Supplement to "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence From the HighScope Perry Preschool Program": Web Appendices

James Heckman, Seong Hyeok Moon, Rodrigo Pinto,

Peter Savelyev, and Adam Yavitz[1]

University of Chicago

July 22, 2010

[1]James Heckman is Henry Schultz Distinguished Service Professor of Economics at the University of Chicago, Professor of Science and Society, University College Dublin, Alfred Cowles Distinguished Visiting Professor, Cowles Foundation, Yale University, and Senior Fellow, American Bar Foundation. Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev and Adam Yavitz are graduate students at the University of Chicago. A version of this paper was presented at a seminar at the HighScope Perry Foundation, Ypsilanti, Michigan, December 2006; at a conference at the Minneapolis Federal Reserve in December 2007; at a conference on the role of early life conditions at the Michigan Poverty Research Center, University of Michigan, December 2007; at a Jacobs Foundation conference at Castle Marbach, April 2008; at the Leibniz Network Conference on Noncognitive Skills in Mannheim, Germany, May 2008; at an Institute for Research on Poverty conference, Madison, Wisconsin, June 2008; and at a conference on early childhood at the Brazilian National Academy of Sciences, Rio de Janeiro, Brazil, December 2009. We thank the editor and two anonymous referees for helpful comments which greatly improved this draft of the paper. We have benefited from comments received on early drafts of this paper at two brown bag lunches at the Statistics Department, University of Chicago, hosted by Stephen Stigler. We thank all of the workshop participants. In addition, we thank Amanda Agan, Mathilde Almlund, Joseph Altonji, Ricardo Barros, Dan Black, Steve Durlauf, Chris Hansman, Tim Kautz, Paul LaFontaine, Devesh Raval, Azeem Shaikh, Jeff Smith, and Steve Stigler for helpful comments. Our collaboration with Azeem Shaikh on related work has greatly strengthened the analysis in this paper. 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; Mr. Robert Dugger; and NICHD R01HD043411. The views expressed in this presentation are those of the authors and not necessarily those of the funders listed here. Supplementary materials for this paper may be found at http://jenni.uchicago.edu/Perry/.

Contents

A The Perry Preschool Program Experiment and Curriculum

Preschool Overview During each wave of the experiment, each preschool class consisted of 20–25 children of ages 3 to 4. The first wave admitted 4-year-olds who only received 1 year of treatment. The last wave was taught alongside a group of 3-year-olds who were not included in the Perry study. Classes were 2.5 hours every weekday during the regular school year (mid-October through May). The preschool teaching staff of four produced a child-teacher ratio ranging from 5 to 6.25 over the course of the program, with teaching positions filled by former public-school teachers. Teachers had special training for tutoring disadvantaged children and were "certified in elementary, early childhood, and special education" (Schweinhart, Barnes, and Weikart, 1993, p. 32).

Home Visits Weekly home visits lasting $1\frac{1}{2}$ hours were conducted by the preschool teachers. The purpose of these visits was to "involve the mother in the educational process," and "implement the curriculum in the home," (Schweinhart, Barnes, and Weikart, 1993, p. 32). By way of encouraging the mothers' participation, teachers also helped with problems arising in the home during the visit. Occasionally, these visits took the form of field trips to stimulating environments, such as a zoo.

Curriculum The Perry Preschool curriculum was based on the concept of *active learning*, which is centered around play that is based on problem-solving and guided by open-ended questions. Children were encouraged to plan, carry out, and then reflect on their own activities. 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. These factors were then organized into 10 topical categories, such as "creative representation," "classification" (recognizing similarities and differences), "number," and "time."[1] These educational principles were reflected in the types of open-ended questions asked by teachers: for example,"What happened? How did you make that? Can you show me? Can you help another child?" (Schweinhart, Barnes, and Weikart, 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, Barnes, and Weikart, 1993, pp. 32–33).

[1] For a full list, see Schweinhart, Barnes, and Weikart (1993).

B The Basic Evaluation Model

A standard model of program evaluation describes the observed outcome Y_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 treatment and control status for agent i, respectively, and D_i is an 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, $\mathrm{E}(Y_{i,1} \mid D_i = 1) - \mathrm{E}(Y_{i,0} \mid D_i = 0)$, is not generally equal to the average treatment effect, $E(Y_{i,1} - Y_{i,0})$, or to the treatment effect conditional on participation, $\mathrm{E}(Y_{i,1} - Y_{i,0} \mid D_i = 1)$. 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 $(Y_{i,0}, Y_{i,1})$ and D_i, interpreted as a treatment assignment indicator, $(Y_{i,0}, Y_{i,1}) \perp\!\!\!\perp D_i$, where $\perp\!\!\!\perp$ 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_i and the joint distribution of counterfactual outcomes $(Y_{i,0}, Y_{i,1})$. A common feature of compromised experiments is reassignment of treatment and control status by a method different from an ideal randomization. Randomization for the Perry experiment 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 sub-section.

C 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.

C.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 the context. We use lowercase 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 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 $D_i = 0$ if not. 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 D that permutes the position of its elements is represented by gD and is defined as

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

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 are indexed by $\mathcal{I}$. In the main text, 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[2] that swaps the elements of any variable D according to the permutation g.

[2] 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 1 and all other elements equal to 0 within the same row or column. Formally, $a_{ij} \in \{0, 1\}$, $\sum_{j=1}^{L} a_{ij} = 1$, and $\sum_{i=1}^{L} a_{ij} = 1$ for all i, j.

The Randomization Hypothesis Permutation-based inference seeks to test the randomization hypothesis, which states that the joint distribution of some outcome Y is invariant under permutations $g \in \mathscr{G}$, that is, 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 $\forall g \in \mathscr{G},\ (Y, gD) \overset{d}{=} (Y, D)$, where, as in the text, $\overset{d}{=}$ means equality in distribution.

Interpreting the Randomization Hypothesis The hypothesis of no treatment effect for randomized trials is equivalent to the hypothesis of independence between treatment assignments D and outcome Y, as noted in Section 4.3. Suppose $(Y, gD) \overset{d}{=} (Y, D)$ holds. Define $T(Y, D)$ as our test statistic. We assume that it 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 treatment and control participants are swapped in accordance with permutations $g \in \mathscr{G}$. Equivalently, we can write $T(Y, D) \overset{d}{=} T(gY, D)$.

C.2 Conditional Exchangeability and Independence under the Randomization Hypothesis

An idealized randomization generates 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 (see Section 2 of the main text). 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 (by the economist) 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. In this notation, treatment assignments D can be written as

$$D = \boldsymbol{M}(R, X, V),$$

where $\boldsymbol{M}$ is a deterministic vector-valued function.

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

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

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.[3] $\mathbf{1}[\cdot]$ is an indicator function.

In Section 4.2, we assume 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:

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

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

Theorem C.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, $gX = X \ \forall \ g \in \mathscr{G}_X$.

[3] This is an element-wise product.

By assumption, $D = \boldsymbol{M}(R, X)$, so $\forall\ g \in \mathscr{G}_X$,

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

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

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

□

Conditional Independence Another consequence of the randomization protocol $\boldsymbol{M}$ is independence between D and (Y_0, Y_1), conditional on X. This follows from the observation that R is independent of (Y_0, Y_1) by construction. The following theorem proves the conditional independence $(Y_0, Y_1) \perp\!\!\!\perp R \mid X$, assuming that D is generated by (R, X) via $\boldsymbol{M}$ and that X is observed:

Theorem C.2. *Assuming that $D = \boldsymbol{M}(X, R)$, $(Y_1, Y_0) \perp\!\!\!\perp D \mid X$.*

Proof. We have

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

□

This result justifies the following assumption:

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

The assumption justifies matching as a method to correct for compromises in the randomization protocol.

Defining the Hypothesis of No Treatment Effect The null hypothesis of no treatment effect states that the distribution of treatment outcomes Y_1 and control outcomes Y_0 is equivalent: $Y_1 \stackrel{d}{=} Y_0$. Likewise,

in non-compromised experiments, treatment assignments D are independent of outcomes: $(Y_1, Y_0) \perp\!\!\!\perp D$. As noted in Section 4.2, these two statements imply unconditional independence between observed outcomes Y and treatment assignments: $Y \perp\!\!\!\perp D$.[4]

However, compromised randomization precludes the use of this statement of the null hypothesis of unconditional independence ($Y \perp\!\!\!\perp D$) for treatment effect inference. To understand why, first recall that compromised randomization means that treatment assignments D are not independent of covariates X. Now, suppose that these X impact outcomes. In this case, a relationship between Y and D may be induced via X regardless of whether any real treatment effect exists. Such an induced dependence between Y and D would invalidate unconditional independence, even under the null hypothesis of no treatment effect, and would render this representation of the null hypothesis unsuitable as a basis for testing.

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:

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

As stated in Section 4.2, Assumption **A-1** and Hypothesis **H-1** together imply that $Y \perp\!\!\!\perp D \mid X$, which is the hypothesis of no treatment effect that we seek to test.

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 C.3. *Suppose that the randomization is as described in Theorem C.1. Under Hypothesis **H-1**, the joint distribution of outcomes Y and treatment assignments D is 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 C.1, $D \overset{d}{=} gD$ conditional on X. Moreover, Theorem C.2 shows that $(Y_1, Y_0) \perp\!\!\!\perp D \mid X$. Thus, for

[4] The proof is omitted for reasons of brevity, although the proof of a similar fact can be found in Section 4.2.

all $g \in \mathscr{G}_X$ we can write

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

□

Appendix C.5 provides detailed information on how to use Theorem C.3 to design a testing procedure. 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 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 in this setting.

C.3 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 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 $\mathscr{G}_X$ can be written as

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

This definition says 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

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

Alternatively, the permutation g' defined by

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

permutes the index across partition sets and thus it does *not* satisfy the conditions required for inclusion in $\mathscr{G}_X$. 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 $\mathscr{G}_X$ 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 matrixes 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 use permutations that 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 (see Appendix F for a discussion).

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.

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 the control groups) using the simulated permutation distribution. An example of a permutation-based p-value is the fraction of the computed permutation distribution that is greater than the statistic computed using the original unpermuted data. We use the mid-p-value described in Appendix C.5. The next section describes the construction of our test statistic in greater detail.

