NBER WORKING PAPER SERIES

ANALYZING SOCIAL EXPERIMENTS AS IMPLEMENTED: A REEXAMINATION OF THE EVIDENCE FROM THE HIGHSCOPE PERRY PRESCHOOL PROGRAM

James J. Heckman
Seong Hyeok Moon
Rodrigo Pinto
Peter A. Savelyev
Adam Yavitz

Working Paper 16238
http://www.nber.org/papers/w16238

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, or of the National Bureau of Economic Research. Supplementary materials for this paper may be found at http://jenni.uchicago.edu/Perry/

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2010

Analyzing Social Experiments as Implemented: A Reexamination of the Evidence From the HighScope Perry Preschool Program

James J. Heckman, Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz

NBER Working Paper No. 16238
July 2010
JEL No. C93,I21,J15

ABSTRACT

Social experiments are powerful sources of information about the effectiveness of interventions. In practice, initial randomization plans are almost always compromised. Multiple hypotheses are frequently tested. "Significant" effects are often reported with p-values that do not account for preliminary screening from a large candidate pool of possible effects. This paper develops tools for analyzing data from experiments as they are actually implemented.

We apply these tools to analyze the influential HighScope Perry Preschool Program. The Perry program was a social experiment that provided preschool education and home visits to disadvantaged children during their preschool years. It was evaluated by the method of random assignment. Both treatments and controls have been followed from age 3 through age 40.

Previous analyses of the Perry data assume that the planned randomization protocol was implemented. In fact, as in many social experiments, the intended randomization protocol was compromised. Accounting for compromised randomization, multiple-hypothesis testing, and small sample sizes, we find statistically significant and economically important program effects for both males and females. We also examine the representativeness of the Perry study.

James J. Heckman
Department of Economics
The University of Chicago
1126 E. 59th Street
Chicago, IL 60637
and University College Dublin and IZA
and also NBER
jjh@uchicago.edu

Seong Hyeok Moon
Department of Economics
The University of Chicago
1126 E. 59th Street
Chicago, IL 60637
moon@uchicago.edu

Rodrigo Pinto
Department of Economics
The University of Chicago
1126 E. 59th Street
Chicago, IL 60637
rodrig@uchicago.edu

Peter A. Savelyev
Department of Economics
1126 E. 59th Street
Chicago, IL 60637
psavel@uchicago.edu

Adam Yavitz
Department of Economics
1126 E. 59th Street
Chicago, IL 60637
adamy@uchicago.edu

An online appendix is available at:
http://www.nber.org/data-appendix/w16238

1 Introduction

Social experiments can produce valuable information about the effectiveness of interventions. However, many social experiments are compromised by departures from initial randomization plans.[1] Many have small sample sizes. Applications of large sample statistical procedures may produce misleading inferences. In addition, most social experiments have multiple outcomes. This creates the danger of selective reporting of "significant" effects from a large pool of possible effects, biasing downward reported p-values. This paper develops tools for analyzing the evidence from experiments with multiple outcomes as they are implemented rather than as they are planned. We apply these tools to reanalyze an influential social experiment.

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

Many analysts discount the reliability of the Perry study. For example, Hanushek and Lindseth (2009), among others, claim that the sample size of the study is too small to make valid inferences about the program. Herrnstein and Murray (1994) claim that estimated effects of the program are small and that many are not statistically significant. Others express the concern that previous analyses selectively report statistically significant estimates, biasing the inference about the program (Anderson, 2008).

There is a potentially more devastating critique. As happens in many social experiments, the proposed randomization protocol for the Perry study was compromised. This compromise casts doubt on the validity of evaluation methods that do not account for the compromised randomization and calls into question the validity of the simple statistical

[1] See the discussion in Heckman (1992); Hotz (1992); and Heckman, LaLonde, and Smith (1999).

procedures previously applied to analyze the Perry study.[2]

In addition, there is the question of how representative the Perry population is of the general African-American population. Those who advocate access to universal early childhood programs often appeal to the evidence from the Perry study, even though the project only targeted a disadvantaged segment of the population.[3]

This paper develops and applies small-sample permutation procedures that are tailored to test hypotheses on samples generated from the less-than-ideal randomizations conducted in many social experiments. We apply these tools to the data from the Perry experiment. We correct estimated treatment effects for imbalances that arose in implementing the randomization protocol and from post-randomization reassignment. We address the potential problem that arises from arbitrarily selecting "significant" hypotheses from a set of possible hypotheses using recently developed stepdown multiple-hypothesis testing procedures. The procedures we use minimize the probability of falsely rejecting any true null hypotheses.

