The Econometrics Journal

*Econometrics Journal* (2021), volume **24**, pp. C1–C39.
https://doi.org/10.1093/ectj/utab009

# Using a satisficing model of experimenter decision-making to guide finite-sample inference for compromised experiments

JAMES J. HECKMAN[†] AND GANESH KARAPAKULA[‡]

[†]*Center for the Economics of Human Development, The University of Chicago, Chicago, IL 60637, USA.*
Email: jjh@uchicago.edu

[‡]*Yale University, New Haven, CT 06511, USA.*
Email: ganesh.karapakula@yale.edu

First version received: 29 December 2020; final version accepted: 23 March 2021.

**Summary:** This paper presents a simple decision-theoretic economic approach for analysing social experiments with compromised random assignment protocols that are only partially documented. We model administratively constrained experimenters who satisfice in seeking covariate balance. We develop design-based small-sample hypothesis tests that use worst-case (least favourable) randomization null distributions. Our approach accommodates a variety of compromised experiments, including imperfectly documented rerandomization designs. To make our analysis concrete, we focus much of our discussion on the influential Perry Preschool Project. We reexamine previous estimates of programme effectiveness using our methods. The choice of how to model reassignment vitally affects inference.

**Keywords:** *Randomized controlled trial*, *rerandomization*, *satisficing*, *worst-case randomization tests*, *partial identification*, *design-based least favourable small-sample inference*.

**JEL codes:** *D91*.

## 1. INTRODUCTION

This paper develops a finite-sample, design-based approach for analysing data from compromised social experiments using a satisficing model of experimenter behaviour. Compromises can take many forms, including exchanges or transfers of subjects across the experimental groups based on post-randomization considerations that are not fully documented. For specificity, we motivate our approach drawing on the Perry Preschool Project, an experimental high-quality preschool programme targeted toward disadvantaged African American children in the 1960s.[1]

Previous studies of the Perry programme report substantial treatment effects on numerous outcomes.[2] These studies have greatly influenced discussions about the benefits of early childhood programmes.[3] However, critics of the Perry programme question the validity of these conclusions. They point to the small sample size of the experiment—just over a hundred observations. They also mention incomplete knowledge of, and compromises in, the randomization protocol used

[1] See Schweinhart et al. (1985; 1993; 2005), Heckman et al. (2010a), and Appendix A for more background.
[2] See, e.g., Heckman et al. (2010a) and Heckman et al. (2020).
[3] See Obama (2013).

to form control and treatment groups. Problems with attrition and nonresponse are also cited. Previous research (Heckman et al., 2010a; Heckman et al., 2020) addresses some of these concerns. We offer an alternative approach that models experimenter decision-making in conducting the experiment. We compare our approach with that of Heckman et al. (2020) in Section 4.4.

The Perry randomization protocol was a multi-stage process. Its main compromised feature is shared by many randomized controlled trials: undocumented rerandomization. This involves reassignment of treatment status after initial random assignment in order to improve balance between experimental groups with respect to baseline covariates, but without a pre-specified, fully documented reassignment plan.

This practice occurs often. Bruhn and McKenzie (2009) survey 25 leading researchers using randomized experiments and report a typical response:

> "*[Experimenters] regressed variables like education on assignment to treatment, and then re-did the assignment if these coefficients were too big*."

Some 52% admit to "subjectively deciding whether to redraw" and 15% admit to "using a statistical rule to decide whether to redraw" the treatment assignment vector in at least one of the experiments they conducted.[4] The authors conclude that

> "*this reveals common use of methods to improve baseline balance, including several rerandomization methods not discussed in print*."

The approach developed in this paper applies to experiments conducted in such a subjective and incompletely documented manner. If rerandomization criteria are specified and adhered to before carrying out final treatment assignment, there exist simpler methods for conducting valid inference.[5] We supplement the literature by considering the case where the reassignment rule is only partially documented. We build on and complement the analysis of Heckman et al. (2020) with an explicit model of experimenter behaviour.

We model experimenters as decision-makers who satisfice in seeking to achieve covariate balance with a "suitable" metric. Implicit decision rules underlie all covariate balancing procedures. The decision-makers forming the experimental groups do not necessarily have a precise rule in mind but satisfice in the sense of Simon (1955). Even if experimenters have a specific rule in mind, it may not be carefully documented.

