a multiple-hypothesis testing based on the stepdown procedure (Romano and Wolf, 2005). We use small sample permutation methods to compute familywise error rates that account for the multiplicity of experimental outcomes. 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 multiple hypotheses.

In summary, we find that treated females have fewer negative effects of drug/alcohol usage on a range of later-life activities and use fewer hard drugs, such as cocaine. Our results show that the treatment had less pronounced impacts on males than on females. Still, treated males use fewer hard drugs, such as heroin and hashish, than control ones.

References

Aitkin, M. A. (1969). Some tests for correlation matrices. *Biometrika*, 56(2):443–446.

Aitkin, M. A. (1971). Correction: Some tests for correlation matrices. *Biometrika*, 58(1):245.

Anderson, M. (2008). 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. & Legendre, P. (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. & Robinson, J. (2001). Permutation tests for linear models. *The Australian and New Zealand Journal of Statistics*, 43(1):75–88.

Barros, R., Carvalho, M., Franco, S., Mendon¸ca, R., & Rosalém, A. (2011). Uma avaliação do impacto da qualidade da creche no desenvolvimento infantil. *Pesquisa e Planejamento Econômico*, 41(2).

Barros, R. & Olinto, O. (2008). Early child development program of Rio de Janeiro – Impact evaluation design and challenges. *IPEA, Rio de Janeiro Brazil and, DECVP*.

Begun, J. & Gabriel, K. R. (1981). Closure of the Newman-Keuls multiple comparison procedure. *Journal of the American Statistical Association*, 76(1):241–45.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Association, Series B*, 57(1):289–300.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507.

Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83(403):687–697.

Einot, I. & Gabriel, K. R. (1975). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association*, 70(351):574–583.

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

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

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. Q. (2010b). Reanalysis of the perry preschool program: Multiple-hypothesis and permutation tests applied to a quasi-randomized experiment. *Quantitative Economics*, 1(1).

Keuls, M. (1952). The use of the "studentized range" in connection with an analysis of variance. *Euphytica*, 1(2):112–122.

Lehmann, E. L. & Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer Science and Business Media, New York, third edition.

Marcus, R., Peritz, E., & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660.

Newman, D. (1939). The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. *Biometrika*, 31(1/2):20–30.

Romano, J. P. & Shaikh, A. M. (2004). On control of the false discovery proportion. Technical Report 2004-31, Department of Statistics, Stanford University.

Romano, J. P. & Shaikh, A. M. (2006). Stepup procedures for control of generalizations of the familywise error rate. *Annals of Statistics*, 34(4):1850–1873.

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

Ryan, T. A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin*, 56(1):26–47.

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

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

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