Inference with Imperfect Randomization: The Case of the Perry Preschool Program*

James Heckman
Department of Economics
University of Chicago
jjh@uchicago.edu

Rodrigo Pinto
Department of Economics
University of California at Los Angeles
rodrig@econ.ucla.edu

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

December 4, 2023

Abstract

This paper considers the problem of making inferences about the effects of a program on multiple outcomes when the assignment of treatment status is imperfectly randomized. By imperfect randomization we mean that treatment status is reassigned after an initial randomization on the basis of characteristics that may be observed or unobserved by the analyst. We develop a partial identification approach to this problem that makes use of information limiting the extent to which randomization is imperfect to show that it is still possible to make nontrivial inferences about the effects of the program in such settings. We consider a family of null hypotheses in which each null hypothesis specifies that the program has no effect on one of several outcomes of interest. Under weak assumptions, we construct a procedure for testing this family of null hypotheses in a way that controls the familywise error rate – the probability of even one false rejection – in finite samples. We develop our methodology in the context of a reanalysis of the HighScope Perry Preschool program. We find statistically significant effects of the program on a number of different outcomes of interest, including outcomes related to criminal activity for males and females, even after accounting for the imperfectness of the randomization and the multiplicity of null hypotheses.

KEYWORDS: Exact Inference, Experiments, Familywise Error Rate, Imperfect Randomization, Multiple Testing, Multiple Outcomes, Permutation Tests, Perry Preschool Program, Program Evaluation

JEL CODES: C31, I21, J13

*This research was supported by the American Bar Foundation, the Committee for Economic Development, Pew Charitable Trusts and the Partnership for America's Economic Success, the JB and MK Pritzker Family Foundation, the Susan Thompson Buffett Foundation, Robert Dugger, the National Institute for Child Health and Human Development (Grants R01-HD043411 and R01-HD065072), the Institute for New Economic Thinking (Grant 262), and the National Science Foundation (Grants DMS-0820310 and SES-1530661). The views expressed in this paper are those of the authors and not necessarily those of the funders listed here. We thank Patrick Kline, Aprajit Mahajan, Joseph Romano, Andres Santos, Edward Vytlacil and Daniel Wilhelm for helpful comments. This paper was first circulated as National Bureau of Economic Research Working Paper w16935 in April 2011.

1 Introduction

This paper considers the problem of making inferences about the effects of a program on multiple outcomes when assignment of treatment status is imperfectly randomized. By imperfect randomization we mean that treatment status is reassigned after an initial randomization on the basis of characteristics that may be observed or unobserved by the analyst. As noted by Heckman et al. (2010a), such post-randomization reassignment of treatment status often occurs in real-world field experiments. Since these characteristics may affect outcomes, differences in outcomes between the treated and untreated groups may be due to the imperfectness of the randomization instead of the treatment itself.

We develop a partial identification approach to this problem that makes use of information limiting the extent to which randomization is imperfect to show that it is still possible to make nontrivial inferences about the effects of the program in such settings. We consider a family of null hypotheses in which each null hypothesis specifies that the program has no effect on one of several outcomes of interest. Under weak assumptions, we construct a procedure for testing this family of null hypotheses in a way that controls the familywise error rate – the probability of even one false rejection – in finite samples.

Our methodology depends on a detailed understanding of the way in which treatment status was assigned. For this reason, we develop it in the context of a specific application – a reanalysis of the HighScope Perry Preschool program – and our assumptions are tightly connected to the specific way in which treatment status was assigned in this program. We emphasize, however, that the underlying approach applies not only to this program, but more generally to the analysis of other experiments with imperfect randomization.

The HighScope Perry Preschool program is an influential preschool intervention that targeted disadvantaged African-American youth in Ypsilanti, Michigan in the early 1960s. The reported beneficial long-term effects of the program are a cornerstone in the argument for early childhood intervention in the United States. Most analyses of the HighScope Perry Preschool program have failed to account for the limited sample size of the study, the multiplicity of null hypotheses being tested, as well as the way in which treatment status in the program was imperfectly randomized. For an exposition of some of these criticisms, see, e.g., Herrnstein and Murray (1994), and Hanushek and Lindseth (2009). Two notable exceptions are Heckman et al. (2010a) and, more recently, Heckman and Karapakula (2019), who both acknowledge these concerns and address them in different ways than we do here. We postpone a detailed comparison of our approach with theirs to Remarks 2.1 and 4.5 below, but emphasize that both approaches do not permit in the way that we do here for post-randomization reassignment of treatment status on the basis of characteristics that are unobserved by the analyst. In particular, as explained further in Section 2 below, a key part of the intervention required families to be available for weekly home visits, and some families for whom this was not possible were removed from the treatment group and placed in the control group. In our analysis, we treat the availability of families for these weekly home visits as an unobserved characteristic that, importantly, may be related to potential outcomes. With our approach, we still find, like these other two studies, statistically significant effects of the program on a wide variety of outcomes, including outcomes related to criminal activity for males and females, and thereby contribute to the cumulative evidence of the favorable effects of this intervention.

The remainder of the paper is organized in the following way. Section 2 describes the HighScope Perry

Preschool program, focusing on the way in which treatment status was reassigned after the initial randomization on the basis of characteristics both observed and unobserved by the analyst. Section 3 formally describes our setup and assumptions, which are motivated by the description in the preceding section of the way in which treatment status was assigned in the program. We present our testing procedures in Section 4. We first discuss the problem of testing a single (joint) null hypothesis, before considering the problem of testing multiple null hypotheses. Section 5 presents the results of applying our methodology to the data from the HighScope Perry Preschool program. Section 6 concludes.

2 Empirical Setting

2.1 HighScope Perry Preschool Program

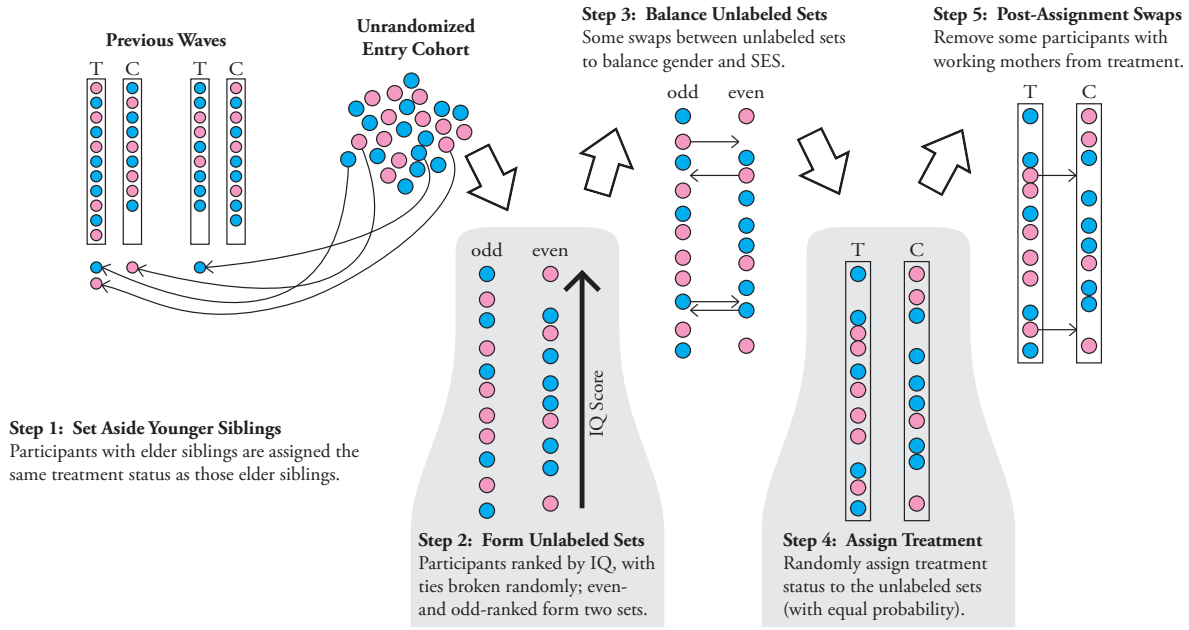
The HighScope Perry Preschool program was a prominent early childhood intervention conducted at the Perry elementary school in Ypsilanti, Michigan during the early 1960s. Beginning at age three and lasting for two years, treatment consisted of a 2.5-hour preschool program on weekdays during the school year supplemented by weekly home visits from teachers. The preschool curriculum was organized around the principle of active learning, guiding students through key learning experiences with open-ended questions. Social and emotional skills were also fostered. See Heckman et al. (2013). The purpose of the weekly home visits was to involve the parents in the learning process. Further details about the program are described in Schweinhart et al. (1993).

Program eligibility was determined by the child's Stanford-Binet IQ score and a measure of the family's socio-economic status. The measure of socio-economic status used was constructed as a weighted linear combination of father's skill level and educational attainment and the number of rooms per person in the family's home. With a few exceptions, those with Stanford-Binet IQ scores less than 70 or greater than 85 were excluded from the program. Likewise, with a few exceptions, those with a sufficiently high socio-economic status were excluded from the program.