This paper proceeds in the following way. Section 2 illustrates the class of problems addressed in this paper by reexamining the reassignment protocols of an influential compromised small-sample social experiment. Section 3 presents a satisficing model of experimenter behaviour consistent with the available information on it from published and informal accounts. We partially identify the set of randomization protocols consistent with our model. We consider the generality of our approach by discussing the class of experiments to which our model applies. In Section 4, we first discuss hypotheses of interest and conventional testing procedures used in the literature. We then construct worst-case randomization tests using stochastic approximations of

[4] These percentages are calculated by weighting each survey respondent by the number of experiments in which the respondent had participated.

[5] See, e.g., Morgan and Rubin (2012; 2015) and Li et al. (2018). Morgan and Rubin (2012) state that they "only advocate rerandomization if the decision to rerandomize or not is based on a pre-specified criterion." Their inferential methods require knowledge of such pre-specified criteria. Although rerandomization methods have the property that they reduce variance of the null distribution *asymptotically* in certain settings (Morgan and Rubin, 2012; 2015; Li et al., 2018), this property is not guaranteed in the finite-sample setting we consider.

least favourable randomization null distributions. We also compare our approach with an existing method for inference with imperfect randomization. Section 5 presents our test statistics and uses our methodology to reexamine the inference reported by Heckman et al. (2020). Section 6 concludes.

## 2. THE MOTIVATING PROBLEM

To give specificity to our analysis we draw on the Perry Preschool Project, a prototypical social experiment that was conducted in the early 1960s. The original sample for the experiment consisted of 128 children. Five of these children were dropped from the study due to extraneous reasons.[6] Starting at age 3, treatment in the following two years included preschool for 2.5 hours per day on weekdays during the academic year. The programme also offered 1.5-hour weekly home visits by the Perry teachers to promote parental engagement with the child.[7] For more details on the background and eligibility criteria of the Perry programme, see Heckman et al. (2010a) and Appendix A.

### *2.1. Randomization protocol*

Understanding the randomization protocol is essential for constructing valid frequentist inference for any experiment. As Bruhn and McKenzie (2009) emphasize, many experimental studies in economics do not report the complete set of rules (e.g., balancing criteria) used to form experimental samples. They conduct hypothesis tests that ignore the randomization protocols actually used. In analysing the Perry data, this issue is salient. Reports vary about the procedure used and the exact rules followed in creating experimental samples. We discuss the various descriptions of the randomization protocols. While the core descriptions of the procedure followed are broadly consistent across texts, some of the details provided are vague and inconsistent, even those by the same authors. We account for this ambiguity in designing and interpreting our hypothesis tests. While the details are Perry-specific, the general principles involved are not.

Before the initiation of the randomization procedure by the Perry staff in each of the last four Perry cohorts, any younger siblings of participants enrolled in previous waves are separated from children of freshly recruited families, whom we term "singletons" (Schweinhart et al., 1985; Schweinhart, 2013). As Schweinhart et al. (1985) explain,

> "*[A]ny siblings [are] assigned to the same group [either treatment or control] as their older siblings in order to maintain the independence of the groups.*"

[6] According to Schweinhart et al. (2005), "4 children did not complete the preschool programme because they moved away and 1 child died [in a fire accident] shortly after the study began." We are missing the following data (on some of these children) that are necessary for inference procedures. We do not know the mother's working status at baseline of a subject in wave 0 (who has a sibling in wave 1) among the five children who dropped out of the original sample of 128 for extraneous reasons. We also do not know the gender of a subject in wave 1. (We use the Perry convention that wave 0 is the first wave and wave 4 is the last one.) The baseline information on these subjects is important in our formal model of the randomization protocol. We do not make assumptions regarding the mother's working status at baseline of the subject in wave 0 and the gender of the other subject in wave 1. We run our testing procedures for each of the possible values of the variables. While we use the data on the five dropped children in our simulations of the randomization protocol for our worst-case tests, we treat the five participants as ignorable in our estimation of the treatment effects. Thus, our effective sample for estimation and inference is the core sample of 123 children.

[7] Those in the treatment group of the first entry cohort (wave 0) were provided with the intervention for only one year, starting at age 4, and thus constitute an exception. Our estimates of treatment effects pool all five cohorts, even though the lower programme intensity in the first cohort might in principle attenuate the magnitudes of the effects downward.