C.4 The Test Statistic

Conditional Inference in Small Samples As the Perry experiment has a sample of size 123, partitioning participants into detailed categories based on the five pre-program variables is impractical. 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 relationship with Y, and the remaining variables $X^{[N]}$, whose relationship with Y is unconstrained. Using this partition, write $X = [X^{[L]}, X^{[N]}]$. The model for outcomes can be written as $Y = \delta X^{[L]} + f(X^{[N]}, \varepsilon)$, where ε is an error term assumed to be independent of $X^{[L]}$ and $X^{[N]}$.

Linearity Define $\tilde{Y} = Y - \delta X^{[L]}$. 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 $(\tilde{Y}, D) \overset{d}{=} (\tilde{Y}, gD);\ g \in \mathscr{G}_{X^{[N]}}$. As a result, 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 due to 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,

$$Y = f(X, D(X), \varepsilon) = \delta X + \Delta D + \varepsilon,$$

where ε is independent of X. As previously noted, 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 has to be estimated. The Freedman-Lane procedure assumes exchangeability of errors under the null, that is, 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.[5] We account for the non-linear 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

[5] 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 seems unreasonable.

from permutation g as $\widetilde{\varepsilon}_g$ 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 parameter, $(D'Q_X D)^{-1} D'Q_X \tilde{Y}_g$, is the Freedman-Lane coefficient for permutation g.[6] We denote by Δ^j 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 Anderson and Robinson 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 our primary test statistic.

[6] Observe that

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

C.5 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.[7]

Following the notation of Section 4.4, suppose that we have a set of $J+1$ permutations g_j, test statistics Δ^j computed for each permutation, and ranks $T^j = \sum_{l=1}^{J+1} \mathbf{1}[\Delta^j \geqslant \Delta^l]/(J+1)$ for those test statistics.[8] Then 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

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

where $\lceil \alpha(J+1) \rceil$ denotes the largest integer less than or equal to $\alpha(J+1)$. Let the ordered values of $T^j;\ j = 1, \ldots, J+1$, be represented by $T^{(1)}, \ldots, T^{(J+1)}$. 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 is greater 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 if $p \in (\alpha_1, 1]$, and we reject with probability $\frac{\alpha - \alpha_0}{\alpha_1 - \alpha_0}$, if $p \in [\alpha_0, \alpha_1]$. We reject the null hypothesis with probability τ, where τ is given by

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

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

[7] 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).

[8] 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.

Theorem C.4. *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 $\Pr\{\text{reject } H_0 \mid X\} = \alpha$ whenever H_0 is true.*

Proof. We have

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

□

The cardinality of the set $\mathscr{G}$ can be so large that computing p-values over all elements becomes infeasible. In this case, we employ a test that uses random samples of J permutations $g \in \mathscr{G}$ plus the identity permutation as the $J+1$ draw.[9] 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.

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

D Multiple-Hypothesis Testing with Stepdown[10]

D.1 Introduction

In multiple-hypothesis testing, there are two generalized Type-I errors: the *familywise error rate* (FWER), which is the probability of rejecting any true null hypothesis, and 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, Romano, and Shaffer, 2005).[11] 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 large number of hypotheses, such as tens or hundreds of hypotheses, a common occurrence in fields such as genomics.

D.2 Overview of Multiple-Hypothesis Testing

Two traditional but conservative methods for multiple-hypothesis testing are the Bonferroni and the 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 $\leq \alpha$, the traditional procedures use the following rejection rules:

Bonferroni Rejection Rule:

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

Holm Rejection Rule:

(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.

[10] The structure and examples in this appendix are developed by Romano and Wolf (2005). Readers are advised to consult this primary source.

[11] 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), and Benjamini, Krieger, and Yekutieli (2006).

Modern work is based on the procedure of "closure methods."[12] 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 by 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

$$\alpha_{\mathcal{K}'} \equiv \Pr(T_{\mathcal{K}'} > c_{\mathcal{K}'}(\alpha_{\mathcal{K}'})),$$

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 hypotheses 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 insures 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.

[12] See Lehmann and Romano (2005).

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 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, Peritz, and Gabriel (1976).

D.3 Subset Pivotality and 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 sub-vector $p_{\mathcal{L}} = (p_l \; ; \; l \in \mathcal{L})$; for an $\mathcal{L} \subset \mathcal{K}$ 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 sub-vector of p-values is unaffected by the truth or falsehood of hypotheses corresponding to the p-values that are not included in the sub-vector.

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) 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, with all others 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 on 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.

D.3.1 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 can be 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 as follows:

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 $k = K$ being 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 less than $\hat{p}_k$.

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 follows:

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 less than $\hat{p}_k$.

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.

D.4 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 contain 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.

D.5 The Stepdown Algorithm

Data and Hypotheses Assume that we start with outcomes $Y^k;\ k \in \mathcal{K} \equiv \{1, \ldots, K\}$, which have the 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,

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

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 Hypothesis The stepdown method starts by testing the full set of joint null hypotheses $H_{\mathcal{K}}$. For notational 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

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

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

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

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

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 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 largest integer less 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 to compute 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 its full enumeration.

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

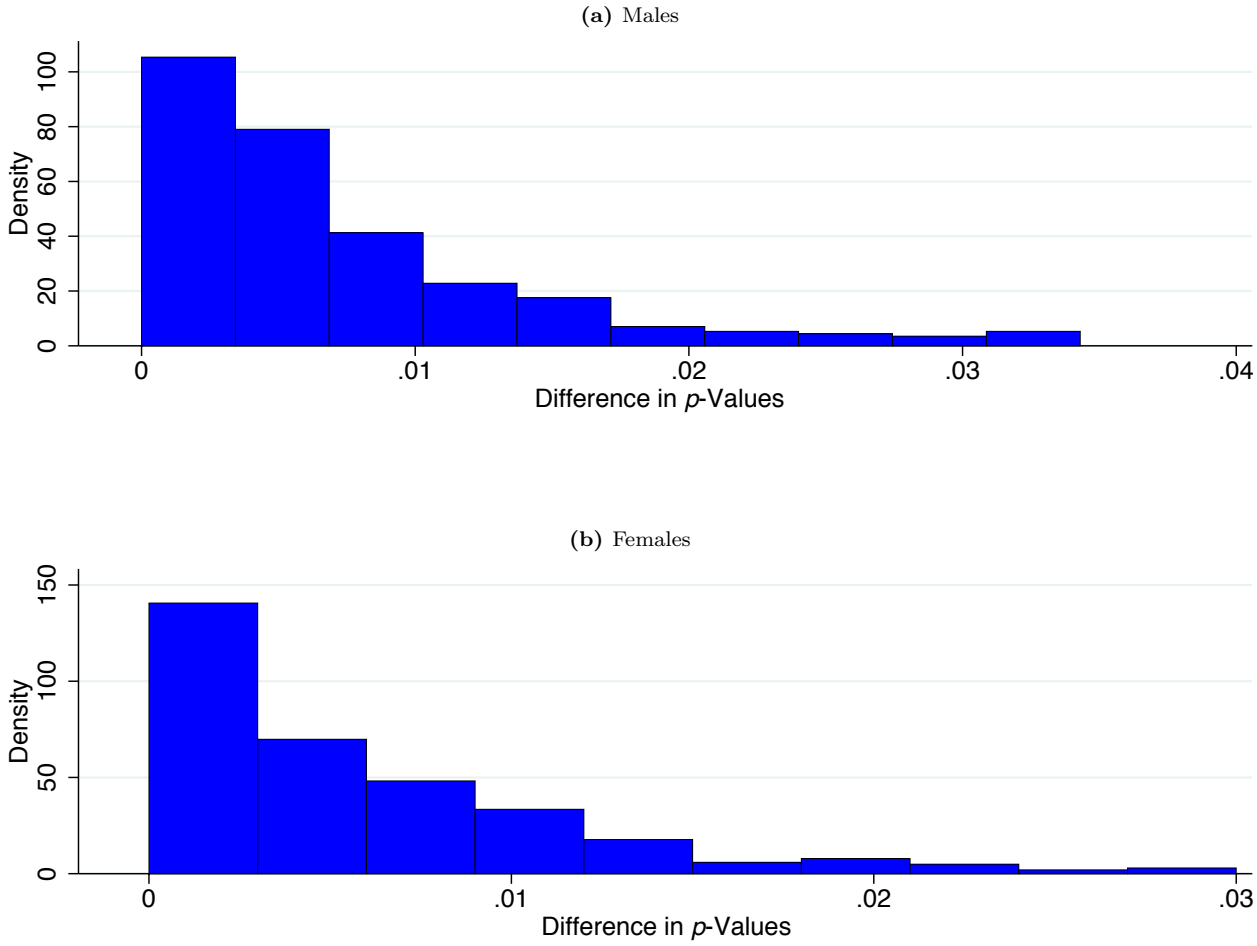
$$
\begin{array}{cl}
[s = 1] & \text{If } T_{\mathcal{K}_1} \leq c_{\mathcal{K}_1}(\alpha), \text{ accept all } H_k,\ k \in \mathcal{K}_1 \text{ and stop.} \\
 & \text{Otherwise, let } \mathcal{K}_2 = \mathcal{K}_1 \setminus \{k^*\},\ k^* = \arg\max(T_k; k \in \mathcal{K}_1). \\
\vdots & \\
[1 < s <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

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.

E Asymptotic and Permutation p-Values

The Perry study has about 60 observations per outcome for each gender. Our analysis confirms a well known fact about the validity of asymptotic statistics in samples of this size (Good, 2000). Traditional resampling techniques, such as the bootstrap and unrestricted permutation, produce distributions of some common sufficient statistics which are very close to their asymptotic versions. As an example, we compute the asymptotic p-value for the t-statistics of the difference in means between treatment groups. We also compute two comparison p-values: the p-value based on an unrestricted permutation method and another based on the usual bootstrap procedure. All three p-values are computed for 350 Perry outcomes chosen for their reliability and relevance to the topic of study. There is little difference between asymptotic p-values and the values based on resampling. Indeed, in 50% of the outcomes, the absolute difference between the asymptotic p-values and resampling values was less than 0.5 percentage point; in 95% of the cases, it was less than 4 percentage points. Figure E.1 shows the histograms of the outcomes with respect to the absolute difference between the asymptotic and permutation p-values.