The study enrolled a total of five cohorts over the years 1962-1965; two cohorts were admitted in the first year and one in each subsequent year. The first cohort is exceptional in that treated children only received one year of treatment beginning at age four. Altogether 123 children from 104 families were admitted to the program. Siblings are distributed among families as follows: 82 singletons, 17 pairs, 1 triple and 1 quadruple.

Follow-up interviews were conducted yearly from 3 to 15 years old. Additional interviews were conducted in three waves that cover persons in age intervals centered at ages 19, 27, and 40 years. Program attrition remained low through age 40. Indeed, over 91% of the participants were accounted for in the final survey. Moreover, two-thirds of those who did not were dead. Interviews covered a variety of outcomes. See Schweinhart et al. (1993) and Heckman et al. (2010a) for further discussion. For the purposes of our analysis, we focus on outcomes that have attracted considerable attention in the literature on the HighScope Perry Preschool program: IQ, achievement test scores, educational attainment, criminal behavior, and employment at three different stages of the life cycle.

Figure 1: Graphical Description of the Randomization Procedure



Notes: T and C refer to treatment and control groups respectively. Blue circles represent males. Pink circles represent females.

2.2 Randomization Procedure

Our methodology relies on a detailed understanding of the randomization procedure. According to Schweinhart et al. (1993), treatment status was assigned for each cohort of children in the following way:

> **Step 1**: Younger siblings of earlier program participants were assigned the same treatment status as their elder siblings.
>
> **Step 2**: Remaining participants were ranked according to their Stanford-Binet IQ scores at study entry. Those with the same Stanford-Binet IQ scores were ordered at random with all orderings equally likely. Two groups were defined by the odd-ranked and even-ranked participants.
>
> **Step 3**: Some participants were exchanged between the two groups in order to "balance" gender and the socio-economic status scores while keeping Stanford-Binet IQ scores roughly constant.
>
> **Step 4**: The two groups defined in this way were labeled treatment and control with equal probability.
>
> **Step 5**: Some participants with single mothers who were working and unavailable for the weekly home visits were moved from the treatment group to the control group.

This procedure is depicted graphically in Figure 1. The rationale for assigning younger siblings of earlier program participants to the same treatment status as their elder siblings was to avoid "spillovers" within a family, that might weaken the estimated treatment effect. For our purposes, it is most important to note that Step 5 depends on a characteristic we do not observe – whether the family has a single mother who

is working and unavailable for the weekly home visits – but was observed and used by those determining treatment status (at least for families who were offered treatment). To the extent that the availability of the mother is related to the outcomes of interest, it is important to account for this feature of the randomization procedure in analyzing experimental outcomes.

Note that by symmetry we may without loss of generality interchange Steps 3 and 4 of the randomization procedure without affecting the distribution of treatment status. Thus, the randomization procedure may be described equivalently as follows:

> **Step 1′**: Younger siblings of earlier program participants were assigned the same treatment status as their elder siblings.
>
> **Step 2′**: Remaining participants were ranked according to their Stanford-Binet IQ scores at study entry. Those with the same Stanford-Binet IQ score were ordered at random with all orderings equally likely. Two groups were defined by the odd-ranked and even-ranked participants.
>
> **Step 3′**: The two groups defined in this way were labeled treatment and control with equal probability.
>
> **Step 4′**: Some participants were exchanged between the treatment and control groups in order to "balance" gender and socio-economic status score while keeping Stanford-Binet IQ score roughly constant.
>
> **Step 5′**: Some participants with single mothers who were working and unavailable for the weekly home visits were moved from the treatment group to the control group.

This observation will be useful below when modeling the distribution of treatment status.

Remark 2.1. Heckman and Karapakula (2019) interpret "balance" in Step 3 to be defined in terms of Hotelling's multivariate two-sample t-squared statistic being less than some threshold. They also discipline Step 5 by assuming that there were (at most) a certain number of participants with single, working mothers for whom special accommodations could be made for the weekly home visits, and that the analyst chose which families to accommodate at random. Both the threshold and the number of participants for whom special accommodations could be made are treated as unknown, but can be partially identified from the observed data. For testing procedures that exploit this different model of the way in which treatment status was determined, we refer the reader to Heckman and Karapakula (2019). We emphasize, however, that our approach below allows, in particular, that single, working mothers were unavailable for these weekly home visits for reasons that may be important in that they are related to potential outcomes. ■

3 Setup and Assumptions

3.1 Setup

We index outcomes of interest by $k \in K$, families by $j \in J$ and siblings in the jth family by $i \in I_j$. Denote by $Y_{i,j,k}(0)$ the kth (potential) outcome of the ith sibling in the jth family if the jth family were not treated

and by $Y_{i,j,k}(1)$ the kth (potential) outcome of the ith sibling in the jth family if the jth family were treated. Let D_j be the treatment status of the jth family. Denote by $Z_{i,j}$ the vector of observed characteristics of the ith sibling in the jth family used in determining treatment status and by $U_{i,j}$ the vector of unobserved characteristics of the ith sibling in the jth family used in determining treatment status. In our empirical analysis,

$$Z_{i,j} = (G_{i,j}, SES_j, IQ_{i,j}, W_{i,j}) \ ,$$

where $G_{i,j}$ is the gender of the ith sibling in the jth family, $SES_{i,j}$ is the measure of socio-economic status of the jth family, $IQ_{i,j}$ is the Stanford-Binet IQ score at study entry of the ith sibling in the jth family, and $W_{i,j}$ is the cohort or wave of the ith sibling in the jth family. In this notation, the kth observed outcome of the ith sibling in the jth family is

$$Y_{i,j,k} = D_j Y_{i,j,k}(1) + (1 - D_j) Y_{i,j,k}(0) \ .$$

Recall that only the characteristics of the eldest sibling in each family matter for determining treatment status. We will therefore drop the dependence on i and henceforth simply write Z_j in place of $Z_{i^*,j}$ where

$$i^* = \arg\min_{i \in I_j} W_{i,j} \ .$$

In light of the description of the randomization procedure in Section 2.2, we interpret $U_{i,j}$ as an indicator of whether the ith sibling in the jth family has a single mother who (at the date of enrollment of the eldest sibling) was working and unavailable for weekly home visits. Since this variable does not depend on i, we will henceforth drop the dependence on i and simply write U_j. Further define MW_j to be an indicator for whether the jth family has a single mother who (at the date of enrollment of the eldest sibling) was working. Although this variable is not used directly in the assignment of treatment status, we must, of course, have $U_j = 0$ whenever $MW_j = 0$.

It is useful to introduce the following shorthand notation. Define

$$\begin{aligned} D &= (D_j : j \in J) \\ Z &= (Z_j : j \in J) \\ U &= (U_j : j \in J) \\ MW &= (MW_j : j \in J) \ . \end{aligned}$$

For $d \in \text{supp}(D)$ and $k \in K$, further define

$$\begin{aligned} Y_k &= (Y_{i,j,k} : i \in I_j, j \in J) \\ Y_k(d) &= (Y_{i,j,k}(d_j) : i \in I_j, j \in J) \ . \end{aligned}$$

Denote by P the distribution of

$$((Y_k(d) : d \in \text{supp}(D), k \in K), D, Z, U, MW) \ ,$$

which is assumed to lie in a class of distributions Ω, i.e.,

$$((Y_k(d) : d \in \text{supp}(D), k \in K), D, Z, U, MW) \sim P \in \Omega \ .$$

The assumptions we impose on Ω are presented in Section 3.2 below. For $k \in K$, let

$$\omega_k = \{P \in \Omega : Y_k(d) \text{ does not depend on } d\} \ .$$

In this notation, our goal is to test the family of null hypotheses

$$H_k : P \in \omega_k \text{ for } k \in K \tag{1}$$

in a way that controls in finite samples the familywise error rate – the probability of even one false rejection.[1] More formally, let $K_0(P)$ denote the set of true null hypotheses, i.e.,

$$K_0(P) = \{k \in K : P \in \omega_k\} \ ,$$

and define

$$FWER_P = P\{\text{reject } \geq 1 \text{ hypothesis } H_k \text{ with } k \in K_0(P)\} \ .$$

In this notation, our goal is to test the family of null hypotheses (1) in a way that satisfies

$$FWER_P \leq \alpha \text{ for all } P \in \Omega \tag{2}$$

for some pre-specified value of $\alpha \in (0, 1)$.