Using these tools, this paper demonstrates the following points: (a) Statistically significant Perry treatment effects survive analyses that account for the small sample size of the study. (b) Correcting for the effect of selectively reporting statistically significant responses, there are substantial impacts of the program on males and females. Results are stronger for females at younger adult ages and for males at older adult ages. (c) Accounting for the compromised randomization of the program strengthens the evidence for important program effects compared to the evidence reported in the previous literature that neglects the imbalances created by compromised randomization. (d) Perry participants are representative of a low-ability, disadvantaged African-American population.

This paper proceeds as follows. Section 2 describes the Perry experiment. Section 3

[2] This problem is pervasive in the literature. For example, in the Abecedarian program, randomization was also compromised as some initially enrolled in the experiment were later dropped (Campbell and Ramey, 1994). In the SIME-DIME experiment, the randomization protocol was never clearly described. See Kurz and Spiegelman (1972). Heckman, LaLonde, and Smith (1999) chronicle the variety of "threats to validity" encountered in many social experiments.

[3] See, for example, The Pew Center on the States (2009) for one statement about the benefits of universal programs.

discusses the statistical challenges confronted in analyzing the Perry experiment. Section 4 presents our methodology. Our main empirical analysis is presented in Section 5. Section 6 examines the representativeness of the Perry sample. Section 7 compares our analysis to previous analyses of Perry. Section 8 concludes. Supplementary material is placed in the Web Appendix.[4]

2 Perry: Experimental Design and Background

The HighScope Perry Program was conducted during the early- to mid-1960s in the district of the Perry Elementary School, a public school in Ypsilanti, Michigan, a town near Detroit. The sample size was small: 123 children allocated over five entry cohorts. Data were collected at age 3, the entry age, and through annual surveys until age 15, with additional follow-ups conducted at ages 19, 27, and 40. Program attrition remained low through age 40, with over 91% of the original subjects interviewed. Two-thirds of the attrited were dead. The rest were missing.[5] Numerous measures were collected on economic, criminal, and educational outcomes over this span as well as on cognition and personality. Program intensity was low compared to that in many subsequent early childhood development programs.[6] Beginning at age 3, and lasting 2 years, treatment consisted of a 2.5-hour educational preschool on weekdays during the school year, supplemented by weekly home visits by teachers.[7] HighScope's innovative curriculum, developed over the course of the Perry experiment, was based on the principle of active learning, guiding students through the formation of key developmental factors using intensive child-teacher interactions (Schweinhart et al. 1993, pp. 34–36; Weikart et al. 1978, pp. 5–6, 21–23). A more complete description of the Perry program curriculum is given in Web Appendix A.[8]

[4] http://jenni.uchicago.edu/Perry/

[5] There are two missing controls and two missing treatments. Five controls and two treatments are dead.

[6] The Abecedarian program is an example (see, e.g., Campbell et al., 2002). Cunha, Heckman, Lochner, and Masterov (2006) and Reynolds and Temple (2008) discuss a variety of these programs and compare their intensity.

[7] An exception is that the first entry cohort received only 1 year of treatment, beginning at age 4.

[8] The website can be accessed at http://jenni.uchicago.edu/Perry/.

Eligibility Criteria The program admitted five entry cohorts in the early 1960s, drawn from the population surrounding the Perry Elementary school. Candidate families for the study were identified from a survey of the families of the students attending the elementary school, by neighborhood group referrals, and through door-to-door canvassing. The eligibility rules for participation were that the participants should (i) be African-American; (ii) have a low IQ (between 70 and 85) at study entry,[9] and (iii) be disadvantaged as measured by parental employment level, parental education, and housing density (persons per room). The Perry study targeted families who were more disadvantaged than most other African-American families in the United States. but were representative of a large segment of the disadvantaged African-American population. We discuss the issue of the representativeness of the program compared to the general African-American population in Section 6.