By construction this does not apply to the very first cohort.

The singletons from new families are then randomized into the two experimental groups as follows. Weikart et al. (1978) detail the second step of the randomization protocol:

> "*First, all [singletons] are rank-ordered according to Stanford–Binet [IQ] scores. Next, they are sorted (odd/even) into two groups.*"

Singletons are then divided into two groups, one comprising those with even IQ ranks and another with odd IQ ranks. The latter group has one additional person if the singletons are odd in number; otherwise, the sizes of the two groups are equal.

In the third step, children are exchanged between the two groups to balance the vector of means of an index of socioeconomic status (SES), the proportions of boys and girls, and the proportion of children with working mothers, in addition to mean IQ (Weikart et al., 1964; Schweinhart et al., 1993). The exact balancing criteria and the number of exchanges are not specified, and the exchanges are not necessarily restricted to those between consecutively ranked IQ pairs,[8] as is sometimes assumed, e.g., in Heckman et al. (2020). After the first three steps, there are two undesignated groups that differ in number by at most one, and the two groups are balanced with respect to mean IQ, mean SES, percentage of boys, and the proportion of children with working mothers, in a manner acceptable to the staff, using balancing rules that are undocumented.

All sources agree that in the fourth step a toss of a fair coin decides assignment of the two groups to treatment and control conditions. The fifth step concerns children with working mothers who are placed in the treatment group after the fourth step. In the fifth step, some of these children are transferred to the control group.[9] Although there is no consistent account of the number of transfers, the sources describe the fifth step as involving one-way transfers of some children of working mothers from the treatment group to the control group.[10] Weikart et al. (1978) provide reasons for the transfers: "no funds were available [to provide all working mothers with logistical support, and] special arrangements could not always be made." We interpret this statement as implying that special arrangements could be made for at least some working mothers to enable their children to attend preschool and participate in home visits if placed in the treatment group. The constraints facing programme administrators in doing so likely vary across cohorts. We assume that the Perry staff are impartial as to which working mothers get special arrangements.

Table 1 summarizes the randomization protocol. The main sources of ambiguity are in boldface: (a) the undocumented balancing criteria and rules used to satisfactorily balance the two undesignated groups with respect to the mean levels of baseline variables in the third step; and (b) the nature of constraints on the provision of special home visitation arrangements for children of working mothers in the fifth step.

[8] See Appendix B. According to Schweinhart et al. (1993), "[The staff] exchanged several similarly ranked pair members so the two groups would be matched on [the baseline variables]." Even though the phrase "similarly ranked pair members" might suggest consecutively ranked members, this is not necessarily the case. In Appendix B, we use Perry data from wave 4 to demonstrate that the exchanges were not necessarily between consecutively ranked pairs.

[9] See Schweinhart and Weikart (1980); Schweinhart et al. (1985; 1993); Zigler and Weikart (1993).

[10] This is also manifested in the observed data. For example, as explained later in Section 3.2, the number of singletons in wave 2 is 22, with 12 in the control group and 10 in the treatment group. If there were exchanges between the initial experimental groups instead of one-way transfers to the control group, there would have been 11 singletons in both the control and treatment groups instead of 12 and 10, respectively.

**Table 1.** Schematic of the actual randomization protocol.

| |
|---|
| (1) Recruit participants and separate any younger siblings of participants enrolled in previous waves from singletons (children of freshly recruited families) |
| ↓ |
| (2) Rank singletons by IQ and split into two groups based on whether the rank is even or odd |
| ↓ |
| (3) Exchange singletons between the two groups to **satisfactorily balance** the mean levels of a vector of IQ, SES, gender, and mother's working status |
| ↓ |
| (4) Toss a fair coin to determine which of the two groups becomes the initial treatment group |
| ↓ |
| (5) Transfer **some children** of working mothers from the treatment group to the control group impartially if special arrangements for home visits can be made for only a limited number |
| ↓ |
| (6) Assign any eligible younger siblings to the same group as their enrolled older siblings |

## 3. MODELLING AND PARTIALLY IDENTIFYING THE RANDOMIZATION PROTOCOL

Since no precise description of the full Perry randomization protocol exists, we do not know who was exchanged in the third step and who was transferred in the fifth step, making a standard bounding analysis intractable. To address this problem, we assume that experimenters *satisfice*[11] in seeking "balance" in the baseline covariate means of treatment and control groups, while facing capacity constraints on special home visits for children of working mothers.