Before proceeding to a formal description of our testing procedure, it is useful to model the distribution of D. Let $\tilde{D}$ be a vector of treatment assignments produced from Steps 1′-3′ above, i.e., according to the initial randomization before any reassignment of treatment status. Let

$$\delta : \{0, 1\}^{|J|} \times \text{supp}(Z, U) \to \{0, 1\}^{|J|}$$

be the rule used to exchange participants from the treatment group to the control group in Steps 4′ and 5′. It is helpful to decompose δ into two functions in the following way. Let

$$\delta_1 : \{0, 1\}^{|J|} \times \text{supp}(Z) \to \{0, 1\}^{|J|}$$

be the rule used to exchange participants from the treatment group to the control group in Step 4′. In an analogous fashion, let

$$\delta_2 : \{0, 1\}^{|J|} \times \text{supp}(U) \to \{0, 1\}^{|J|}$$

be the rule used to move participants with single mothers who were working and unavailable for the weekly home visits from the treatment group to the control group in Step 5′. In this notation, D can be written as

[1] The null hypotheses specified in (1) are sometimes referred to as "sharp" null hypotheses to distinguish them from "weak" null hypotheses that specify instead that $E[Y_k(d)]$ does not depend on d. For a discussion of how randomization tests may be used to test such null hypotheses, see, e.g., Chung and Romano (2013), Bugni et al. (2018) and Bai et al. (2022a).

the composition of two functions:

$$D = \delta_2(\delta_1(\tilde{D}, Z), U) = \delta(\tilde{D}, Z, U) \ .$$

Remark 3.1. By requiring that our testing procedure satisfy criterion (2), all of the null hypotheses rejected by our procedure are false with probability at least $1 - \alpha$. The recent literature on multiple testing has considered error rates less stringent than the familywise error rate (see, e.g., Romano, Shaikh, and Wolf, 2010). One example is the m-familywise error rate – the probability of m or more false rejections for some $m \geq 1$. Another example is the false discovery proportion —the ratio of false rejections to total rejections (defined to be zero when there are no rejections at all)—where $P\{FDP > \gamma\}$ for some $\gamma \in [0, 1)$, and here FDP is the false discovery proportion. With such error rates, one is only guaranteed that, with probability at least $1 - \alpha$, "most" of the null hypotheses rejected by the procedure are false. However, such procedures may have much greater ability to detect false null hypotheses. This feature may be especially valuable when the number of null hypotheses under consideration is very large. See Romano and Shaikh (2006a), Romano and Shaikh (2006b) and Romano et al. (2008) for a discussion of some procedures for control of such error rates. We do not pursue such error rates here because in our application the number of null hypotheses under consideration is relatively small. ■

3.2 Assumptions

In this section, we describe the assumptions we impose on Ω. These assumptions are connected tightly to our description of the randomization procedure in Section 2.2. We first state our assumptions formally and then relate them briefly to our description of the way in which treatment status was assigned.

Some of our assumptions are most succinctly stated in terms of groups of transformations. Here, we use the term group as it is used in mathematics. See, e.g., Dummit and Foote (1999) or any other standard reference. To this end, let $\mathbf{G}$ be the set of permutations of $|J|$ elements. This set forms a group under the usual composition of functions. Define the action of $g \in \mathbf{G}$ on $|J|$-dimensional vectors v by

$$gv = (v_{g(1)}, \ldots, v_{g(|J|)}) \ .$$

Let $\mathbf{H} = \{-1, 1\}^{|J|}$. This set forms a group under component-wise multiplication. Define the action of $h \in \mathbf{H}$ on $|J|$-dimensional vectors v by the rule that the jth element of hv equals v_j if $h_j = 1$ and $1 - v_j$ if $h_j = -1$. For $z \in \text{supp}(Z)$, let

$$\mathbf{H}_z = \{h \in \mathbf{H} : h_j = h_{j'} \text{ whenever } w_j = w_{j'}\} \ .$$

Here, w_j is the component of z corresponding to the wave in which the eldest sibling in the jth family was enrolled in the program. In other words, $\mathbf{H}_z$ is the subgroup of $\mathbf{H}$ that is constant across families whose treatment status was determined in the same wave. Using this notation, we may now state the assumptions that will underlie our analysis.

Assumption 3.1. For any $P \in \Omega$, $(Y_k(d) : d \in \text{supp}(D), k \in K) \perp\!\!\!\perp D | Z, U$ under P.

Assumption 3.2. For any $g \in \mathbf{G}$, $\delta_1(gd, gz) = g\delta_1(d, z)$.

Assumption 3.3. For any $h \in \mathbf{H}_z$, $h\delta_1(d,z) = \delta_1(hd,z)$.

Assumption 3.4. The jth component of $\delta_2(d,u)$ equals zero if $d_j = 1$ and $u_j = 1$; otherwise, jth component of $\delta_2(d,u)$ equals d_j.

Assumption 3.5. For any $P \in \Omega$, $U_j = 0$ if $MW_j = 0$ w.p.1. under P.

Our first assumption simply states that our description of the way in which treatment status was assigned in Section 2.2 is accurate in the sense that the only variables used to determine treatment status that affect potential outcomes are Z and U. Hence, potential outcomes are independent of treatment status conditional on Z and U. Assumption 3.2 is a mild equivariance restriction that will be satisfied provided that the way in which treatment status is reassigned in Step 4′ does not depend on the order of the participants themselves. Informally, it says that "ordering of participants doesn't matter." Assumption 3.3 further imposes a mild symmetry requirement on the way in which treatment status is reassigned in Step 4′. Informally, it says that "the 'odd' and 'even' labels don't matter." Assumption 3.4 simply defines the function δ_2 so that it agrees with Step 5′ in the description of the randomization procedure in Section 2.2, i.e., participants in the treatment group with single mothers who were working and unavailable for the weekly home visits are moved to the control group. Finally, Assumption 3.5 imposes the logical restriction that U_j and MW_j described in Section 3.1, i.e., $U_j = 0$ whenever $MW_j = 0$. In other words, for a family to have a single mother who is working and unavailable for the weekly home visits, it must obviously be the case that the family has a single mother who is working.

4 Testing Procedures

In Section 4.2 below, we develop methods for testing a single (joint) null hypothesis of the form

$$H_L : P \in \omega_L \ , \tag{3}$$

where

$$\omega_L = \bigcap_{k \in L} \omega_k$$

for $L \subseteq K$, in a way that controls the usual probability of a Type I error at level α. In Section 4.3, we extend these methods to test the family of null hypotheses (1) so that it satisfies (2).

Our methods for testing (3) in a way that controls the usual probability of a Type I error will be based on the general principle behind randomization tests of exploiting certain symmetries in the distribution of the observed data. Here, by a symmetry in the distribution of the observed data we mean that there is a group of transformations of the observed data that leave its distribution unchanged whenever the null hypothesis is true. When this is the case, it is possible to construct a test of the null hypothesis that controls the usual probability of a Type I error in finite samples. Perhaps the most familiar example of a randomization test is a permutation test, which may be used to test the null hypothesis that two i.i.d. samples from possibly distinct distributions are in fact from the same underlying distribution, but, as explained in Section 15.2 of Lehmann and Romano (2005), the principle applies more generally. Recently, randomization tests have been employed

in a wide variety of settings, including settings with staggered treatment adoption (Shaikh and Toulis, 2021), experiments with covariate-adaptive randomization (Bugni et al., 2018; Bai et al., 2022b), experiments with interference (Basse et al., 2019), settings with a "small" number of clusters (Canay et al., 2017, 2021; Cai et al., 2023), regression kink designs (Ganong and Jäger, 2018) and regression discontinuity designs (Canay and Kamat, 2017). The main challenge in applying these ideas in our setting lies in finding symmetries in the distribution of treatment status that persist despite the complicated way in which treatment status was assigned in the HighScope Perry Preschool program. These symmetries are developed in Lemma 4.1, which is presented in Section 4.1 below, by exploiting Assumptions 3.2–3.3 in conjunction with Assumption 3.4.

4.1 A Useful Lemma

In order to describe the symmetries in the distribution of the observed data that we will exploit formally, we require some further notation. For $(z,u) \in \text{supp}(Z,U)$, let $\mathbf{G}_{z,u}$ be the subgroup of $\mathbf{G}$ that only contains $g \in \mathbf{G}$ such that

$$g(j) = j' \Longrightarrow (z_j, u_j) = (z_{j'}, u_{j'}) \ .$$

In particular, $g \in \mathbf{G}_{Z,U}$ will therefore act on a $|J|$-dimensional binary vector of treatment statuses by permuting treatment status among those families with the same observed and unobserved characteristics (defined by the characteristics of the eldest child in the case of families with multiple children). For $(z,u) \in \text{supp}(Z,U)$, let

$$\mathbf{H}_{z,u} = \{uh : h \in \mathbf{H}_z\} \ ,$$

where the jth element of uh equals h_j if $u_j = 0$ and 1 if $u_j = 1$. The action of $h \in \mathbf{H}_{z,u}$ on $|J|$-dimensional vectors v is defined as it was for $\mathbf{H}$ and $\mathbf{H}_z$. In particular, $h \in \mathbf{H}_{Z,U}$ will therefore act on a $|J|$-dimensional binary vector of treatment statuses by possibly "flipping" treatment status for all families whose treatment status was determined in the same wave except for those with single mothers who were working and unavailable for the weekly home visits (at the date of enrollment of the eldest sibling). Using this notation, we may now state the lemma.