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

We do not know whether, among eligible families in the Perry catchment, those who volunteered to participate in the program were more motivated than other families, and whether this greater motivation would have translated into better child outcomes. However, according to Weikart, Bond, and McNeil (1978, p. 16), "virtually all eligible children were enrolled in the project," so this potential concern appears to be unimportant.

Randomization Protocol The randomization protocol used in the Perry study was complex. According to Weikart et al. (1978, p. 16), for each designated eligible entry cohort,

[9] Measured by the Stanford-Binet IQ test (1960s norming). The average IQ in the general population is 100 by construction. IQ range for Perry participants is 1–2 standard deviations below the average.

Table 1: Comparing Families of Participants with Other Families with Children in the Perry Elementary School Catchment and a Nearby School in Ypsilanti, Michigan

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

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

children were assigned to treatment and control groups in the following way, which is graphically illustrated in Figure 1:

1. In any entering cohort, younger siblings of previously enrolled families were assigned the same treatment status as their older siblings.[10]

2. Those remaining were ranked by their entry IQ scores.[11] Odd- and even-ranked subjects were assigned to two separate unlabeled groups.

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

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

4. A flip of a coin (a single toss) labeled one group as "treatment" and the other as "control."

5. Some individuals provisionally assigned to treatment, whose mothers were employed at the time of the assignment, were swapped with control individuals whose mothers were not employed. The rationale for these swaps was that it was difficult for working mothers to participate in home visits assigned to the treatment group and because of transportation difficulties.[12] A total of five children of working mothers initially assigned to treatment were reassigned to control.

[10] The rationale for excluding younger siblings from the randomization process was that enrolling children in the same family in different treatment groups would weaken the observed treatment effect due to within-family spillovers.

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

[12] The following quotation from an early monograph on Perry summarizes the logic of the study planners: "*Occasional exchanges of children between groups also had to be made because of the inconvenience of half-day preschool for working mothers and the transportation difficulties of some families. No funds were available for transportation or full-day care, and special arrangements could not always be made.*" (Weikart, Bond, and McNeil, 1978, p. 17)

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

3 Statistical Challenges in Analyzing the Perry Program

Drawing valid inference from the Perry study requires meeting three statistical challenges: (i) small sample size; (ii) compromise in the randomization protocol; and (iii) the large number of outcomes and associated hypotheses, which creates the danger of selectively reporting "significant" estimates out of a large candidate pool of estimates, thereby biasing downward reported p-values.