Figure E.1: Difference between Asymptotic and Permutation p-Values



Notes: 350 observations of differences between one-sided asymptotic and one-sided permutation p-values were used. No blocks or clusters were used while permuting. The 350 observations are p-values for the null hypothesis of no treatment effect based on 350 outcomes of Perry subjects such as wages, test scores, number of arrests, and so forth, chosen by the authors for Perry reanalysis for their reliability and relevance to the outcomes studied in this paper.

F Sensitivity Analysis

The test results reported in this paper rely on Freedman-Lane linear parametric approximations. We can choose either parametric or non-parametric conditioning for each covariate (see Section 4.5 of the paper). In calculating our main results (Tables 3–6), we use non-parametric conditioning on an indicator for whether the socio-economic status (SES) index is above or below the median and use parametric conditioning on the remaining covariates: Stanford-Binet IQ, mother's employment status, and father's presence in the home, all measured at the time of entry into the study.

The purpose of this appendix is to examine the sensitivity of our estimates to different choices of conditioning variables. We focus on two aspects of our procedure. First, what happens when additional covariates are introduced into the nonparametric conditioning set. Second, since some discretization of the continuous variable SES index is necessary to make possible non-parametric conditioning, we examine the sensitivity of our inferences to alternative plausible discretizations. The results of this analysis are described below. We conclude that our main results are robust to alternative choices of the conditioning variables.

Parametric vs. Non-Parametric Conditioning Columns (1)–(4) of Tables F.1–F.4 show the sensitivity of the p-values derived from the Freedman-Lane procedure to shifting additional covariates from the parametric portion of the model to the non-parametric portion. Column (1) shows partial linearity results comparable to the "Partial Linearity" column of our main results (Tables 3–6), while columns (2), (3), and (4) show the effect of shifting mother's employment status, father's presence, and Stanford-Binet IQ, respectively, from the parametric portion to the non-parametric portion of the regression function. In each case, we condition parametrically on the two remaining covariates.

The p-values are quite comparable across columns. Only rarely does inference vary, depending on choices of conditioning variables. Similarly, p-values for the other outcomes—that is, the outcomes for which no column indicates statistical significance—are comparable across alternative conditioning sets, although in cases of nonsignificance, p-values vary greatly. The differences that arise do not exhibit an obvious pattern. These results support the analysis of the text by indicating that the choice of the conditioning variables does not greatly affect the main results reported in Section 5.

Discretizing Non-Parametric Conditioning Variables Columns (1), (5), and (6) of Tables F.1–F.4 show the sensitivity of our results to using different discretizations of the SES index for non-parametric conditioning. To use non-parametric conditioning for a continuous covariate, that variable must be transformed into a discrete covariate on which the permutation orbits used in testing can be restricted (see

Section 4.4).[13] We examine three possible transformations of the non-parametric conditioning covariate SES index: Column (1)—comparable to the "Partial Linearity" column of our main results (Tables 3–6)—conditions non-parametrically on an indicator for whether SES index is above or below the median, column (5) conditions on terciles, and column (6) conditions on quartiles. In all cases, we continue to condition parametrically on the remaining covariates (mother's employment status, father's presence, and Stanford-Binet IQ, measured at study entry).

As with our comparison of parametric vs. non-parametric conditioning, p-values are comparable across the columns. Inference varies across approaches for only a handful of outcomes. These results further reinforce the conclusion that our choice of conditioning sets does not substantially affect the results reported in Section 5.

[13] Kernel methods would be impractical in samples of the size analyzed in this paper.

Table F.1: Main Outcomes by Conditioning: Females, Part 1

	Restriction / Conditioning[a]	**IQ**[b]	—	—	—	Median	—	—	
		M. Work[c]	—	Yes	—	—	—	—	
		F. Pres.[d]	—	—	Yes	—	—	—	
		SES[e]	Median	Median	Median	Median	Tercile	Quartile	
	Outcome	**Age**	**Freedman-Lane p-Values**[f]						**N**
			(1)	(2)	(3)	(4)	(5)	(6)	
Education	Learning Disabled?	≤19	**.009**	**.026**	**.007**	**.011**	**.019**	**.008**	46
	Mentally Impaired?	≤19	**.004**	**.008**	**.007**	**.004**	**.007**	**.005**	46
	Yrs. in Disciplinary Program	≤19	**.070**	**.082**	**.070**	**.071**	.117	**.066**	46
	Yrs. of Special Services	≤14	**.012**	**.016**	**.012**	**.016**	**.024**	**.016**	51
	HS Graduation	19	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	51
	# Years Held Back	≤19	**.098**	.100	.161	**.088**	.135	**.092**	46
	Highest Grade Completed	19	**.002**	**.005**	**.005**	**.002**	**.005**	**.002**	49
	GPA	19	**.000**	**.000**	**.003**	**.001**	**.000**	**.000**	30
	Vocational Training Certificate	≤40	.107	.100	.134	.100	.146	.115	51
Health	No Health Problems	19	.139	.167	.122	.130	.119	.147	49
	No Doctors for Illness, Past Yr.	19	.543	.551	.465	.455	.535	.510	49
	No Non-Routine Care, Past Yr.	27	.491	.487	.662	.411	.551	.549	44
	No Sick Days in Bed, Past Yr.	27	.523	.571	.545	.458	.498	.537	47
	No Treat. for Illness, Past 5 Yrs.	27	.242	.247	.195	.180	.226	.243	47
	Routine Annual Health Exam	27	.734	.717	.734	.654	.777	.728	47
	No Tobacco Use	27	.292	.322	.319	.325	.402	.249	47
	Infrequent Alcohol Use	27	.364	.231	.372	.329	.425	.365	45
	Alive	40	.193	.240	.225	.146	.249	.184	51
	Has Any Children	≤ 19	.331	.324	.327	.288	.410	.328	48
	# Out-of-Wedlock Births	≤ 40	.401	.474	.386	.392	.317	.433	42
Crime	Any Non-Juv. Arrests	≤27	.128	.161	.155	.111	.165	.109	51
	# Non-Juv. Arrests	≤27	**.003**	**.009**	**.010**	**.002**	**.007**	**.002**	51
	Any Misd. Arrests	≤40	.519	.498	.580	.465	.577	.538	51
	# Misd. Arrests	≤40	**.086**	**.093**	.117	**.066**	.138	**.074**	51
	Any Non-Juv. Arrests	≤40	.519	.498	.580	.465	.577	.538	51
	# Non-Juv. Arrests	≤40	**.052**	**.059**	**.077**	**.039**	**.089**	**.046**	51
	Any Arrests	≤40	.240	.244	.268	.230	.316	.250	51
	# Total Arrests	≤40	**.043**	**.041**	**.059**	**.029**	**.069**	**.038**	51

Notes: Monetary values adjusted to thousands of year-2006 dollars using annual national CPI. (a) "—" indicates parametric conditioning, and all others indicate non-parametric conditioning: "yes" if the covariate is discrete, in which case that direct non-parametric conditioning is possible, and otherwise (e.g. "Tercile") to indicate the levels used in conditioning on a continuous covariate; (b) Stanford-Binet IQ, at study entry; (c) Maternal working status at study entry; (d) Father's presence in the home at study entry; (e) Socio-economic status (SES) index at study entry; (f) one-sided p-values for the significance of the treatment coefficient, computed using the Freedman-Lane procedure, with non-parametric conditioning as indicated at the top. p-values below 0.1 are in bold..

Table F.2: Main Outcomes by Conditioning: Females, Part 2

	Restriction / Conditioning[a]							
		IQ[b]	—	—	—	Median	—	—
		M. Work[c]	—	Yes	—	—	—	—
		F. Pres.[d]	—	—	Yes	—	—	—
		SES[e]	Median	Median	Median	Median	Tercile	Quartile
	Outcome	**Age**	**Freedman-Lane p-Values[f]**					**N**
			(1)	(2)	(3)	(4)	(5)	(6)	
Employment	Current Employment	19	**.033**	**.031**	**.048**	**.015**	**.063**	**.018**	51
	No Job in Past Year	19	**.005**	**.004**	**.005**	**.000**	**.009**	**.002**	51
	Jobless Months in Past 2 Yrs.	19	**.019**	**.022**	**.027**	**.013**	**.050**	**.010**	42
	Current Employment	27	**.042**	**.056**	**.058**	**.050**	**.062**	**.034**	47
	No Job in Past Year	27	**.038**	**.038**	**.044**	**.023**	**.041**	**.039**	48
	Jobless Months in Past 2 Yrs.	27	.170	.152	.173	.130	.197	.156	47
	Current Employment	40	.617	.627	.591	.743	.681	.622	46
	No Job in Past Year	40	**.057**	**.051**	**.052**	**.052**	**.087**	**.060**	47
	Jobless Months in Past 2 Yrs.	40	.533	.488	.504	.571	.601	.524	46
Earnings[g]	Monthly Earn., Current Job	19	.216	.187	.261	.129	.266	.195	51
	Monthly Earn., Current Job	27	.116	.114	.121	**.084**	.121	.120	47
	Yearly Earn., Current Job	27	.282	.266	.244	.242	.282	.303	47
	Monthly Earn., Current Job	40	.269	.239	.180	.291	.328	.269	46
	Yearly Earn., Current Job	40	.228	.224	.180	.240	.273	.219	46
Economic	Car Ownership	27	.148	.167	.149	**.075**	.155	.157	47
	Checking Account	27	.474	.447	.484	.366	.589	.476	47
	Savings Account	27	**.052**	**.051**	**.067**	**.029**	**.075**	**.042**	47
	Car Ownership	40	.264	.255	.255	.350	.342	.261	46
	Checking Account	40	.236	.260	.219	.245	.333	.230	46
	Credit Card	40	.236	.213	.162	.202	.285	.245	46
	Savings Account	40	.522	.479	.502	.516	.556	.518	46
	# Months on Welfare	18–27	.122	.113	.148	.106	.168	.108	47
	$>$ 30 Mos. on Welfare	18–27	**.073**	**.081**	.103	**.052**	**.099**	**.060**	47
	Ever on Welfare	18–27	**.048**	**.051**	.112	**.048**	**.062**	**.047**	47
	Never on Welfare	16–40	.138	.131	.157	.157	.168	.148	51
	Never on Welfare (Self Rep.)	26–40	.670	.633	.644	.709	.716	.647	46

Notes: Monetary values adjusted to thousands of year-2006 dollars using annual national CPI. (a) "—" indicates parametric conditioning, and all others indicate non-parametric conditioning: "yes" if the covariate is discrete, in which case that direct non-parametric conditioning is possible, and otherwise (e.g. "Tercile") to indicate the levels used in conditioning on a continuous covariate; (b) Stanford-Binet IQ, at study entry; (c) Maternal working status at study entry; (d) Father's presence in the home at study entry; (e) Socio-economic status (SES) index at study entry; (f) one-sided p-values for the significance of the treatment coefficient, computed using the Freedman-Lane procedure, with non-parametric conditioning as indicated at the top. p-values below 0.1 are in bold.; (g) Age-19 measures are conditional on at least some earnings during the period specified — observations with zero earnings are omitted in computing means and regressions. p-values below 0.1 are in bold.