Lemma 4.1. *Let $g \in \mathbf{G}_{Z,U}$ and $h \in \mathbf{H}_{Z,U}$. Suppose $\tilde{D}$ is distributed as described in Section 3. Then, the following statements hold:*

(i) If Assumptions 3.2 and 3.4 hold, then

$$g\delta(\tilde{D}, Z, U)|Z, U \overset{d}{=} \delta(\tilde{D}, Z, U)|Z, U \ . \tag{4}$$

(ii) If Assumptions 3.3 and 3.4 hold, then

$$h\delta(\tilde{D}, Z, U)|Z, U \overset{d}{=} \delta(\tilde{D}, Z, U)|Z, U \ . \tag{5}$$

(iii) If Assumptions 3.2–3.4 hold, then

$$hg\delta(\tilde{D}, Z, U)|Z, U \overset{d}{=} \delta(\tilde{D}, Z, U)|Z, U \ . \tag{6}$$

PROOF: In order to establish (i), first note that by definition of $\tilde{D}$ and $\mathbf{G}_{Z,U}$, we have that

$$g\tilde{D}|Z,U \stackrel{d}{=} \tilde{D}|Z,U \ . \tag{7}$$

Next, note for any $g' \in \mathbf{G}$, we have that

$$\begin{aligned} \delta(g'd, g'z, g'u) &= \delta_2(\delta_1(g'd, g'z), g'u) \\ &= \delta_2(g'\delta_1(d,z), g'u) \\ &= g'\delta_2(\delta_1(d,z), u) \\ &= g'\delta(d,z,u) \ , \end{aligned} \tag{8}$$

where the first and fourth equalities follow from the definition of δ, the second equality follows from Assumption 3.2, and the third equality follows from Assumption 3.4. Finally, for any $A \subseteq \{0,1\}^{|J|}$, note that

$$\begin{aligned} P\{g\delta(\tilde{D},Z,U) \in A|Z,U\} &= P\{\delta(g\tilde{D},gZ,gU) \in A|Z,U\} \\ &= P\{\delta(g\tilde{D},Z,U) \in A|Z,U\} \\ &= P\{\delta(\tilde{D},Z,U) \in A|Z,U\} \ , \end{aligned}$$

where the first equality follows from (8), the second follows from the definition of $\mathbf{G}_{Z,U}$, and the third from (7).

In order to establish (ii), first choose $h^*(h') \in \mathbf{H}_z$ for each $h' \in \mathbf{H}_{z,u}$ such that $uh^*(h') = h'$. Next, note that by the definition of $\tilde{D}$ and $\mathbf{H}_Z$, we have that

$$h^*(h)\tilde{D}|Z,U \stackrel{d}{=} \tilde{D}|Z,U \ . \tag{9}$$

Further observe that Assumption 3.4 implies for any $h' \in \mathbf{H}_{z,u}$ that

$$h'\delta_2(d,u) = \delta_2(h^*(h')d, u) \ . \tag{10}$$

Hence, for any $h' \in \mathbf{H}_{z,u}$,

$$\begin{aligned} h'\delta(d,z,u) &= h'\delta_2(\delta_1(d,z),u) \\ &= \delta_2(h^*(h')\delta_1(d,z),u) \\ &= \delta_2(\delta_1(h^*(h')d,z),u) \\ &= \delta(h^*(h')d,z,u) \ , \end{aligned} \tag{11}$$

where the first and fourth equalities follow from the definition of δ, the second equality follows from (10), and the third equality follows from Assumption 3.3. Finally, for any $A \subseteq \{0,1\}^{|J|}$, note that

$$\begin{aligned} P\{h\delta(\tilde{D},Z,U) \in A|Z,U\} &= P\{\delta(h^*(h)\tilde{D},Z,U) \in A|Z,U\} \\ &= P\{\delta(\tilde{D},Z,U) \in A|Z,U\} \ , \end{aligned}$$

where the first equality follows from (11) and the second follows from (9).

Part (iii) follows immediately from parts (i) and (ii), which completes the proof. ■

In Sections 4.2 and 4.3 below, we employ the symmetries in the distribution of treatment status described in Lemma 4.1 above to develop tests of (3) and (1).

4.2 Testing a Single (Joint) Null Hypothesis

In order to describe our test of the single (joint) null hypothesis (3) for $L \subseteq K$, we first require a test statistic. To this end, define

$$X_L = ((Y_k : k \in L), D, Z)$$

and let

$$T_L = T_L(X_L)$$

be a test statistic for testing (3). Note that we impose the mild requirement that T_L only depends on X_L. In particular, we assume that it does not depend on Y_k with $k \notin L$. We assume further that large values of T_L provide evidence against the null hypothesis.

We now describe the construction of a critical value for our test. For this purpose, the following lemma is useful:

Lemma 4.2. *If $P \in \omega_L$ and Assumption 3.1 holds, then*

$$(Y_k : k \in L) \perp\!\!\!\perp D | Z, U$$

under P.

Proof: Consider $P \in \omega_L$. Assumption 3.1 implies that

$$(Y_k(d) : d \in \text{supp}(D), k \in L) \perp\!\!\!\perp D | Z, U$$

under P. Since $P \in \omega_L$, we have further that $Y_k(d) = Y_k$ for all $k \in L$. The desired result thus follows. ■

In order to describe an important implication of Lemma 4.2, it is useful to define

$$hgX_L = ((Y_k : k \in L), hgD, Z)$$

for $g \in \mathbf{G}_{Z,u}$ and $h \in \mathbf{H}_{Z,u}$. If Assumptions 3.1–3.4 hold, then Lemmas 4.1–4.2 together imply that

$$(X_L, U) | Z, U \overset{d}{=} (hgX_L, U) | Z, U \tag{12}$$

whenever $P \in \omega_L$, $g \in \mathbf{G}_{Z,U}$ and $h \in \mathbf{H}_{Z,U}$. This symmetry suggests that we can construct a critical value with which to compare our test statistic by re-evaluating it at hgX_L for each $g \in \mathbf{G}_{Z,U}$ and $h \in \mathbf{H}_{Z,U}$. As mentioned previously, U is unknown, but its possible values can be limited by Assumptions 3.4–3.5 to the

set $\mathbf{U}(D, MW)$, where

$$\mathbf{U}(d, mw) = \{u \in \{0,1\}^{|J|} : u_j = 0 \text{ whenever } d_j = 1 \text{ or } mw_j = 0\} \ .$$

In other words, we may use as our critical value

$$\bar{c}_L(X_L, 1-\alpha) = \max_{u \in \mathbf{U}(D,MW)} c_L(X_L, u, 1-\alpha) \ , \tag{13}$$

where

$$c_L(X_L, u, 1-\alpha) = \inf \left\{ t \in \mathbf{R} : \frac{1}{|\mathbf{G}_{Z,u}||\mathbf{H}_{Z,u}|} \sum_{g \in \mathbf{G}_{Z,u}, h \in \mathbf{H}_{Z,u}} I\{T_L(hgX_L) \leq t\} \geq 1-\alpha \right\} \ ,$$

where $I\{\cdot\}$ is the indicator function. It is worth noting that in our setting $|\mathbf{U}(D, MW)| = 2^{18}$. This idea is formalized in the following theorem:

Theorem 4.1. *Under Assumptions 3.1–3.5, the test that rejects H_L whenever*

$$T_L(X_L) > \bar{c}_L(X_L, 1-\alpha) \ ,$$

where $\bar{c}_L(X_L, 1-\alpha)$ is defined by (13) controls the usual probability of a Type I error at level α, i.e.,

$$P\{T_L(X_L) > \bar{c}_L(X_L, 1-\alpha)\} \leq \alpha$$

for all $P \in \omega_L$.