Using this model, we bound the level of covariate balance deemed acceptable by the experimenters at the end of the first three stages of the protocol. We also bound the number of possible transfers at the fifth stage of the assignment procedure. Our model and the identified bounds are used to construct worst-case randomization tests using least favourable null distributions for treatment effects. While the details differ, the approach readily generalizes to the class of compromised rerandomization designs discussed by Bruhn and McKenzie (2009).

### *3.1. Formalizing the randomization protocol*

We first model the Perry randomization protocol and later discuss its generalizability. Let $\mathcal{S}_c$ be the set of unique identifiers of participants in cohort[12] $c \in \{0, 1, 2, 3, 4\}$ with no elder siblings already enrolled in the Perry Preschool Project. The cardinality of the set of singletons is $|\mathcal{S}_c|$.[13] The participants in the set $\mathcal{S}_c$ are ranked according to their IQs by the Perry staff, using an undocumented method to break any ties. The participants with odd and even ranks are then

[11] See Simon (1955), an early paper in behavioural economics that analyses satisficing behaviour.

[12] Each of the cohorts corresponds to one of the five waves (labelled 0 to 4) of study participants recruited from the autumn seasons of 1962 to 1965. Waves 0 and 1 were randomized in the autumn of 1962, while the waves 2, 3, and 4 were randomized in the autumn of 1963, 1964, and 1965, respectively. We follow the labelling convention for the cohorts by the Perry analysts who designate the first cohort as "0."

[13] Note that the other participants in cohort $c$ who are not singletons have older siblings already enrolled in the Perry experiment in a previous wave. The nonsingletons are not randomized but, rather, assigned to the same treatment status as their elder siblings already enrolled in the study.

split into two undesignated groups, with $\lceil |\mathcal{S}_c|/2 \rceil$ and $\lfloor |\mathcal{S}_c|/2 \rfloor$ members, respectively.[14] Staff exchange participants between the two groups until the mean levels of four variables (Stanford–Binet IQ, index of SES, gender, and mother's working status) are balanced to their satisfaction.[15] The exact metric the staff used to determine satisfactory covariate balance is not documented.

We assume that they use Hotelling's two-sample $t$-squared statistic $\tau_c^2$, which is closely related to the Mahalanobis distance metric often used in matching.[16] However, for each cohort's initial groups (partially identified in Section 3.2), the Hotelling statistic and raw mean differences do not correspond to their possible minimum values and are sometimes far away from them.[17] Thus, it appears in terms of this model that programme officials were satisficing rather than optimizing (minimizing covariate imbalance) in constructing the two groups.

This process results in a partition $(\mathcal{A}_c^*, \mathcal{B}_c^*)$ of the set $\mathcal{S}_c$ chosen uniformly from

$$\mathbb{U}_c(\delta_c) = \{(\mathcal{A}, \mathcal{B}) : \mathcal{A} \subset \mathcal{S}_c,\ \mathcal{B} = \mathcal{S}_c \setminus \mathcal{A},\ |\mathcal{A}| = \lceil |\mathcal{S}_c|/2 \rceil,\ \tau_c^2(\mathcal{A}, \mathcal{B}) \leq \delta_c\}, \tag{3.1}$$

where $\delta_c$ is a satisficing threshold that captures how stringent or lax the Perry staff were in trying to balance the mean levels of the two groups.[18] Note that the above set is invariant to the choice of any strictly increasing transformation of the Hotelling statistic and the corresponding satisficing threshold. Define $D_{i,c}^{(0)}$ as an indicator of whether participant $i \in \mathcal{S}_c$ belongs to $\mathcal{A}_c^*$. In other words, $D_{i,c}^{(0)} = \mathbb{I}\{i \in \mathcal{A}_c^*\}$.

In the next stage, the Perry staff flip a fair coin to determine whether $\mathcal{A}_c^*$ or $\mathcal{B}_c^*$ becomes the preliminary treatment group. Let $Q_c$ be an indicator of whether the coin flip results in a head. If $Q_c = 1$, then $\mathcal{B}_c^*$ becomes the treatment group. If $Q_c = 0$, then $\mathcal{A}_c^*$ becomes the treatment group. Let $D_{i,c}^{(1)}$ denote membership in the preliminary treatment group. Thus

$$D_{i,c}^{(1)} = Q_c\,(1 - D_{i,c}^{(0)}) + (1 - Q_c)\,D_{i,c}^{(0)}. \tag{3.2}$$

[14] Note that $\lceil \cdot \rceil \equiv \text{ceil}(\cdot)$ is the ceiling function and $\lfloor \cdot \rfloor \equiv \text{floor}(\cdot)$ is the floor function. They assign the least upper integer bound and greatest lower integer bound to the argument in the function, respectively.