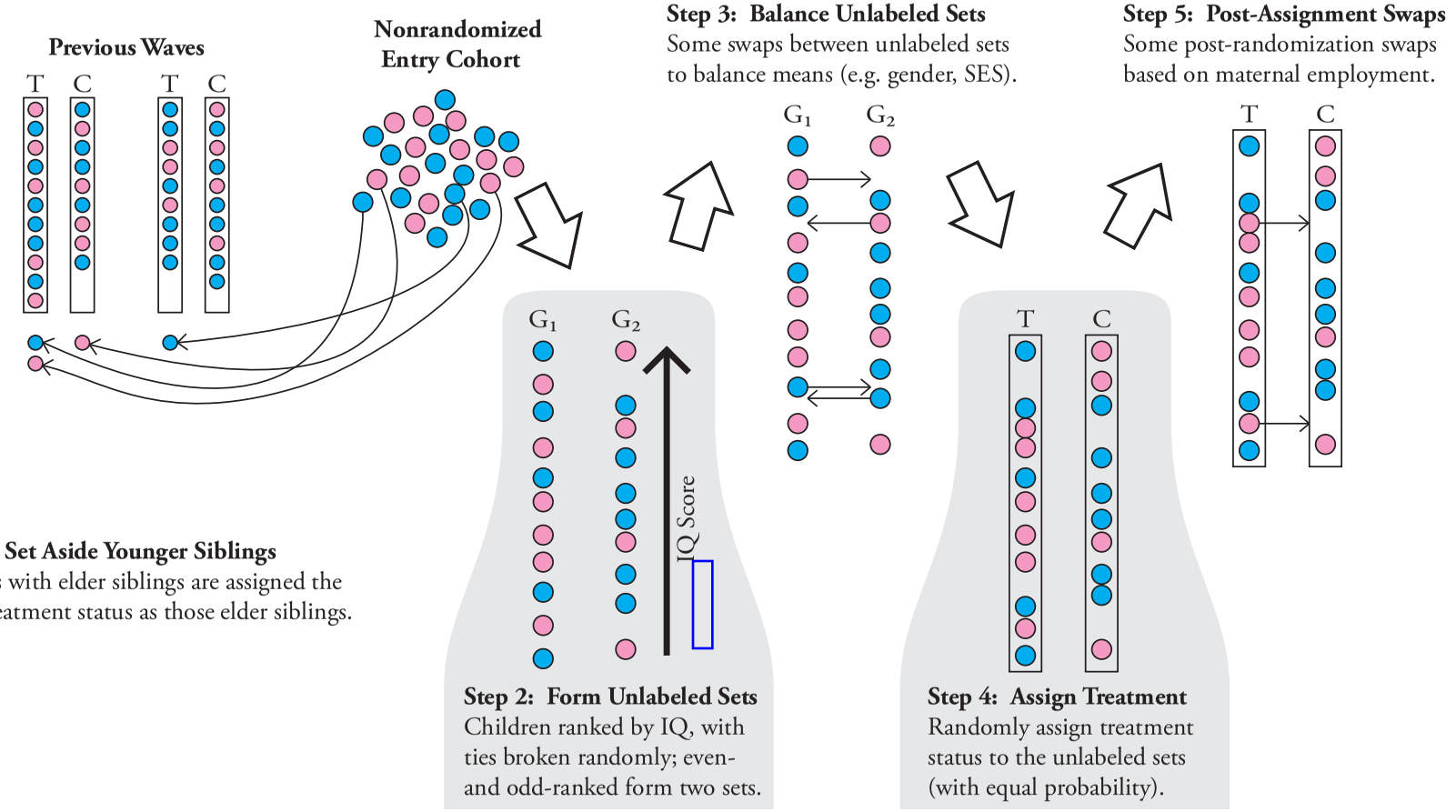
Small Sample Size The small sample size of the Perry study and the non-normality of many outcome measures call into question the validity of classical tests, such as those based on the t-, F-, and χ^2-statistics.[13] Classical statistical tests rely on central limit theorems and produce inferences based on p-values that are only asymptotically valid.

A substantial literature demonstrates that classical testing procedures can be unreliable when sample sizes are small and the data are non-normal.[14] Both features characterize the Perry study. There are approximately 25 observations per gender in each treatment assignment group and the distribution of observed measures is often highly skewed.[15] Our paper addresses the problem of small sample size by using permutation-based inference procedures that are valid in small samples.

The Treatment Assignment Protocol The randomization protocol implemented in the Perry study diverged from the original design. Treatment and control statuses were