PROOF: Consider $P \in \omega_L$. Define

$$\phi(X_L, u) = I\{T_L(X_L) > c_L(X_L, u, 1-\alpha)\} \ .$$

From Assumptions 3.4 and 3.5, we have that $U \in \mathbf{U}(D, MW)$. Hence,

$$\bar{c}_L(X_L, 1-\alpha) \geq c_L(X_L, U, 1-\alpha) \ . \tag{14}$$

It therefore suffices to show that

$$E_P[\phi(X_L, U)] \leq \alpha \ . \tag{15}$$

To this end, first note under Assumptions 3.1–3.4 that it follows from Lemmas 4.1–4.2 for any $g \in \mathbf{G}_{Z,U}$ and $h \in \mathbf{H}_{Z,U}$ that (12) holds under any such P. Next, note that

$$\begin{aligned} E_P\left[\sum_{g \in \mathbf{G}_{Z,U}, h \in \mathbf{H}_{Z,U}} \phi(hgX_L, U)|Z, U\right] &= \sum_{g \in \mathbf{G}_{Z,U}, h \in \mathbf{H}_{Z,U}} E_P[\phi(hgX_L, U)|Z, U] \\ &= \sum_{g \in \mathbf{G}_{Z,U}, h \in \mathbf{H}_{Z,U}} E_P[\phi(X_L, U)|Z, U] \\ &= |\mathbf{G}_{Z,U}||\mathbf{H}_{Z,U}| E_P[\phi(X_L, U)|Z, U] \ , \end{aligned} \tag{16}$$

On the other hand, since

$$c_L(hgX_L, U, 1-\alpha) = c_L(X_L, U, 1-\alpha)$$

for any $g \in \mathbf{G}_{Z,U}$ and $h \in \mathbf{H}_{Z,U}$, we also have that

$$E_P\left[\sum_{g\in\mathbf{G}_{Z,U}, h\in\mathbf{H}_{Z,U}} \phi(hgX_L, U)|Z, U\right] \leq |\mathbf{G}_{Z,U}||\mathbf{H}_{Z,U}|\alpha \ . \tag{17}$$

It follows from (16) and (17) that

$$E_P[\phi(X_L, U)|Z, U] \leq \alpha \ ,$$

from which the desired conclusion (15) follows immediately. ■

Remark 4.1. Once (12) is established, the proof of Theorem 4.1 follows the usual arguments that underlie the validity of randomization tests. See, e.g., Chapter 15 of Lehmann and Romano (2005) for a textbook discussion of such methods. Nevertheless, we include the details of the argument for completeness. ■

Remark 4.2. Note that $c_L(X_L, u, 1-\alpha)$ defined in (4.2) requires computing $T_L(hgX_L)$ for every $g \in \mathbf{G}_{Z,u}$ and $h \in \mathbf{H}_{Z,u}$. In our setting, the sets $\mathbf{G}_{Z,u}$ and $\mathbf{H}_{Z,u}$ are sufficiently small that the construction of the critical value is computationally feasible. In other settings, this may not be the case and one may need to resort to a stochastic approximation to the critical value. This can be done without affecting the finite-sample validity of the resulting test. See Section 15.2 of Lehmann and Romano (2005) for details. ■

Remark 4.3. It is straightforward to include additional "exogenous" variation in the way that treatment status was reassigned. Here, by "exogenous" variation we mean variation unrelated to outcomes, but used in determining treatment status. Such variation could be useful, for instance, if in Step 3 of the randomization procedure there was more than one way to exchange participants across the two groups in order to "balance" gender and socio-economic status scores. For example, we could allow δ to depend on an additional random variable V that enters δ_1 if

$$gV|Z, U \stackrel{d}{=} V|Z, U$$

for any $g \in \mathbf{G}$, Assumption 3.2 were strengthened so that

$$\delta_1(gd, gz, gv) = g\delta_1(d, z, v)$$

for any $g \in \mathbf{G}$, and Assumption 3.3 were strengthened so that $h\delta_1(d, z, v) = \delta_1(hd, z, v)$ for any $h \in \mathbf{H}_z$. Under these conditions, it follows by arguing as in the proof of Lemma 4.1 that (6) holds, from which the rest of our arguments would follow. In particular, our testing procedures would remain unchanged even if we were to allow for this type of additional variation. ■

Remark 4.4. An inspection of the proof of Theorem 4.1 reveals that the validity of our test hinges crucially on part (iii) of Lemma 4.1. On the other hand, there is no reason to suspect that

$$g\delta(\tilde{D}, Z, U)|Z, U \stackrel{d}{=} \delta(\tilde{D}, Z, U)|Z, U$$

for $g \in \mathbf{G}$. For this reason, a test of (3) based simply off of permutations from $\mathbf{G}$ does not necessarily control the usual probability of a Type I error. Nevertheless, because such a test has been applied in earlier analyses

of the HighScope Perry Preschool program, we include it in our comparisons below. ■

Remark 4.5. In addition to the "naïve" permutation test described in Remark 4.4, Heckman et al. (2010a) consider a test of (3) based on permutations from $\mathbf{G}_z$, where, by analogy with the definition of $\mathbf{G}_{z,u}$ given earlier, $\mathbf{G}_z$ is the subgroup of $\mathbf{G}$ that contains only $g \in \mathbf{G}$ such that

$$g(j) = j' \Longrightarrow z_j = z_{j'} \ .$$

It is possible to justify such an approach using Lemma 4.1 provided that one assumes that the way in which treatment status was reassigned in Step 5 of the randomization procedure depended only on whether the participant had a single mother who was working. If one were willing to make such an assumption, then one could simply expand Z so as to include MW and ignore the effect of δ_2 on treatment status (e.g., by setting all elements of U equal to zero). Under Assumptions 3.2 and 3.4, it then follows from part (i) of Lemma 4.1 that

$$gD|Z \stackrel{d}{=} D|Z$$

for $g \in \mathbf{G}_Z$. On the other hand, because MW was used in an asymmetric fashion to reassign treatment status, Assumption 3.3 is no longer plausible, so it is not reasonable to expect parts (ii) and (iii) of Lemma 4.1 to apply. Unfortunately, the number of permutations in $\mathbf{G}_Z$ alone is too small to be useful. Heckman et al. (2010a) therefore impose additional assumptions, such as parametric restrictions about the way in which certain observed characteristics affect outcomes, to make use of this limited number of permutations. Note further that the resulting approach does not have the finite-sample validity of the approach developed here. ■

4.3 Testing Multiple Null Hypotheses

We now return to the problem of testing the family of null hypotheses (1) in a way that satisfies (2). Under Assumptions 3.2–3.5, it is straightforward to calculate a p-value $\hat{p}_k$ for each H_k using Theorem 4.1 by simply applying the theorem with $L = \{k\}$ and computing the smallest value of α for which the null hypothesis is rejected. The resulting p-values will satisfy

$$P\{\hat{p}_k \leq u\} \leq u$$

for all $u \in (0, 1)$ and $P \in \omega_k$. A crude solution to the multiplicity problem would therefore be to apply a Bonferroni or Holm-type correction. Such an approach would indeed satisfy (2), as desired, but implicitly relies upon a "least favorable" dependence structure among the p-values. To the extent that the true dependence structure differs from this "least favorable" one, improvements may be possible. For that reason, we apply a stepwise multiple testing procedure developed by Romano and Wolf (2005) for control of the familywise error rate that implicitly incorporates information about the dependence structure when deciding which null hypotheses to reject. Our discussion follows that in Romano and Shaikh (2010), wherein the algorithm is generalized to allow for possibly uncountably many null hypotheses.

In order to describe our testing procedure, we first require a test statistic for each null hypothesis such that large values of the test statistic provide evidence against the null hypothesis. As before, we impose the

requirement that the test statistic for H_k depends only on $X_{\{k\}}$. Denote such a test statistic by $T_k(X_{\{k\}})$. Next, for $L \subseteq K$, define

$$T_L(X_L) = \max_{k \in L} T_k(X_{\{k\}}) \ .$$

Finally, for $L \subseteq K$, denote by $\bar{c}_L(X_L, 1-\alpha)$ the critical value defined in (13) with this choice of $T_L(X_L)$.

Our testing procedure is summarized in the following algorithm:

Algorithm 4.1.

Step 1: Set $L_1 = K$. If

$$\max_{k \in L_1} T_k(X_{\{k\}}) \leq \bar{c}_{L_1}(1-\alpha) \ ,$$

then stop and reject no null hypotheses; otherwise, reject any H_k with

$$T_k(X_{\{k\}}) > \bar{c}_{L_1}(X_{L_1}, 1-\alpha)$$

and go to Step 2.

$\vdots$

Step j: Let L_j denote the indices of remaining null hypotheses. If

$$\max_{k \in L_j} T_k(X_{\{k\}}) \leq \bar{c}_{L_j}(X_{L_j}, 1-\alpha) \ ,$$

then stop and reject no further null hypotheses; otherwise, reject any H_k with

$$T_k(X_{\{k\}}) > \bar{c}_{L_j}(X_{L_j}, 1-\alpha)$$

and go to Step $j+1$.

$\vdots$

Theorem 4.2. *Under Assumptions 3.1–3.5, Algorithm 4.1 satisfies* (2).

Proof: The claim follows from Theorem 4.1 and arguments given in Romano and Wolf (2005) or Romano and Shaikh (2010). Since the argument is brief, we include it here for completeness.