[15] An exchange means a swap between two participants belonging to different undesignated groups. Since the Perry experiment did not use a matched pair design, an exchange or swap is not restricted to occur between participants with consecutive IQ ranks. Exchanges between participants with nonconsecutive IQ ranks can occur. See Appendix B.

[16] The Hotelling's multivariate two-sample $t$-squared statistic $\tau_c^2$ maps a partition $(\mathcal{A}, \mathcal{B})$ of $\mathcal{S}_c$ (such that $|\mathcal{A}| = \lceil |\mathcal{S}_c|/2 \rceil$ and $\mathcal{B} = \mathcal{S}_c \setminus \mathcal{A}$) to $\mathbb{R}_{\geq 0}$ and is given by $\tau_c^2(\mathcal{A}, \mathcal{B}) = \left(\bar{Z}_{\mathcal{A}} - \bar{Z}_{\mathcal{B}}\right)' \left(|\mathcal{A}|^{-1} \hat{\Sigma}_{\mathcal{A}} + |\mathcal{B}|^{-1} \hat{\Sigma}_{\mathcal{B}}\right)^{-1} \left(\bar{Z}_{\mathcal{A}} - \bar{Z}_{\mathcal{B}}\right)$, where $\bar{Z}_{\mathcal{A}} = |\mathcal{A}|^{-1} \sum_{i \in \mathcal{A}} Z_i$, with $Z_i$ as the pre-programme covariate vector containing the $i$-th participant's IQ, SES index, gender, and mother's working status, $\bar{Z}_{\mathcal{B}} = |\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} Z_i$, and $\hat{\Sigma}_{\mathcal{A}} = (|\mathcal{A}| - 1)^{-1} \sum_{i \in \mathcal{A}} (Z_i - \bar{Z}_{\mathcal{A}})(Z_i - \bar{Z}_{\mathcal{A}})'$, while $\hat{\Sigma}_{\mathcal{B}} = (|\mathcal{B}| - 1)^{-1} \sum_{i \in \mathcal{B}} (Z_i - \bar{Z}_{\mathcal{B}})(Z_i - \bar{Z}_{\mathcal{B}})'$. We use this metric for dimensionality reduction and computational feasibility. Chung and Romano (2016) show, without assuming normality, that the permutation distribution of $\tau_c^2$ is asymptotically chi-squared. If adequate computational power were available, we could also incorporate into our model the raw mean differences in the four variables, their studentized versions, or other measures of mean differences between two groups. Of course, it is possible that the Perry staff were just looking at mean differences and did not use any formal metric.

[17] For cohort 0, the proportion of possible group formations with a lower Hotelling statistic is at least 29.24%. The corresponding numbers for cohorts 1, 2, 3, and 4 are 64.51%, 14.79%, 9.76%, and 75.56%, respectively. Similarly, the raw mean differences in baseline covariates for the initial groups also do not correspond to their minimum possible values.

[18] The satisficing threshold $\delta_c$ is the maximum level of covariate imbalance that satisficed Perry staff. The threshold $\delta_c$ is unknown to the analyst but can be partially identified, as explained later. We assume a uniform probability over $\mathbb{U}_c$ for the choice of the partition $(\mathcal{A}_c^*, \mathcal{B}_c^*)$ for the purpose of keeping the model simple and computationally feasible. In general, we might suspect the following: given two partitions of $\mathcal{S}_c$ with the same level of Hotelling's statistic, there might have been a higher probability mass on the partition closer to the initial grouping based on odd and even IQ ranks. In addition, the staff might have also preferred not to make additional exchanges if they expected relatively insignificant reductions in covariate imbalance. In other words, the probability that the Perry staff chose a particular partition $(\mathcal{A}_c^*, \mathcal{B}_c^*)$ could have depended on their preferences over substitution between two things: similarity of $(\mathcal{A}_c^*, \mathcal{B}_c^*)$ to the initial IQ rank-based grouping; and the level of covariate imbalance (as measured by Hotelling's statistic) resulting from the partition $(\mathcal{A}_c^*, \mathcal{B}_c^*)$. However, there is no unique way to formalize this notion. Such a general model may not even be computationally feasible.