Table F.3: Main Outcomes by Conditioning: Males, Part 1

	Restriction / Conditioning[a]		IQ[b]	—	—	—	Median	—	—	
			M. Work[c]	—	Yes	—	—	—	—	
			F. Pres.[d]	—	—	Yes	—	—	—	
			SES[e]	Median	Median	Median	Median	Tercile	Quartile	
	Outcome	Age		Freedman-Lane p-Values[f]						N
				(1)	(2)	(3)	(4)	(5)	(6)	
Education	Learning Disabled?	≤19		.768	.760	.763	.794	.906	.777	66
	Mentally Impaired?	≤19		**.055**	**.058**	**.053**	**.037**	**.053**	**.046**	66
	Yrs. in Disciplinary Program	≤19		.130	.149	.142	.124	.145	.227	66
	Yrs. of Special Services	≤14		.212	.222	.209	.194	.286	.163	72
	HS Graduation	19		.418	.438	.426	.437	.431	.577	72
	# Years Held Back	≤19		.740	.753	.738	.721	.803	.714	66
	Highest Grade Completed	19		.304	.321	.315	.261	.336	.414	72
	GPA	19		.338	.355	.327	.336	.403	.360	47
	Vocational Training Certificate	≤40		.411	.395	.416	.368	.308	.316	72
Health	No Health Problems	19		.863	.859	.870	.813	.814	.830	72
	No Doctors for Illness, Past Yr.	19		.448	.433	.468	.373	.432	.470	72
	No Non-Routine Care, Past Yr.	27		.547	.527	.544	.568	.358	.529	63
	No Sick Days in Bed, Past Yr.	27		.162	.158	.162	.127	.188	.168	70
	No Treat. for Illness, Past 5 Yrs.	27		.381	.388	.379	.422	.411	.276	70
	Routine Annual Health Exam	27		.441	.445	.457	.530	.490	.433	68
	No Tobacco Use	27		.253	.252	.264	.238	.189	.275	70
	Infrequent Alcohol Use	27		**.049**	**.050**	**.054**	**.083**	**.020**	**.043**	66
	Alive	40		.155	.197	.164	.125	.197	.250	72
Crime	Any Fel. Arrests	≤27		.436	.464	.436	.350	.603	.629	72
	# Fel. Arrests	≤27		**.042**	**.046**	**.044**	**.041**	**.093**	**.069**	72
	Any Non-Juv. Arrests	≤27		.291	.307	.296	.210	.412	.431	72
	# Non-Juv. Arrests	≤27		**.015**	**.018**	**.016**	**.013**	**.040**	**.043**	72
	Any Misd. Arrests	≤40		.193	.207	.195	.127	.351	.340	72
	# Misd. Arrests	≤40		**.023**	**.021**	**.026**	**.017**	**.069**	**.063**	72
	Any Fel. Arrests	≤40		**.082**	**.086**	**.082**	**.068**	**.088**	.131	72
	# Fel. Arrests	≤40		**.086**	**.099**	**.089**	**.081**	.152	.150	72
	Any Non-Juv. Arrests	≤40		**.076**	**.088**	**.080**	**.053**	.154	.150	72
	# Non-Juv. Arrests	≤40		**.025**	**.025**	**.028**	**.020**	**.072**	**.069**	72
	Any Arrests	≤40		.125	.140	.120	**.096**	.185	.197	72
	# Total Arrests	≤40		**.035**	**.038**	**.038**	**.032**	**.098**	**.089**	72
	Ever Incarcerated	≤40		.112	.134	.109	.112	.112	.157	72

Notes: Monetary values adjusted to thousands of year-2006 dollars using annual national CPI. (a) "—" indicates parametric conditioning, and all others indicate non-parametric conditioning: "yes" if the covariate is discrete, in which case that direct non-parametric conditioning is possible, and otherwise (e.g. "Tercile") to indicate the levels used in conditioning on a continuous covariate; (b) Stanford-Binet IQ, at study entry; (c) Maternal working status at study entry; (d) Father's presence in the home at study entry; (e) Socio-economic status (SES) index at study entry; (f) one-sided p-values for the significance of the treatment coefficient, computed using the Freedman-Lane procedure, with non-parametric conditioning as indicated at the top. p-values below 0.1 are in bold..

Table F.4: Main Outcomes by Conditioning: Males, Part 2

	Restriction / Conditioning[a]							
	IQ[b]		—	—	—	Median	—	—
	M. Work[c]		—	Yes	—	—	—	—
	F. Pres.[d]		—	—	Yes	—	—	—
	SES[e]		Median	Median	Median	Median	Tercile	Quartile
	Outcome	**Age**	**Freedman-Lane p-Values[f]**					**N**
			(1)	(2)	(3)	(4)	(5)	(6)	
Employment	Current Employment	19	.104	.108	.112	.121	.166	.229	72
	No Job in Past Year	19	.858	.843	.849	.841	.893	.894	72
	Jobless Months in Past 2 Yrs.	19	.779	.767	.769	.763	.800	.794	70
	Current Employment	27	.220	.228	.215	.196	.190	.246	69
	No Job in Past Year	27	.187	.213	.177	.178	.184	.237	72
	Jobless Months in Past 2 Yrs.	27	**.029**	**.038**	**.029**	**.028**	**.037**	**.044**	69
	Current Employment	40	**.010**	**.012**	**.011**	**.009**	**.011**	**.016**	66
	No Job in Past Year	40	**.068**	**.082**	**.073**	**.058**	**.086**	.100	72
	Jobless Months in Past 2 Yrs.	40	**.018**	**.021**	**.019**	**.014**	**.021**	**.036**	66
Earnings[g]	Monthly Earn., Current Job	19	**.088**	**.097**	**.092**	.112	.100	.102	72
	Monthly Earn., Current Job	27	**.011**	**.009**	**.010**	**.007**	**.009**	**.016**	68
	Yearly Earn., Current Job	27	.186	.177	.178	.153	.183	.295	66
	Monthly Earn., Current Job	40	.192	.187	.180	.162	.237	.365	66
	Yearly Earn., Current Job	40	.145	.161	.141	.136	.208	.314	66
Economic	Car Ownership	27	**.059**	**.061**	**.060**	**.048**	**.029**	.104	70
	Checking Account	27	.576	.586	.575	.567	.519	.617	70
	Savings Account	27	.395	.395	.392	.388	.396	.586	70
	Car Ownership	40	**.001**	**.002**	**.001**	**.001**	**.002**	**.004**	66
	Checking Account	40	.486	.496	.473	.429	.504	.559	66
	Credit Card	40	.198	.205	.201	.178	.194	.297	66
	Savings Account	40	**.001**	**.001**	**.001**	**.001**	**.001**	**.002**	66
	# Months on Welfare	18–27	.520	.514	.520	.460	.606	.550	66
	$>$ 30 Mos. on Welfare	18–27	.431	.433	.440	.359	.449	.466	66
	Ever on Welfare	18–27	.595	.596	.598	.554	.710	.682	66
	Never on Welfare	16–40	**.027**	**.035**	**.030**	**.020**	**.068**	**.064**	72
	Never on Welfare (Self Rep.)	26–40	**.052**	**.051**	**.052**	**.047**	**.059**	.116	64

Notes: Monetary values adjusted to thousands of year-2006 dollars using annual national CPI. (a) "—" indicates parametric conditioning, and all others indicate non-parametric conditioning: "yes" if the covariate is discrete, in which case that direct non-parametric conditioning is possible, and otherwise (e.g. "Tercile") to indicate the levels used in conditioning on a continuous covariate; (b) Stanford-Binet IQ, at study entry; (c) Maternal working status at study entry; (d) Father's presence in the home at study entry; (e) Socio-economic status (SES) index at study entry; (f) one-sided p-values for the significance of the treatment coefficient, computed using the Freedman-Lane procedure, with non-parametric conditioning as indicated at the top. p-values below 0.1 are in bold.; (g) Age-19 measures are conditional on at least some earnings during the period specified — observations with zero earnings are omitted in computing means and regressions.