Suppose that a false rejection occurs. Let $\hat{j}$ be the *smallest* step at which a false rejection occurs. By the minimality of $\hat{j}$, we must have that

$$L_{\hat{j}} \supseteq K_0(P) \ . \tag{18}$$

It follows that

$$\bar{c}_{L_{\hat{j}}}(X_{L_{\hat{j}}}, 1-\alpha) \geq \bar{c}_{K_0(P)}(X_{K_0(P)}, 1-\alpha) \ . \tag{19}$$

Since a false rejection occurred, we must also have that

$$\max_{k \in K_0(P)} T_k(X_{\{k\}}) > \bar{c}_{L_{\hat{j}}}(X_{L_{\hat{j}}}, 1-\alpha) \ .$$

Hence,

$$\max_{k \in K_0(P)} T_k(X_{\{k\}}) > \bar{c}_{K_0(P)}(X_{K_0(P)}, 1-\alpha) \ ,$$

and the probability of this event is bounded above by α by Theorem 4.1. ■

Remark 4.6. It is straightforward to calculate a multiplicity-adjusted p-value $\hat{p}_k^{\text{adj}}$ for each H_k using Theorem 4.2 by simply computing the smallest value of α for which each null hypothesis is rejected. The resulting p-values have the property that the procedure that rejects any H_k with $\hat{p}_k^{\text{adj}} \leq \alpha$ satisfies (2). ■

Remark 4.7. The choice of $T_k(X_{\{k\}})$ in Algorithm 4.1 is arbitrary, but we apply it to the HighScope Perry Preschool data with $T_k(X_{\{k\}})$ given by a Studentized difference in means between the treatment and control groups for all outcomes except cognitive outcomes, in which case we use a Mann-Whitney U-statistic. Of course, one could just as well use a more omnibus statistic, such as a Kolmogorov-Smirnov statistic. ■

5 Results

We now apply the methodology developed in the preceding section to the HighScope Perry Preschool data. We find that the program has statistically significant effects on a wide range of outcomes even after controlling for (i) the imperfectness of the randomization and (ii) the multiplicity of the null hypotheses under consideration. Recall that (i) involves (a) the way in which treatment status was reassigned to "balance" certain observed characteristics as well as (b) the way in which some participants were removed from the treatment group and placed in the control group on the basis of unobserved characteristics. We address (i) by exploiting symmetries in the distribution of treatment status that remain valid in the presence of both (a) and (b) together with information limiting the extent of (b). We address (ii) by demanding control of the familywise error rate, thereby eliminating concerns about selectively reporting results for only a subset of these null hypotheses.

When applying Theorem 4.1 and Theorem 4.2 in this empirical setting, we discretize SES_j as an indicator denoting whether SES_j exceeds the median value among all families in the same wave. There is no loss of generality with this approach if we assume that the goal of Step 3 of the randomization procedure was to "balance" the two groups so that their respective median SES_j values were the same. We note, however, that because we exploit $\mathbf{H}_{Z,u}$ as well as $\mathbf{G}_{Z,u}$, our inferences would remain nontrivial even if we were to adopt a much finer discretization of SES_j. Indeed, they would remain valid even if the discretization were so fine that $\mathbf{G}_{Z,u}$ became a singleton consisting of only the identity permutation for all $u \in \mathbf{U}(D, MW)$.

Following Heckman et al. (2010a), we analyze seven conceptually distinct "blocks" of outcomes, each of which is of independent interest: one is related to IQ measures, a second to achievement measures, a third to educational attainment, a fourth to criminal activity, and three to employment at ages 19, 27, and 40. We divide the data further by gender. We correct for the multiplicity of outcomes within each of these fourteen blocks of outcomes. Because of our limited sample size, we adopt the convention that null hypotheses with p-values less than or equal to .10 are statistically significant.

The results of our analysis are presented in Tables 1 and 2 for males and females, respectively. The first column of each table displays the outcome analyzed. The second column gives the age at which the

outcome is measured. The third and fourth columns contain, respectively, the mean value of the outcome for the control group and the difference in means between the treatment group and the control group. The remaining columns present p-values from various testing procedures:

- The column under the heading "Asymp." presents (multiplicity) unadjusted p-values from a one-sided test based on comparing a Studentized difference of means with a critical value computed from a normal approximation.
- The two columns under the heading "Naïve" display, respectively, the unadjusted and adjusted p-values based on the naïve application of a permutation test in this setting. In other words, these p-values are based on the unrestricted set of permutations $\mathbf{G}$ rather than $\mathbf{G}_{Z,u}$ and $\mathbf{H}_{Z,u}$.
- The two columns under the heading "$U = 0$" display, respectively, the unadjusted and adjusted p-values derived from applying Theorem 4.1 and Theorem 4.2 assuming that $\mathbf{U}(D, MW) = \{\{0\}^{|J|}\}$, i.e., ignoring the effect of Step 5 of the randomization procedure.
- The two columns under the heading "Max-U" display, respectively, the unadjusted and adjusted p-values derived from applying Theorems 4.1 and 4.2.

Note that the "Naïve" p-values do not account for the imperfectness in the randomization stemming from either (a) or (b) above. For that reason, as discussed in Remark 4.4, there is no reason to suspect that these p-values are valid, but they are included here for comparison. Note further that by construction the "Max-U" (un)adjusted p-values are smaller than the "$U = 0$" (un)adjusted p-values. The "Naïve" (un)adjusted p-values, however, may be either larger or smaller than the "$U = 0$" (un)adjusted p-values.

Our findings are broadly consistent with those in Heckman et al. (2010a). They are summarized as follows:

Cognition: The top panels of Tables 1 and 2 present our evidence on cognitive abilities as measured by Stanford-Binet IQ score at different ages and various California Achievement Test (CAT) scores at age 14. The "Naïve" adjusted p-values suggest a statistically significant effect on Stanford-Binet IQ scores for both males and females at young ages. These findings survive the more stringent "Max-U" adjusted p-values for the youngest age. The "Naïve" adjusted p-values also suggest a significant effect on various CAT scores at age 14 for both males and females. These inferences weaken for females in the "Max-U" adjusted p-values, but for males are generally stronger using the "Max-U" adjusted p-values than the "Naïve" adjusted p-values.

Schooling: The third block in Tables 1 and 2 present our findings for four educational attainment measures. None of the adjusted p-values show any significant effect of the program on schooling for males. For females, the "Naïve" and "$U = 0$" adjusted p-values show significant effects for all schooling outcomes, and two of these null hypotheses are rejected even in the "Max-U" adjusted p-values. We find that the effects of the program on High School Graduation and GPA for females remain statistically significant even after accounting for both the imperfectness of the randomization and the multiplicity of null hypotheses.

Crime: The fourth block in Tables 1 and 2 present our findings for four outcomes related to criminal activity. These outcomes are of special importance since reductions in crime are important contributors to the significant rate of return estimates reported in Heckman et al. (2010b). "Total crime cost" includes victimization, police/court, and incarceration costs. See Heckman et al. (2010b) for a more detailed discussion of this variable and its contribution to the rate of return of the program. "Non-victimless charges" refer to felony crimes associated with substantial costs to crime victims. Victimless charges, on the other hand, refer to illegal activities, such as illegal gambling, drug possession, prostitution, and driving without a license plate, that do not produce victims.

The "Naïve" adjusted p-values suggest a statistically significant effect of the program on all outcomes for females and for two outcomes for males. Only one of the significant findings for females survives in the "$U = 0$" and "Max-U" adjusted p-values – "Total charges." On the other hand, we find statistically significant effects on all four outcomes for males in the "$U = 0$" adjusted p-values. Only one of these survives in the "Max-U" adjusted p-values – "Total non-victimless crimes."

Employment: The final three panels in Tables 1 and 2 present our findings for three outcomes related to employment measured at different ages. The "Naïve" adjusted p-values show a statistically significant effect on only one outcome related to employment for males – current employment measured at age 40. The "$U = 0$" adjusted p-values show a statistically significant effect on the "number of jobless months in the past two years measured at age 27." This effect survives even in the "Max-U" adjusted p-values. The "Naïve" adjusted p-values show a significant effect on almost all outcomes for females. The number of statistically significant effects decreases substantially using the "$U = 0$" adjusted p-values, and disappears entirely for outcomes measured at age 27. Only effects on outcomes measured at age 19 persist in the "Max-U" adjusted p-values.

We additionally consider aggregating the outcomes within each of the fourteen blocks described above into a summary index. This index is composed of the average rank of participant i's outcomes across each block of variables. As in our analysis above, we consider males and females separately. In this way, we obtain two families of null hypotheses: one corresponding to the seven summary indices for males, and another corresponding to the seven summary indices for females. The results of this exercise are summarized in Table 3. The "Naïve" adjusted p-values show a statistically significant effect on only achievement scores for males, whereas the "$U = 0$" adjusted p-values show a statistically significant effect on both crime and achievement scores for males. Only the effect on achievement scores for males, however, remains according to the "Max-U" adjusted p-values. Both the "Naïve" adjusted p-values and the "$U = 0$" adjusted p-values show a significant effect on almost all outcomes for females. Effects on IQ, achievement scores and schooling, remain significative in the "Max-U" adjusted p-values.