In the next step, some children of working mothers initially placed in the treatment group are transferred to the control group.[19] To model this process, we introduce additional notation. Define $M_i$ as an indicator of whether participant $i$'s mother was working at baseline. Cohorts 0 and 1 were both randomized in the autumn of 1962, while each of the remaining cohorts were randomized in successive years from 1963 to 1965. For cohorts $c \in \{2, 3, 4\}$, let $m_c$ be the number of children of working mothers initially placed in the treatment group: $m_c = \sum_{i \in \mathcal{S}_c} M_i \, D_{i,c}^{(1)}$. For the entry cohorts, let $m_{0,1}$ be the number of children of working mothers initially placed in the treatment group for cohorts 0 and 1, that is, $m_{0,1} = \sum_{c \in \{0,1\}} \sum_{i \in \mathcal{S}_c} M_i \, D_{i,c}^{(1)}$.

Define $\eta_c$ as a parameter indicating the maximum number of children of working mothers in cohort $c \in \{2, 3, 4\}$ for whom special arrangements could be made to enable special home visits.[20] We define $\eta_{0,1}$ to be the parameter indicating the maximum number of children of working mothers in the pooled cohorts 0 and 1 for whom special home visitation arrangements could be made, averting their transfer to the control group if placed in the initial treatment group.[21]

Special arrangements are made for $\min(\eta_{0,1}, m_{0,1})$ children of working mothers in the entry cohorts and for $\min(\eta_c, m_c)$ such children in each cohort $c \in \{2, 3, 4\}$ to enable special home visits, as opposed to weekday home visits for children of nonworking mothers. If there are any remaining children with working mothers in the initial treatment group, they are transferred to the control group, potentially increasing covariate imbalance.[22] We assume that the Perry staff impartially choose (with equal probability) the children for whom the special accommodations are made.[23] To formalize this assumption, let $V_{i,c}$ be a binary indicator for whether the participant $i \in \mathcal{S}_c$ was placed the initial treatment group, had a working mother, and remained in the treatment group through special accommodations for home visits. The vector $(V_{i,c} : i \in \mathcal{S}_c, \; M_i \, D_{i,c}^{(1)} = 1)$ is assumed to be drawn uniformly from the set $\{v \in \{0, 1\}^{m_c} : ||v||_1 = \min(\eta_c, m_c)\}$ for last three cohorts $c \in \{2, 3, 4\}$. Since the two entry cohorts face a common capacity constraint with respect to special home visitation accommodations, the vector $(V_{i,c} : i \in \mathcal{S}_0 \cup \mathcal{S}_1, \; M_i \, D_{i,c}^{(1)} = 1)$ is assumed to be drawn uniformly from the set $\{v \in \{0, 1\}^{m_{0,1}} : ||v||_1 = \min(\eta_{0,1}, m_{0,1})\}$. In addition, $V_{i,c} = 0$ for a participant $i \in \mathcal{S}_c$ if $M_i \, D_{i,c}^{(1)} = 0$ for all $c \in \{0, 1, 2, 3, 4\}$.[24] In this notation, the participant's final treatment status $D_{i,c}^{(2)}$ is given by

$$D_{i,c}^{(2)} = M_i \, D_{i,c}^{(1)} \, V_{i,c} + (1 - M_i \, D_{i,c}^{(1)}) \, D_{i,c}^{(1)}. \tag{3.3}$$

[19] The Perry teachers conducted special home visits for working mothers at times other than weekday afternoons, when they visited the homes of nonworking mothers. Because of logistical and financial constraints, the teachers were able to visit the homes of only a limited number of working mothers at times other than weekday afternoons. Thus, the children of working mothers in the preliminary treatment group for whom these special arrangements could not be made were transferred to the control group.

[20] Thus, $\eta_c$ can be thought of as slots available for special visits to the homes of working mothers. Equivalently, it is the number of children of working mothers who would remain in the final treatment group if all of them were placed in the preliminary treatment group.