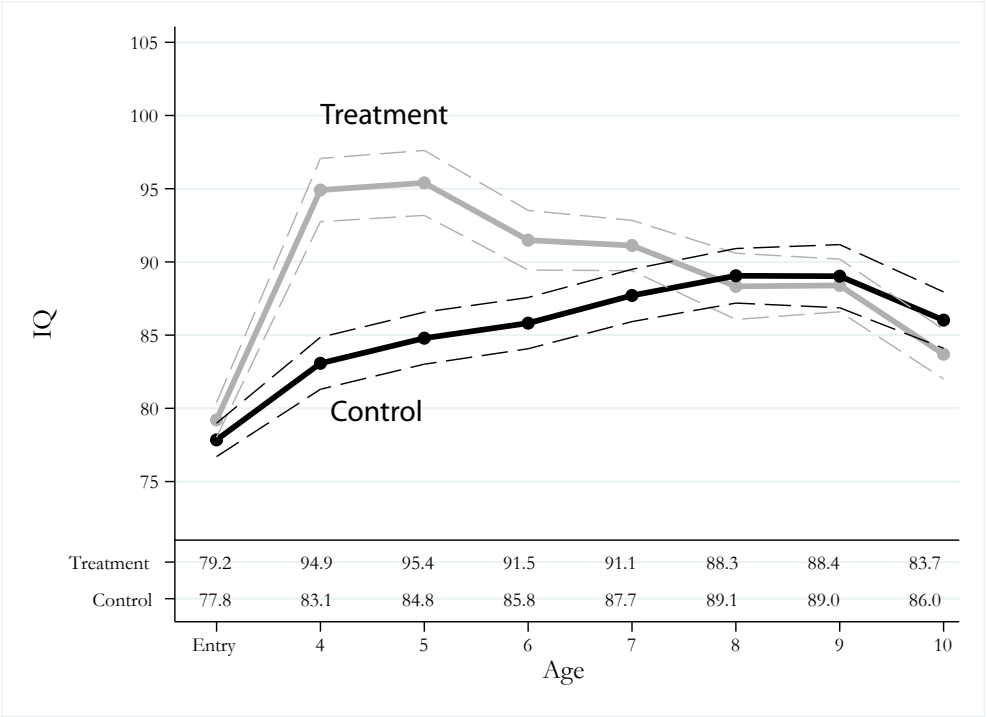
G Analysis of Test Scores

Additional Tables and Figures

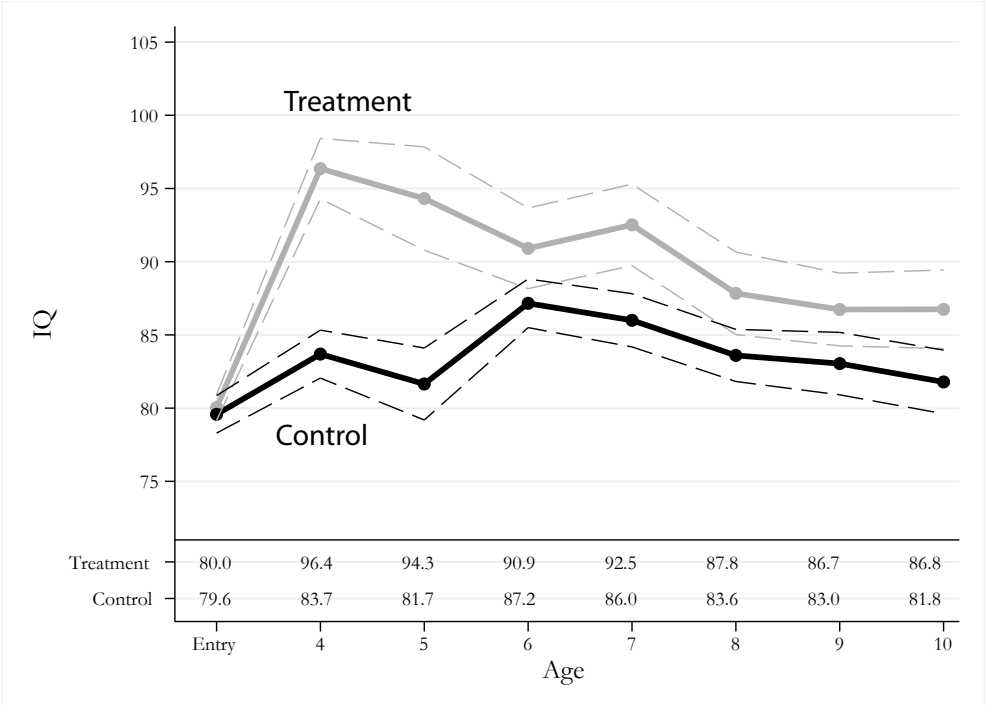
Stanford-Binet IQ Scores During Childhood The graphs in Figure G.1 compare Stanford-Binet IQ scores by gender. IQ effects for the Perry program fade out by age 9. Table G.1 shows that this is especially true for males. For females there is some persistence of the treatment effect, but not in the Stanford-Binet test. (See Table G.1, row labeled "Stanford-Binet".) Yet, strong effects are found for achievement tests for both males and females. Heckman, Malofeeva, Pinto, and Savelyev (2010) analyze this phenomenon in more detail and establish that socioemotional skills were enhanced by the Perry program, driving the boost in test performance. This is consistent with the evidence from Borghans, Golsteyn, Heckman, and Humphries (2010), who report that roughly 50% of the variance in achievement tests is due to variability in noncognitive skills. These results are also consistent with the evidence from Duckworth and Seligman (2005) that higher motivation is predictive of better test scores.

Figure G.1: Perry Subjects' IQ by Gender and Treatment Status

(a) Males



(b) Females



Notes: Data are IQ scores measured using the Stanford-Binet IQ test (1960 revision). The first entry cohort is excluded, as that treatment group received only 1 year of treatment.

Table G.1: Early Cognitive Outcomes by Gender

			Age						
	Measurement	E[a]	3	4	5	6	7	8	9
Males	Stanford-Binet	.191	—	**.000**	**.001**	**.004**	**.049**	.630	.593
	Leiter	—	.103	**.001**	**.009**	.458	.685	.793	.107
	PPVT	—	**.026**	**.001**	**.000**	**.069**	.276	.110	.302
	ITPA	—	.148	—	**.000**	.236	.448	.299	.350
	Joint Test[a]	—	**.073**	**.000**	**.000**	**.014**	.155	.312	.295
Females	Stanford-Binet	.107	—	**.000**	**.002**	**.070**	**.036**	**.039**	.108
	Leiter	—	**.001**	**.001**	**.000**	**.012**	**.035**	**.039**	**.004**
	PPVT	—	**.067**	**.001**	**.001**	**.062**	**.057**	.389	.245
	ITPA	—	**.073**	—	**.000**	**.079**	**.035**	**.063**	**.043**
	Joint Test[a]	—	**.001**	**.000**	**.000**	**.039**	.100	.133	**.015**

Notes: p-Values are for the joint hypothesis consisting of one-sided hypotheses for the significance of treatment effect, corresponding to the first step of a stepdown test on the group of outcomes. Constituent p-values are computed using Mann-Whitney U-statistics, with permutations conditioned on maternal employment and paternal presence, and restricted on SES index and IQ percentiles and maternal employment; siblings were permuted as a block. A complete set of cognitive test scores and detailed tests of California Achievement Test scores can be found in Table G.2. p-values below 0.1 are in bold. (a) For each age, the joint test p-value is the joint-hypothesis test of all available outcomes in the rows above for that gender.

Table G.2: California Achievement Test (CAT) Scores by Gender

	CAT[a] Subscore	Age 7	8	9	10	11	14
Males	Reading	.324	.443	.208	.148	.154	**.086**
	Arithmetic	.207	.114	**.069**	**.082**	.366	**.032**
	Language	.454	.627	**.092**	**.087**	.188	**.013**
	Joint Test	.403	.226	.135	.148	.243	**.032**
Females	Reading	**.024**	**.022**	**.055**	**.042**	.112	**.041**
	Arithmetic	**.020**	**.038**	**.017**	.205	.304	**.063**
	Language	**.085**	**.031**	**.039**	**.078**	.158	**.002**
	Joint Test	**.043**	**.043**	**.030**	**.076**	.180	**.006**

Notes: p-Values are for the joint hypothesis consisting of one-sided hypotheses for the significance of treatment effect, corresponding to the first step of a stepdown test on the group of outcomes. Constituent p-values are computed using Mann-Whitney U-statistics, with permutations conditioned on maternal employment and paternal presence, and restricted on SES index and IQ percentiles and maternal employment; siblings were permuted as a block; each test comprises a single outcome (and hypothesis) in the joint test. p-values below 0.1 are in bold. (a) At ages prior to 14, the CAT hypotheses corresponded to the reading, arithmetic, and language subscores; at age 14, each divided into two further subscores.

H Representativeness of the Perry Sample

Perry Control Group vs. NLSY79 Subsamples Figures H.1–H.5 compare the Perry control group with two comparison groups on selected background characteristics that mimic the Perry eligibility criteria. To extract these comparison groups, we use the National Longitudinal Survey of Youth 1979 (NLSY79), which is a nationally representative longitudinal survey whose respondents represent almost the same birth cohorts as the Perry sample (1956–1964 and 1957–1962, respectively).

The first comparison group is the full black subsample of the NLSY79, while the second is restricted by subject birth order, socio-economic status (SES) index, and Armed Forces Qualification Test (AFQT) score. These restrictions are chosen to mimic the program eligibility criteria of the Perry study.

A practical difficulty in imposing these restrictions on NLSY79 is that we do not have enough information to perfectly mimic the original Perry experiment eligibility criteria. Specifically, we do not know the number of rooms in each NLSY79 respondent's dwelling at age 3, which was used to construct the SES index in the Perry study; neither do we know their IQ scores. Given this lack of information, we construct proxies for these two variables. First, to construct a proxy for the SES index, we first regress the number of rooms in the Perry data set on mother's education, father's occupation, and family size to estimate a linear predictor for the number of rooms. The estimated function is used to predict the number of rooms for each NLSY79 black respondent, which in turn is used to construct a proxy for the SES index. Second, without having IQ scores in the NLSY79, we instead use the AFQT scores as our proxy. While AFQT is an achievement test, not an ability test like the IQ test, it can serve as a proxy for ability as long as achievement and ability are highly correlated. We adjust the AFQT score for age and educational level at the time of testing and use it as our proxy. The method used for adjustment is based on the method of Carneiro, Heckman, and Masterov (2005), which is a simpler version of the method of Hansen, Heckman, and Mullen (2004). This method corrects for reverse causality arising from the effect of education on test scores. The early childhood background characteristics — pre-experimental measures in the Perry sample — that we are comparing in this appendix are parents' average highest grade completed, an SES index, and mother's age at subject's birth, all measured at age 3. Adult outcomes consist of earnings at ages 27 and 40.