6 Conclusion

This paper develops and applies a framework for inference about the effects of a program on multiple outcomes when the assignment of treatment status is imperfectly randomized. The key idea that underlies our approach is to make use of information limiting the extent to which randomization is imperfect. Using this

approach, we have constructed under weak assumptions a procedure for testing the family of null hypotheses in which each null hypothesis specifies that the program had no effect on one of several outcomes of interest that controls the familywise error rate in finite samples. We use our methodology to reanalyze data from the HighScope Perry Preschool program. The reported beneficial long-term effects for the HighScope Perry Preschool program are a cornerstone in the argument for early childhood intervention in the United States. We find statistically significant effects of the program for both males and females, thereby showing that some of the criticisms regarding the reliability of this evidence are not justified. We believe our framework will be useful in analyzing other studies where randomization is imperfect, provided that the information limiting the extent to which randomization is imperfect is available, as it is in the case of the HighScope Perry Preschool program.

Table 1: Results for Males

	Outcome	Age	Control Mean	Diff-in-Means	Asymp.	*p*-Values Naïve Unadj.	Naïve Adj.	$U = 0$ Unadj.	$U = 0$ Adj.	Max-U Unadj.	Max-U Adj.
IQ	Stanford-Binet	4	83.08	11.83	**0.000**	**0.000**	**0.001**	**0.001**	**0.001**	**0.008**	**0.008**
	Stanford-Binet	5	84.79	10.61	**0.000**	**0.001**	**0.004**	**0.022**	0.691	**0.077**	0.800
	Stanford-Binet	6	85.82	5.66	**0.019**	**0.048**	0.138	**0.033**	**0.034**	**0.094**	0.102
	Stanford-Binet	7	87.71	3.41	**0.088**	0.168	0.354	0.103	0.172	0.247	0.374
	Stanford-Binet	8	89.05	-0.72	0.598	0.568	0.705	0.599	0.691	0.733	0.800
	Stanford-Binet	9	89.03	-0.63	0.587	0.469	0.673	0.450	0.548	0.631	0.680
	Stanford-Binet	10	86.03	-2.33	0.814	0.645	0.645	0.684	0.691	0.790	0.800
Achievement	CAT, Reading	14	9.00	4.93	**0.076**	**0.076**	0.118	**0.017**	**0.035**	**0.036**	**0.086**
	CAT, Arithmetic	14	8.11	7.89	**0.059**	**0.058**	0.113	**0.032**	**0.035**	**0.086**	**0.086**
	CAT, Language	14	6.54	7.80	**0.024**	**0.021**	**0.057**	**0.001**	**0.004**	**0.012**	**0.027**
	CAT, Language Mechanics	14	6.96	8.59	**0.020**	**0.014**	**0.042**	**0.006**	**0.007**	**0.023**	**0.035**
	CAT, Spelling	14	11.54	6.98	**0.090**	**0.090**	**0.090**	**0.003**	**0.035**	**0.012**	**0.086**
Educational Attainment	HS Graduation	19	0.51	-0.03	0.592	0.595	0.595	0.614	0.674	0.704	0.716
	Vocational Training Certificate	≤40	0.33	0.06	0.300	0.307	0.651	0.341	0.567	0.547	0.608
	Highest Grade Completed	19	11.28	0.08	0.398	0.393	0.637	0.383	0.622	0.410	0.669
	GPA	19	1.79	0.02	0.462	0.453	0.631	0.457	0.674	0.567	0.716
Crime	# Non-Juv. Arrests	≤40	11.72	-4.26	**0.042**	**0.042**	**0.084**	**0.036**	**0.038**	**0.100**	0.115
	Total Crime Cost	≤40	775.90	-351.22	0.151	0.158	0.158	**0.037**	**0.049**	**0.042**	0.143
	# Total Charges	≤40	13.38	-4.38	**0.068**	**0.068**	0.125	**0.049**	**0.049**	0.143	0.143
	# Non-Victimless Charges	≤40	3.08	-1.59	**0.029**	**0.029**	**0.072**	**0.025**	**0.037**	**0.063**	**0.091**
Employment at 19	Current Employment	19	0.41	0.14	0.129	0.128	0.236	**0.050**	0.164	0.224	0.290
	No Job in Past Year	19	0.13	0.11	0.893	0.888	0.888	0.901	0.901	0.922	0.922
	Jobless Months in Past 2 Yrs.	19	3.82	1.47	0.783	0.768	0.831	0.821	0.849	0.873	0.890
Employment at 27	Current Employment	27	0.56	0.04	0.384	0.372	0.372	0.268	0.295	0.485	0.512
	No Job in Past Year	27	0.31	-0.07	0.272	0.268	0.392	0.235	0.295	0.360	0.512
	Jobless Months in Past 2 Yrs.	27	8.79	-3.66	**0.063**	**0.062**	0.118	**0.020**	**0.020**	**0.036**	**0.051**
Employment at 40	Current Employment	40	0.50	0.20	**0.051**	**0.051**	**0.096**	0.103	0.116	0.130	0.146
	No Job in Past Year	40	0.46	-0.10	0.204	0.205	0.205	0.154	0.154	0.216	0.216
	Jobless Months in Past 2 Yrs.	40	10.75	-3.52	**0.085**	**0.079**	0.108	**0.064**	0.116	**0.070**	0.146

This table reports the results for males. The sample size consists of 72 participants, 33 treated and 39 control. The Table shows seven "blocks" of outcomes: (1) Stanford-Binet IQ for ages 4 – 10; (2) Californian Achievement Test (CAT) measured at age 14; (3) Education achievement outcomes at various ages; (4) Crime outcomes; (5) Employment outcomes at age 19; (6) Employment outcomes at age 27; (7) Employment outcomes at age 40. The first column displays the outcome of interest. The second column displays the age at which the outcome was surveyed. The third and fourth columns contain, respectively, the mean value of the outcome for the control group and the difference in means between the treatment group and the control group. The fifth column displays an asymptotic unadjusted one-sided *p*-value based on the Studentized difference in means. The columns under the heading "Naïve" display, respectively, the unadjusted and adjusted *p*-values based on a naïve permutation test. The two columns under the heading "$U = 0$" display, respectively, the unadjusted and adjusted *p*-values derived from applying Theorem 4.1 and Theorem 4.2 ignoring Step 5 of the randomization procedure by setting $\mathbf{U}(D, MW) = \{\{0\}^{|J|}\}$. The two columns under the heading "Max-U" display, respectively, the unadjusted and adjusted *p*-values derived from applying Theorem 4.1 and Theorem 4.2.

Table 2: Results for Females

	Outcome	Age	Control Mean	Diff-in-Means	Asymp.	*p*-Values Naïve Unadj.	Naïve Adj.	$U = 0$ Unadj.	$U = 0$ Adj.	Max-U Unadj.	Max-U Adj.
IQ	Stanford-Binet	4	83.69	12.67	**0.000**	**0.000**	**0.001**	**0.008**	**0.008**	**0.020**	**0.020**
	Stanford-Binet	5	81.65	12.67	**0.003**	**0.001**	0.243	**0.012**	0.203	**0.014**	0.354
	Stanford-Binet	6	87.16	3.75	0.120	**0.084**	0.200	**0.094**	0.164	0.160	0.346
	Stanford-Binet	7	86.00	6.52	**0.030**	**0.031**	**0.092**	0.133	0.137	0.191	0.222
	Stanford-Binet	8	83.60	4.24	0.105	0.219	0.219	0.152	0.164	0.339	0.346
	Stanford-Binet	9	83.04	3.70	0.133	0.169	0.243	0.203	0.203	0.354	0.354
	Stanford-Binet	10	81.79	4.96	**0.087**	0.149	0.270	0.203	0.203	0.267	0.354
Achievement	CAT, Reading	14	8.44	8.06	**0.021**	**0.018**	**0.061**	**0.078**	**0.082**	0.136	0.167
	CAT, Arithmetic	14	6.89	4.93	**0.059**	**0.061**	**0.061**	**0.035**	**0.082**	**0.074**	0.167
	CAT, Language	14	7.83	11.62	**0.004**	**0.003**	**0.008**	**0.008**	**0.070**	**0.020**	0.144
	CAT, Language Mechanics	14	8.83	11.80	**0.006**	**0.004**	**0.061**	**0.047**	**0.082**	**0.097**	0.167
	CAT, Spelling	14	10.72	18.78	**0.004**	**0.002**	**0.061**	**0.043**	**0.082**	**0.082**	0.167
Educational Attainment	HS Graduation	19	0.23	0.61	**0.000**	**0.000**	**0.001**	**0.008**	**0.008**	**0.020**	**0.020**
	Vocational Training Certificate	≤40	0.08	0.16	**0.057**	**0.064**	**0.064**	**0.078**	**0.078**	0.144	0.144
	Highest Grade Completed	19	10.75	1.01	**0.005**	**0.008**	**0.017**	**0.070**	**0.070**	0.113	0.113
	GPA	19	1.53	0.89	**0.000**	**0.001**	**0.004**	**0.039**	**0.039**	**0.082**	**0.082**
Crime	# Non-Juv. Arrests	≤40	4.42	-2.26	**0.050**	**0.048**	**0.066**	**0.020**	0.133	0.121	0.158
	Total Crime Cost	≤40	293.50	-271.33	0.142	**0.020**	**0.066**	**0.024**	0.133	**0.082**	0.158
	# Total Charges	≤40	4.92	-2.68	**0.033**	**0.032**	**0.036**	**0.020**	**0.067**	**0.043**	**0.090**
	# Non-Victimless Charges	≤40	0.31	-0.27	**0.043**	**0.032**	**0.066**	0.125	0.133	0.158	0.158
Employment at 19	Current Employment	19	0.15	0.29	**0.012**	**0.013**	**0.023**	**0.008**	**0.031**	**0.035**	**0.090**
	No Job in Past Year	19	0.58	-0.34	**0.007**	**0.007**	**0.016**	**0.024**	**0.031**	**0.074**	**0.090**
	Jobless Months in Past 2 Yrs.	19	10.42	-5.20	**0.055**	**0.059**	**0.059**	0.125	0.125	0.206	0.206
Employment at 27	Current Employment	27	0.55	0.25	**0.032**	**0.038**	**0.055**	0.110	0.149	0.175	0.198
	No Job in Past Year	27	0.54	-0.29	**0.020**	**0.023**	**0.043**	**0.078**	0.149	0.128	0.175
	Jobless Months in Past 2 Yrs.	27	10.45	-4.21	**0.081**	**0.082**	**0.082**	0.110	0.149	0.166	0.198
Employment at 40	Current Employment	40	0.82	0.02	0.448	0.456	0.456	0.442	0.442	0.567	0.567
	No Job in Past Year	40	0.41	-0.25	**0.029**	**0.041**	**0.084**	**0.047**	**0.070**	0.113	0.160
	Jobless Months in Past 2 Yrs.	40	5.05	-1.05	0.335	0.342	0.432	0.352	0.367	0.540	0.540