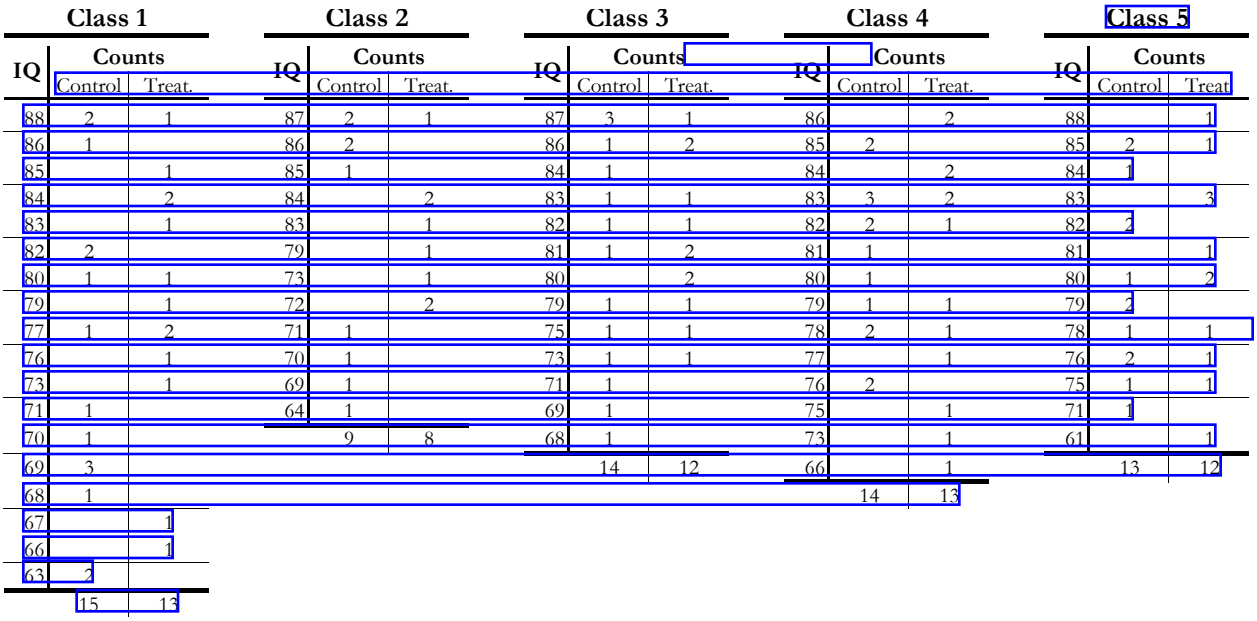
[13] Heckman (2005) raises this concern in the context of the Perry program.
[14] See Micceri (1989) for a survey.
[15] Crime measures are a case in point.

Figure 1: Perry Randomization Protocol*



*This figure is a visual representation of the Perry Randomization Protocol. T and C refer to treatment and control groups respectively. Blue circles represent males. Pink circles represent females. G_1 and G_2 are unlabeled groups of participants.

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



Class 1			Class 2			Class 3			Class 4			Class 5		
IQ	Counts		IQ	Counts		IQ	Counts		IQ	Counts		IQ	Counts	
	Control	Treat.		Control	Treat.		Control	Treat.		Control	Treat.		Control	Treat
88	2	1	87	2	1	87	3	1	86		2	88		1
86	1		86	2		86	1	2	85	2		85	2	1
85		1	85	1		84	1		84		2	84	1	
84		2	84		2	83	1	1	83	3	2	83		3
83		1	83		1	82	1	1	82	2	1	82	2	
82	2		79		1	81	1	2	81	1		81		1
80	1	1	73		1	80		2	80	1		80	1	2
79		1	72		2	79	1	1	79	1	1	79	2	
77	1	2	71	1		75	1	1	78	2	1	78	1	1
76		1	70	1		73	1	1	77		1	76	2	1
73		1	69	1		71	1		76	2		75	1	1
71	1		64	1		69	1		75		1	71	1	
70	1			9	8	68	1		73		1	61		1
69	3						14	12	66		1		13	12
68	1									14	13			
67		1												
66		1												
63	2													
	15	13												

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

reassigned for a subset of persons after an initial random assignment. This creates two potential problems.

First, such reassignments can induce correlation between treatment assignment and baseline characteristics of participants. If the baseline measures affect outcomes, treatment assignment can become correlated with outcomes through an induced common dependence. Such a relationship between outcomes and treatment assignment violates the assumption of independence between treatment assignment and outcomes in the absence of treatment effects. Moreover, reassignment produces an imbalance in the covariates between the treated and the controlled, as documented in Figures 2 and 3. For example, the working status of the mother was one basis for reassignment to the control group. Weikart, Bond, and McNeil (1978, p. 18) note that at baseline, children of working mothers had higher test scores. Not controlling for mother's working status would bias downward estimated treatment effects for schooling and other ability-dependent outcomes. We control for imbalances by conditioning on such covariates.