Relative to the full black NLSY79 subsample, children in the Perry control group have more disadvantaged family backgrounds. This is not surprising, as the Perry program was targeted toward such children through the aforementioned eligibility. One interesting finding is that this disadvantage is also reflected in adult earnings. Compared to the fully restricted NLSY79 subsample (the final column), however, the relative disadvantage disappears in both childhood and adult outcome measures. These restrictions induce broad comparability between the subsample of the NLSY79 constructed using these principles and the controls in

Table H.1: Comparison of Perry Subjects and the US Black Population: Males at Ages 3, 27, and 40

		Perry Subjects		NLSY79: Restricted Black Subsamples				
		Ctl.	Treat.	All[a]	Younger Sibling[b]	Low-Ability[c]	Low-SES[d]	All Re-rictions[e]
	Sample Size	39	33	706	564	352	290	128
	Pop. Represented			2,222,597	1,749,519	1,085,137	879,363	372,004
Age 3	Parents' Education	9.5 (2.0)	9.3 (2.0)	10.7 (2.6)	10.5 (2.7)	9.9 (2.5)	9.8 (2.3)	9.3 (2.4)
	SES Index	8.6 (1.4)	8.9 (1.7)	10.7 (3.0)	10.6 (3.0)	10.0 (2.6)	8.9 (1.3)	8.6 (1.4)
	Mother's Age at Birth	25.6 (6.6)	26.5 (6.5)	25.1 (6.7)	26.2 (6.5)	25.2 (7.0)	25.6 (7.0)	26.7 (6.9)
Age 27	High School Graduation	0.54 (0.51)	0.48 (0.51)	0.71 (0.45)	0.68 (0.47)	0.59 (0.49)	0.71 (0.45)	0.59 (0.49)
	Employed	0.56 (0.50)	0.60 (0.50)	0.82 (0.38)	0.80 (0.40)	0.77 (0.42)	0.84 (0.37)	0.76 (0.43)
	Yearly Earnings	12,495 (11,354)	14,858 (10,572)	20,239 (18,261)	18,799 (15,850)	16,349 (14,835)	19,268 (16,305)	14,579 (11,819)
Age 40	Employed	0.50 (0.51)	0.70 (0.47)	0.84 (0.37)	0.83 (0.37)	0.76 (0.43)	0.82 (0.38)	0.75 (0.43)
	Yearly Earnings	21,119 (23,970)	27,347 (24,224)	28,729 (26,929)	27,581 (26,059)	19,700 (17,947)	26,992 (25,256)	18,860 (21,256)

Notes: All NLSY79 figures weighted by the initial (1979) sampling weights. Numbers in parentheses are standard deviations. All monetary values in year-2000 dollars. (a) No restrictions; (b) Subjects with at least one elder sibling (all Perry subjects also meet this criterion); (c) AFQT scores below the black median; (d) Socio-economic status (SES) index at most 11; (e) Combines the three restrictions to the left.

the Perry sample. This analysis supports the use of this NLSY79 subsample as a comparison group for the Perry control group.

The U.S. population in 1960 was 180 million, of which 10.6% (19 millions) were black.[14] We use NLSY79, a representative sample of the total population that was born between 1957 and 1964, to estimate the number of persons in the United States that resemble the Perry population at study entry (age 3). According to NLSY79, the black cohort born in 1957–1964 is composed of 2.2 million males and 2.3 million females. Our criteria indicate that 712,000 persons out of this 4.5 million black cohort resemble the Perry population. We estimate that 17% of the male cohort and 15% of the female cohort would be eligible for the Perry program if it were applied nationwide.

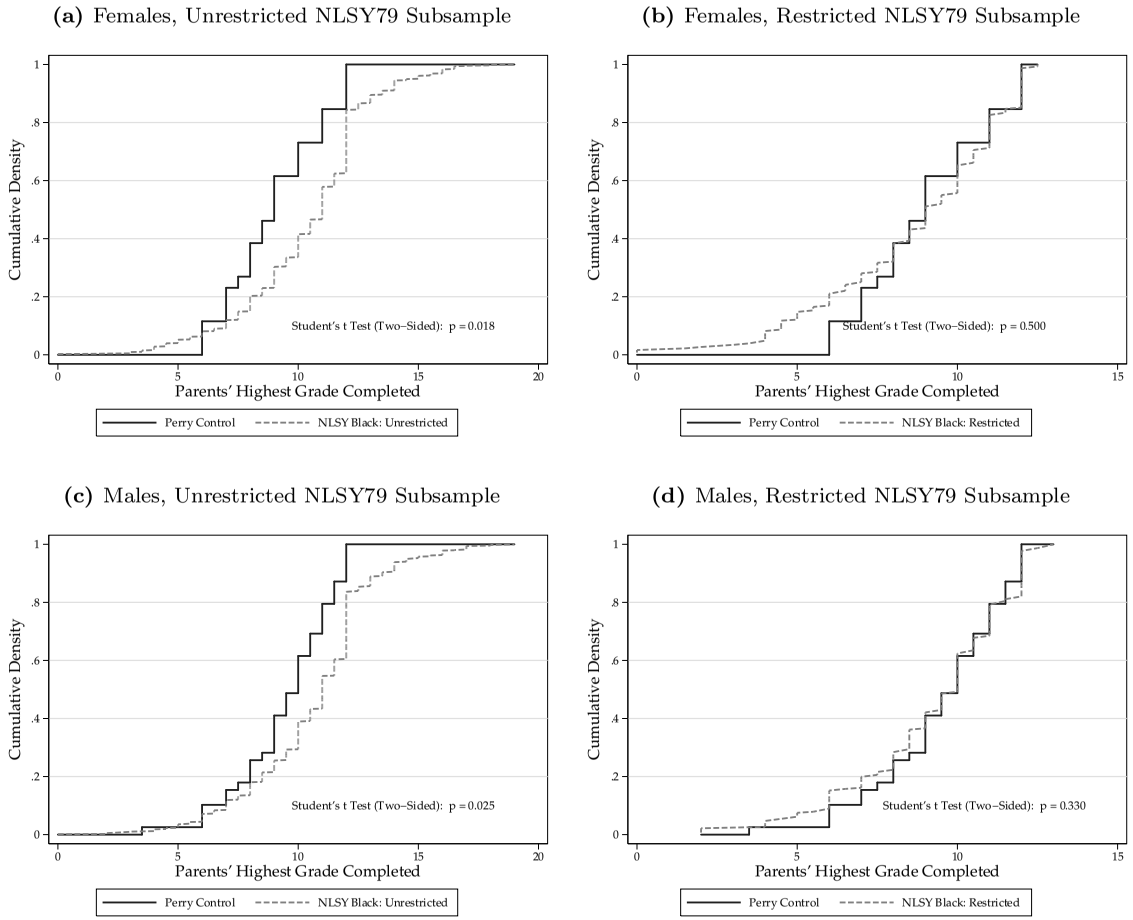
[14] Visit: http://www.census.gov/population/www/documentation/twps0056/twps0056.html for more details.

Table H.2: Comparison of Perry Subjects and the US Black Population: Females at Ages 3, 27, and 40

		Perry Subjects		**NLSY79: Restricted Black Subsamples**				
		Ctl.	Treat.	All[a]	Younger Sibling[b]	Low-Ability[c]	Low-SES[d]	All Re-rictions[e]
	Sample Size	26	25	957	732	434	385	146
	Pop. Represented			2,305,560	1,757,547	1,007,214	902,001	341,721
Age 3	Parents' Education	9.0 *(2.0)*	9.0 *(1.9)*	10.4 *(2.7)*	10.1 *(2.8)*	9.6 *(2.7)*	9.4 *(2.5)*	8.7 *(2.8)*
	SES Index	8.5 *(1.2)*	8.7 *(1.4)*	10.6 *(3.0)*	10.3 *(2.9)*	9.7 *(2.6)*	8.9 *(1.3)*	8.4 *(1.4)*
	Mother's Age at Birth	25.7 *(7.5)*	26.7 *(5.9)*	25.1 *(6.9)*	26.5 *(6.7)*	24.9 *(7.0)*	25.5 *(7.3)*	27.2 *(6.9)*
Age 27	High School Graduation	0.31 *(0.47)*	0.84 *(0.37)*	0.76 *(0.42)*	0.75 *(0.43)*	0.60 *(0.49)*	0.75 *(0.43)*	0.60 *(0.49)*
	Employed	0.55 *(0.51)*	0.80 *(0.41)*	0.65 *(0.48)*	0.62 *(0.48)*	0.50 *(0.50)*	0.60 *(0.49)*	0.45 *(0.50)*
	Yearly Earnings	8,986 *(9,007)*	11,554 *(9,393)*	12,701 *(12,880)*	11,849 *(12,235)*	7,582 *(8,578)*	11,430 *(12,120)*	6,263 *(7,779)*
Age 40	Employed	0.82 *(0.39)*	0.83 *(0.38)*	0.78 *(0.41)*	0.78 *(0.41)*	0.70 *(0.46)*	0.78 *(0.42)*	0.70 *(0.46)*
	Yearly Earnings	17,374 *(16,907)*	20,866 *(20,292)*	20,365 *(18,433)*	19,511 *(17,655)*	12,588 *(11,386)*	19,624 *(18,663)*	11,530 *(10,885)*