This table reports the results for females. The sample size consists of 51 participants, 25 treated and 26 control. The Table shows seven "blocks" of outcomes: (1) Stanford-Binet IQ for ages 4 – 10; (2) Californian Achievement Test (CAT) measured at age 14; (3) Education achievement outcomes at various ages; (4) Crime outcomes; (5) Employment outcomes at age 19; (6) Employment outcomes at age 27; (7) Employment outcomes at age 40. The first column displays the outcome of interest. The second column displays the age at which the outcome was surveyed. The third and fourth columns contain, respectively, the mean value of the outcome for the control group and the difference in means between the treatment group and the control group. The fifth column displays an asymptotic unadjusted one-sided *p*-value based on the Studentized difference in means. The columns under the heading "Naïve" display, respectively, the unadjusted and adjusted *p*-values based on a naïve permutation test. The two columns under the heading "$U = 0$" display, respectively, the unadjusted and adjusted *p*-values derived from applying Theorem 4.1 and Theorem 4.2 ignoring Step 5 of the randomization procedure by setting $\mathbf{U}(D, MW) = \{\{0\}^{|J|}\}$. The two columns under the heading "Max-U" display, respectively, the unadjusted and adjusted *p*-values derived from applying Theorem 4.1 and Theorem 4.2.

Table 3: Results Using Rank Summary Indexes Across Outcome Blocks for Males and Females

	Outcome	Age	Control Mean	Diff-in-Means	Asymp.	p-Values Naïve		$U = 0$		Max-U	
						Unadj.	Adj.	Unadj.	Adj.	Unadj.	Adj.
Males	IQ	4-10	0.476	0.093	**0.040**	**0.039**	0.149	**0.034**	**0.084**	**0.099**	0.135
	Achievement Scores	14	0.483	0.128	**0.015**	**0.016**	**0.080**	**0.001**	**0.001**	**0.008**	**0.008**
	Schooling	19	0.659	0.009	0.414	0.407	0.586	0.358	0.358	0.408	0.408
	Crime	40	0.493	0.128	**0.033**	**0.034**	0.142	**0.036**	**0.062**	**0.084**	0.134
	Employment	19	0.750	0.007	0.459	0.449	0.449	0.244	0.440	0.508	0.546
	Employment	27	0.712	0.062	0.179	0.175	0.360	**0.067**	0.291	0.121	0.361
	Employment	40	0.676	0.082	**0.096**	**0.089**	0.248	**0.079**	0.192	0.101	0.223
Females	IQ	4-10	0.445	0.155	**0.007**	**0.008**	**0.032**	**0.012**	**0.020**	**0.027**	**0.043**
	Achievement Scores	14	0.424	0.208	**0.003**	**0.003**	**0.016**	**0.035**	**0.035**	**0.074**	**0.074**
	Schooling	19	0.568	0.244	**0.000**	**0.000**	**0.000**	**0.008**	**0.008**	**0.019**	**0.019**
	Crime	40	0.617	0.120	**0.075**	**0.076**	0.149	**0.024**	0.129	0.105	0.245
	Employment	19	0.655	0.148	**0.007**	**0.007**	**0.038**	**0.012**	**0.051**	**0.058**	0.128
	Employment	27	0.648	0.171	**0.016**	**0.017**	**0.053**	**0.039**	**0.078**	0.113	0.206
	Employment	40	0.762	0.075	0.176	0.180	0.180	0.149	0.149	0.253	0.253

This table reports the results on the summary index of the average rank across outcome blocks for each gender. There are seven indexes that summarize seven "blocks" of outcomes: (1) IQ; (2) CAT, measured at age 14; (3) Education; (4) Crime; (5) Employment at age 19; (6) Employment at age 27; and (7) Employment at age 40. The first column displays the outcome of interest. The second column displays the age at which the outcome was surveyed. The third and fourth columns contain, respectively, the mean value of the outcome for the control group and the difference in means between the treatment group and the control group. The fifth column displays an asymptotic unadjusted one-sided p-value based on the Studentized difference in means. The columns under the heading "Naïve" display, respectively, the unadjusted and adjusted p-values based on a naïve permutation test. The two columns under the heading "$U = 0$" display, respectively, the unadjusted and adjusted p-values derived from applying Theorem 4.1 and Theorem 4.2 ignoring Step 5 of the randomization procedure by setting $\mathbf{U}(D, MW) = \{\{0\}^{|J|}\}$. The two columns under the heading "Max-U" display, respectively, the unadjusted and adjusted p-values derived from applying Theorem 4.1 and Theorem 4.2.

References

Bai, Y., J. P. Romano, and A. M. Shaikh (2022a). Inference in experiments with matched pairs. *Journal of the American Statistical Association 117*(540), 1726–1737.

Bai, Y., J. P. Romano, and A. M. Shaikh (2022b). Inference in experiments with matched pairs. *Journal of the American Statistical Association 117*(540), 1726–1737.

Basse, G. W., A. Feller, and P. Toulis (2019). Randomization tests of causal effects under interference. *Biometrika 106*(2), 487–494.

Bugni, F. A., I. A. Canay, and A. M. Shaikh (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association 113*(524), 1784–1796.

Cai, Y., I. A. Canay, D. Kim, and A. M. Shaikh (2023). On the implementation of approximate randomization tests in linear models with a small number of clusters. *Journal of Econometric Methods 12*(1), 85–103.

Canay, I. A. and V. Kamat (2017, 10). Approximate permutation tests and induced order statistics in the regression discontinuity design. *The Review of Economic Studies 85*(3), 1577–1608.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017). Randomization tests under an approximate symmetry assumption. *Econometrica 85*(3), 1013–1030.

Canay, I. A., A. Santos, and A. M. Shaikh (2021). The wild bootstrap with a "small" number of "large" clusters. *Review of Economics and Statistics 103*(2), 346–363.

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

Dummit, D. S. and R. Foote (1999). *Abstract Algebra* (2 ed.). Upper Saddle River, NJ: Prentice Hall.

Ganong, P. and S. Jäger (2018). A permutation test for the regression kink design. *Journal of the American Statistical Association 113*(522), 494–504.

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

Heckman, J. J. and G. Karapakula (2019). The Perry pre-sschoolers at late midlife: A study in design specific inference. Unpublished manuscript, University of Chicago.

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

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

Heckman, J. J., R. Pinto, and P. A. Savelyev (2013, October). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review 103*(6), 2052–2086.

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

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses* (3 ed.). New York: Springer-Verlag.

Romano, J. P. and A. M. Shaikh (2006a). On stepdown control of the false discovery proportion. In *Optimality*, pp. 33–50. Institute of Mathematical Statistics.

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

Romano, J. P. and A. M. Shaikh (2010). Inference for the identified set in partially identified econometric models. *Econometrica 78*, 169–211.

Romano, J. P., A. M. Shaikh, and M. Wolf (2008). Formalized data snooping based on generalized error rates. *Econometric Theory 24*(2), 404–447.

Romano, J. P., A. M. Shaikh, and M. Wolf (2010, September). Hypothesis testing in econometrics. *Annual Review of Economics 2*(1), 75–104.

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

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

Shaikh, A. M. and P. Toulis (2021). Randomization tests in observational studies with staggered adoption of treatment. *Journal of the American Statistical Association 116*(536), 1835–1848.