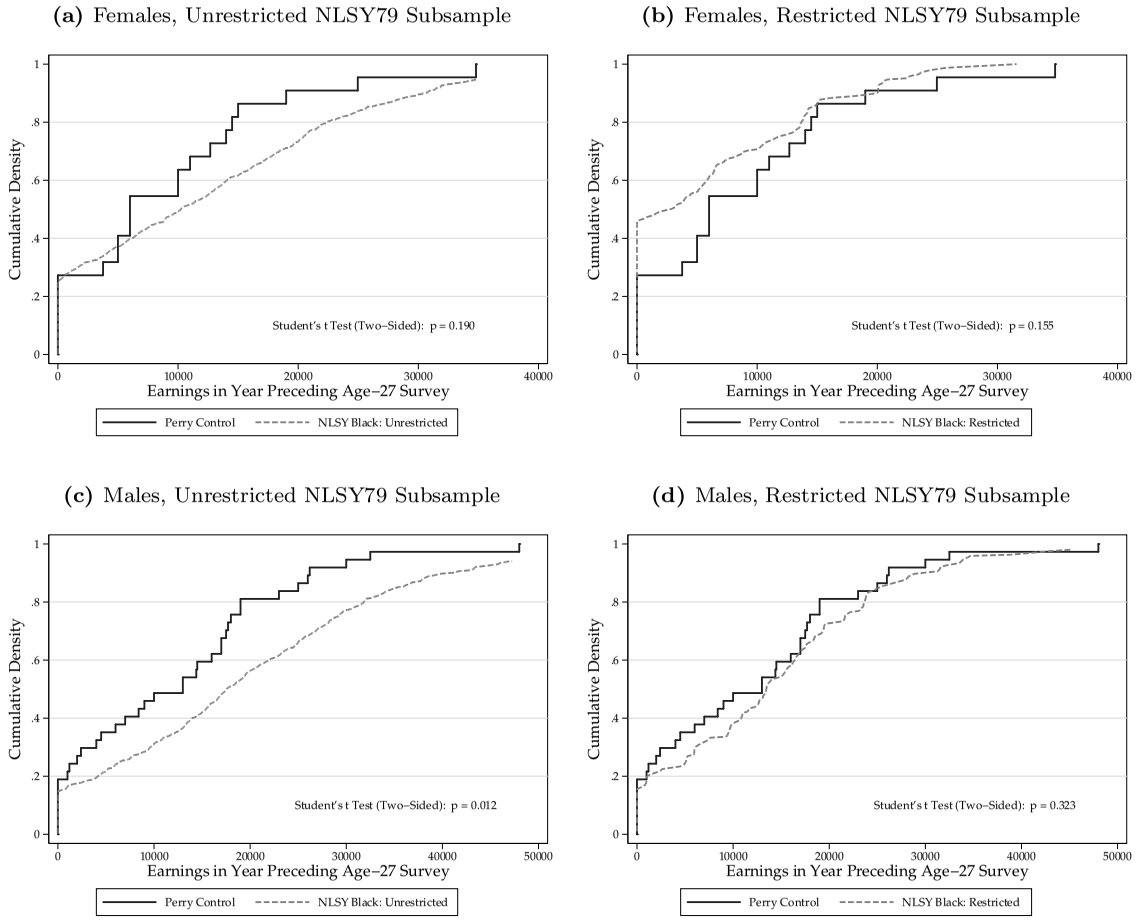
Notes: All NLSY79 figures weighted by the initial (1979) sampling weights. Numbers in parentheses are standard deviations. All monetary values in year-2000 dollars. (a) No restrictions; (b) Subjects with at least one elder sibling (all Perry subjects also meet this criterion); (c) AFQT scores below the black median; (d) Socio-economic status (SES) index at most 11; (e) Combines the three restrictions to the left.

Figure H.1: Perry vs. NLSY79: CDF of Mean Parental Highest Grade Completed



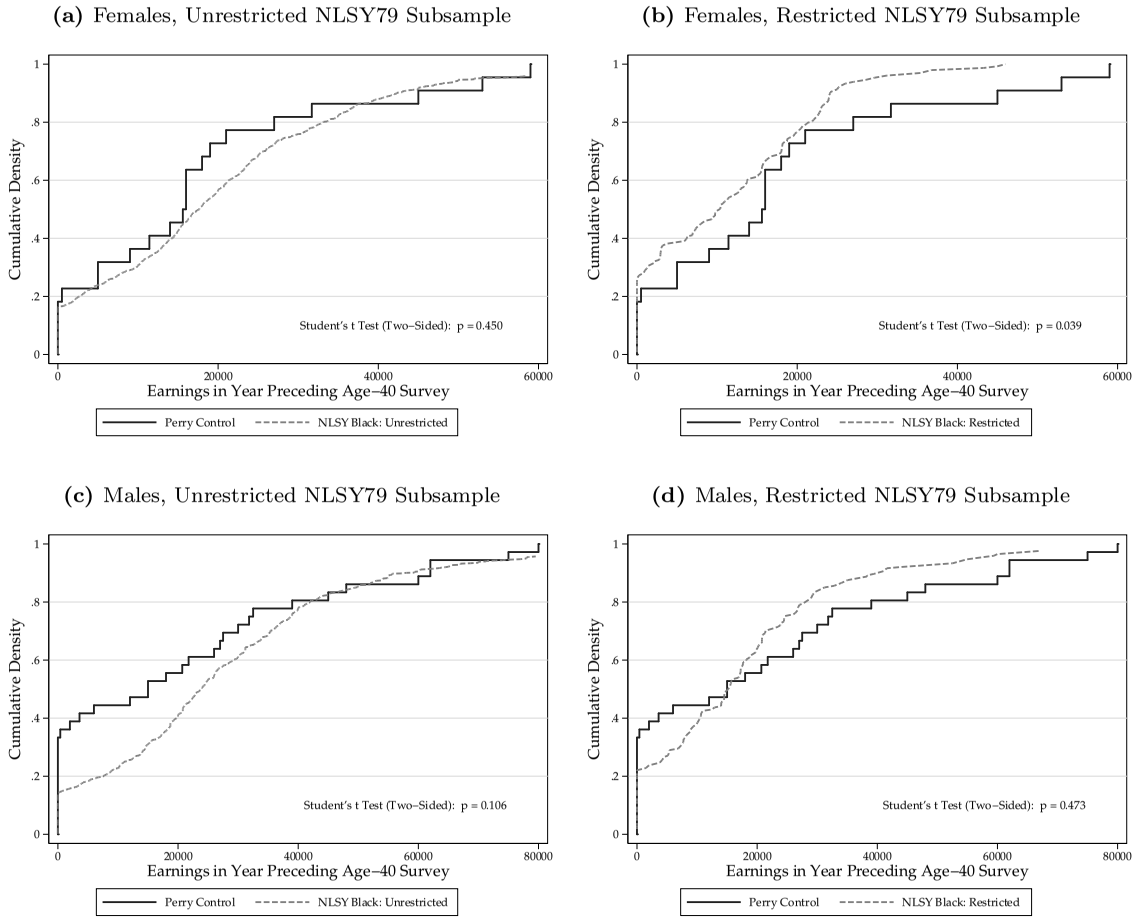
Notes: The full NLSY79 sample is the full black subsample. The restricted NLSY79 sample is the black subsample limited to those with: at least one elder sibling, socio-economic status (SES) index no greater than 11, and 1979 AFQT score less than the black median. SES is a weighted linear combination of average parental highest grade completed, working parent's employment status (father if present), and ratio of rooms to people in household. The t statistic is for the difference in means between the two distributions.

Figure H.2: Perry vs. NLSY79: CDF of Age-27 Earnings



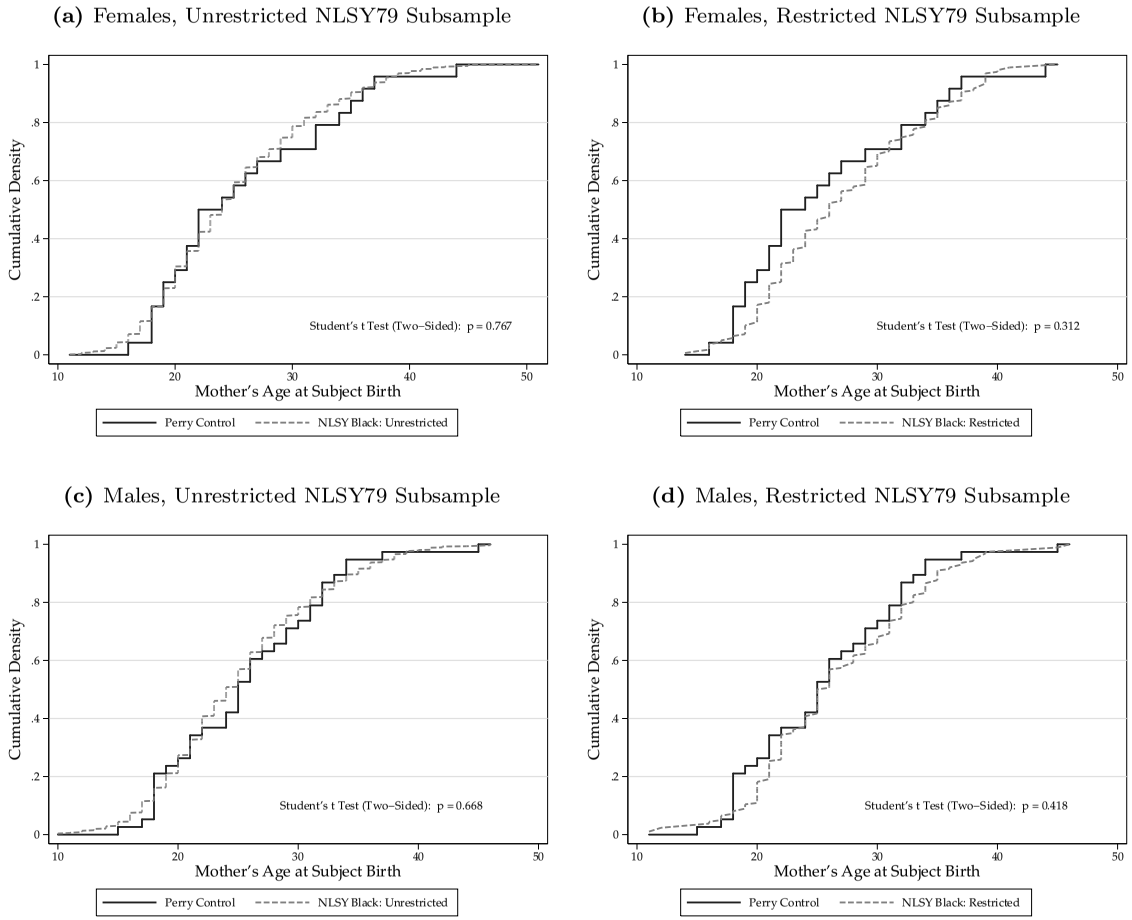
Notes: The full NLSY79 sample is the full black subsample. The restricted NLSY79 sample is the black subsample limited to those with: at least one elder sibling, socio-economic status (SES) index no greater than 11, and 1979 AFQT score less than the black median. SES is a weighted linear combination of average parental highest grade completed, working parent's employment status (father if present), and ratio of rooms to people in household. The t statistic is for the difference in means between the two distributions. Earnings discounted to year-2000 dollars.

Figure H.3: Perry vs. NLSY79: CDF of Age-40 Earnings



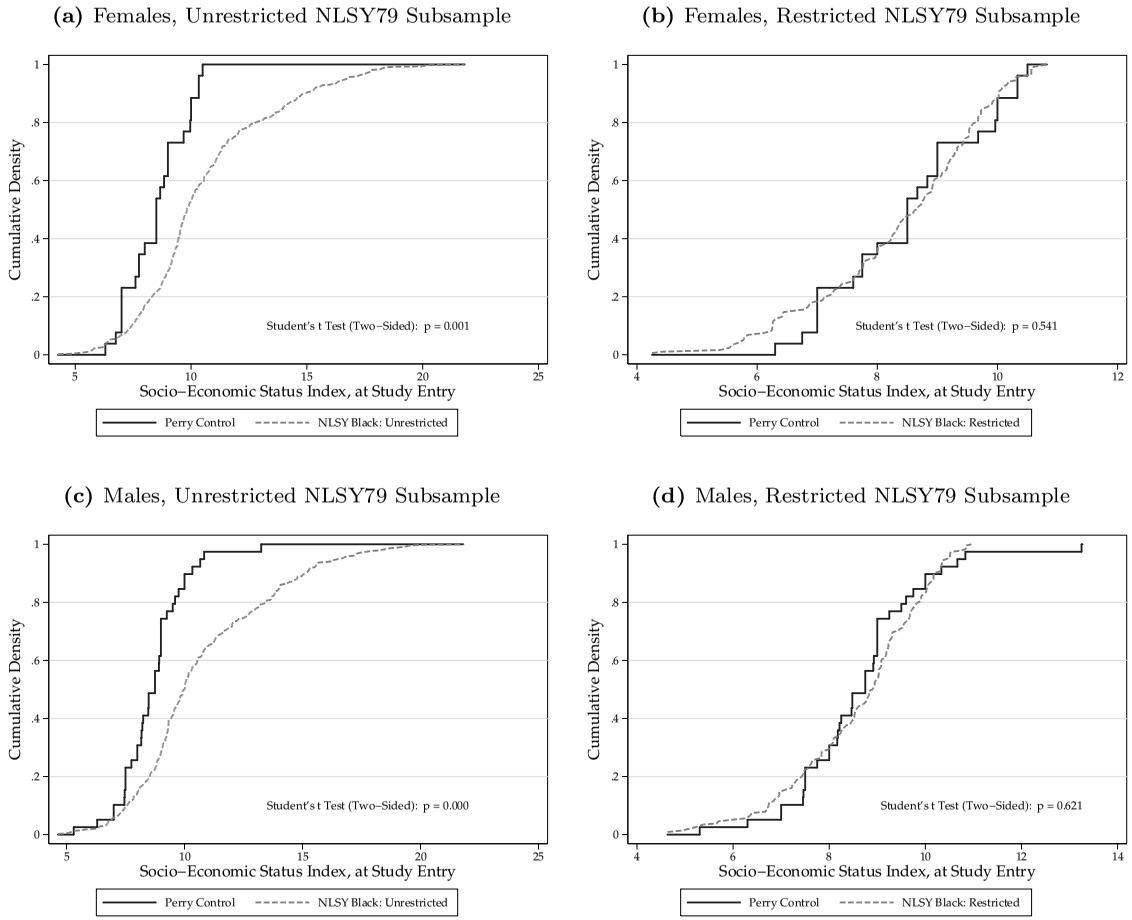
Notes: The full NLSY79 sample is the full black subsample. The restricted NLSY79 sample is the black subsample limited to those with: at least one elder sibling, socio-economic status (SES) index no greater than 11, and 1979 AFQT score less than the black median. SES is a weighted linear combination of average parental highest grade completed, working parent's employment status (father if present), and ratio of rooms to people in household. The t statistic is for the difference in means between the two distributions. Earnings discounted to year-2000 dollars.

Figure H.4: Perry vs. NLSY79: CDF of Mother's Age at Subject Birth



Notes: The full NLSY79 sample is the full black subsample. The restricted NLSY79 sample is the black subsample limited to those with: at least one elder sibling, socio-economic status (SES) index no greater than 11, and 1979 AFQT score less than the black median. SES is a weighted linear combination of average parental highest grade completed, working parent's employment status (father if present), and ratio of rooms to people in household. The t statistic is for the difference in means between the two distributions.

Figure H.5: Perry vs. NLSY79: CDF of Socio-Economic Status Index



Notes: The full NLSY79 sample is the full black subsample. The restricted NLSY79 sample is the black subsample limited to those with: at least one elder sibling, socio-economic status (SES) index no greater than 11, and 1979 AFQT score less than the black median. SES is a weighted linear combination of average parental highest grade completed, working parent's employment status (father if present), and ratio of rooms to people in household. The t statistic is for the difference in means between the two distributions.

I The Role of the Local Economy in Explaining Gender Differences in Treatment Outcomes

The local economic history of Washtenaw County[15] has peculiarities that may explain the age pattern of male treatment effects, and thus explain gender differences in a number of program outcomes. In the 1970s, employment in Ypsilanti and Washtenaw increased by 50% — a much higher rate than for the state (14%) or the country (25%) as a whole (see Table I.1). This rapid growth coincided with a boom in the local manufacturing sector, which subsequently contracted during later decades, although the service sector continued to expand (see Figure I.1). The boom was particularly prevalent in the male-friendly manufacturing sector.[16,17] This economic boom created plentiful jobs during subjects' late teens, increasing the opportunity cost of attending school and resulting in a higher dropout rate for boys. In later decades, as the manufacturing sector shrank, it became more difficult for males to find jobs, while sectors in which females were mostly employed (such as the service sector) expanded.

These labor market dynamics may partially explain the lack of a positive male program treatment effect for high school graduation. Further, the exceptionally rapid employment growth in the Ypsilanti area suggests the possibility that regional economic shocks drive program treatment effects. Therefore, we do not observe a significant treatment effect on male employment at age 19 or for male educational attainment, since at the time Perry participants entered the labor market, manufacturing jobs did not require a high school degree.

While it is not easy to verify this interpretation with any precision, it is consistent with observed patterns of migration out of economically troubled Michigan. At age 27, treatment males were more likely to migrate than their control counterparts, although the difference is not statistically significant at conventional levels (see Table I.2). This evidence is consistent with a positive effect of Perry on the skills of participants. Many studies of migration show a positive link between education and migration (Sjaastad, 1962; Vigdor, 2002a,b). The observed differences in migration between treatments and controls support the interpretation that treatment had some positive effect on skills and motivation, even if we do not observe this directly in terms of its effect on educational attainment of males. This pattern is also consistent with the pattern that males had strong treatment effects on earnings outcomes despite insignificant treatment effects on education, as well as the finding that treatment males had greater noncognitive skills and better achievement test scores than their control counterparts. (See Heckman, Malofeeva, Pinto, and Savelyev, 2010.)

[15] Washtenaw County, which contains Ypsilanti and Ann Arbor, is located in the Detroit metropolitan area.

[16] Goldin and Katz, 2008, discuss the positive relationship between the demand for labor in the manufacturing sector and the high school dropout rate. Manufacturing jobs did not require skilled workers (high school graduates).

[17] At age 19, 12 out of 31 working males reported their jobs as assembly or auto mechanic, while 8 out of 15 working females reported their jobs as cashier, food service, or dishwasher.

Figure I.1: Michigan Employment, by Industry

Source: Southeast Michigan Council of Governments (2002).

Table I.1: Historical Employment Trends in Ypsilanti, Michigan

Year	Ypsilanti		Washtenaw		Michigan		U.S. Total	
	Emp.	Δ %	Emp.	Δ %	Emp.	Δ %	Emp.	Δ %
1970	12,634	-	105,058	-	3,558,467	-	91,281,600	-
1980	19,441	54	164,723	57	4,039,438	14	114,231,200	25
1990	19,773	2	213,928	30	4,826,388	19	139,426,900	22
2000	17,716	-10	232,175	9	5,654,522	17	167,465,300	20

Source: Southeast Michigan Council of Governments (2002).

Table I.2: Migration, by Gender

% Out of Michigan	**Males**			**Females**		
	Ctl.	Trt.	p[a]	Ctl.	Trt.	p[a]
at age 27[b]	12.8	21.2	.174	26.9	8.0	.040
at age 40[c]	25.0	26.7	.440	13.6	4.2	.132
N	39	33		26	25	

Notes: (a) p-values are for asymptotic one-sided tests; (b) At the time of age-27 survey; (c) 1996–2002.

References

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

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

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

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

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

Begun, J. M. and K. R. Gabriel (1981, June). Closure of the Newman-Keuls multiple comparisons procedure. *Journal of the American Statistical Association 76*(374), 241–245.

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

Borghans, L., B. H. H. Golsteyn, J. J. Heckman, and J. E. Humphries (2010). IQ, achievement, and personality. Unpublished manuscript, University of Maastricht and University of Chicago (revised from the 2009 version).

Carneiro, P., J. J. Heckman, and D. V. Masterov (2005, April). Labor market discrimination and racial differences in pre-market factors. *Journal of Law and Economics 48*(1), 1–39.

Duckworth, A. L. and M. E. P. Seligman (2005, November). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science 16*(12), 939–944.

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

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

Goldin, C. and L. F. Katz (2008). *The Race between Education and Technology*. Cambridge, MA: Belknap Press of Harvard University Press.

Good, P. I. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses* (2 ed.). Series in Statistics. New York: Springer Verlag.

Hansen, K. T., J. J. Heckman, and K. J. Mullen (2004, July–August). The effect of schooling and ability on achievement test scores. *Journal of Econometrics 121*(1-2), 39–98.

Heckman, J. J., L. Malofeeva, R. Pinto, and P. A. Savelyev (2010). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. Unpublished manuscript, University of Chicago, Department of Economics.

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

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

Lehmann, E. L., J. P. Romano, and J. P. Shaffer (2005). On optimality of stepdown and stepup multiple test procedures. *Annals of Statistics 33*(3), 10841108.

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

Newman, D. (1939, July). 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. and A. M. Shaikh (2004). On control of the false discovery proportion. Technical Report 2004-31, Department of Statistics, Stanford University.

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

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.

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

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

Sjaastad, L. A. (1962). The costs and returns of human migration. *Journal of Political Economy 70*(5, Part 2: Investment in Human Beings), 80–93.

Southeast Michigan Council of Governments (2002). *Historical Population and Employment by Minor Civil Division, Southeast Michigan*. Detroit, MI: Southeast Michigan Council of Governments.

Vigdor, J. L. (2002a, November). Locations, outcomes, and selective migration. *The Review of Economics and Statistics 84*(4), 751–755.

Vigdor, J. L. (2002b, May). The pursuit of opportunity: Explaining selective black migration. *Journal of Urban Economics 51*(3), 391–417.

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