[21] Since cohorts 0 and 1 had a common set of teachers, they share the number of slots available for the special home visits. Thus, we pool these two cohorts while defining $m_{0,1}$ and $\eta_{0,1}$. However, cohorts 2 to 5 have separate parameters for the slots available for special home visits.

[22] It is possible that the Perry staff engaged in another round of satisficing at this step. In principle, this could be incorporated into our model but would increase its dimensionality. Since the published accounts do not mention another round of balancing, we do not add this feature to our model to keep it computationally feasible.

[23] We are implicitly assuming that all working mothers would be able to send their children to preschool and participate in weekly home visits if special arrangements could be made for them. A model allowing for heterogeneity in availability of working mothers (for special arrangements) does not appear to be computationally feasible.

[24] In other words, $V_{i,c} = 0$ for the participants who were either initially placed in the control group or placed in the initial treatment group but have nonworking mothers.

Any Perry subjects with identifiers not in $\bigcup_{c=0}^{4} \mathcal{S}_c$ receive the same treatment status as their elder siblings already enrolled in the Perry study. Thus, the final treatment status $D_i$ of the $i$-th subject is given by $D_i = D_{i,c}^{(2)}$ if $i \in \bigcup_c \mathcal{S}_c$. Otherwise, if participant $i$ is not from a freshly recruited family, the assignment is given by $D_i = D_h$, where the $h$-th subject is the $i$-th subject's eldest sibling enrolled in the Perry study, if $i \in \mathcal{I} \setminus \bigcup_c \mathcal{S}_c$, where $\mathcal{I}$ is the set of all Perry subjects.

### *3.2. Partially identifying satisficing thresholds and capacity constraints*

Using the Perry data, we now demonstrate how we can partially identify the satisficing thresholds $\delta_c$ and the special home visitation capacity constraints $\eta_c$ using the last three cohorts as examples. We then present a general framework for partially identifying these parameters.

**Example 1:** Wave 2 (A case with 1 transfer in the last stage).

| Wave 2 | $D_i = 0$ | $D_i = 1$ | Total |
|---|---|---|---|
| $M_i = 0$ | 9 | 7 | 16 |
| $M_i = 1$ | 3 | 3 | 6 |
| Total | 12 | 10 | 22 |

Example 1 discusses the steps for bounding the parameters $\delta_2$ and $\eta_2$ in wave 2. Shown is a contingency table of mother's working status $M_i$ and final treatment status $D_i$ for participants $i \in \mathcal{S}_2$ in cohort 2 with no elder siblings already enrolled in the Perry study. There are 22 such participants in total. Since there are an even number of participants, each of the initial two undesignated groups (as well as the initial treatment and control groups in the next stage) would have been $\lceil |\mathcal{S}_2|/2 \rceil = \lfloor |\mathcal{S}_2|/2 \rfloor = 11$ in size. However, we observe only 10 members in the final treatment group but 12 members in the final control group. This implies that there must have been one transfer from the initial treatment group to the control group. Thus, one of the 3 children of working mothers in the final control group was in the initial treatment group. However, we do not know exactly which one of these children was transferred, so there are 3 possibilities for the initial treatment group. Let $\tau_{2,1}^2$, $\tau_{2,2}^2$, $\tau_{2,3}^2$ be the Hotelling two-sample statistics for these 3 possibilities. One of these Hotelling statistics was the actual level of covariate imbalance between the initial treatment and control groups, and this level of imbalance is assumed to be within the satisficing threshold $\delta_2$ of the Perry staff (by construction). Thus, $\delta_2 \geq \min\{\tau_{2,1}^2, \tau_{2,2}^2, \tau_{2,3}^2\}$. In addition, $m_2 = 4$, since there must have been 4 children of working mothers in the initial treatment group, consisting of the 3 participants who remain in the final treatment group and the 1 participant who was transferred to the control group. Since 3 of the initial 4 participants remained in the final treatment group, $\min(\eta_2, m_2) = \min(\eta_2, 4) = 3$, implying that $\eta_2 = 3$, the only solution that satisfies the equality. We next present two other examples.

**Example 2:** Wave 3 (A case with 1 or 2 transfers in the last stage).

| Wave 3 | $D_i = 0$ | $D_i = 1$ | Total |
|---|---|---|---|
| $M_i = 0$ | 7 | 9 | 16 |
| $M_i = 1$ | 5 | 0 | 5 |
| Total | 12 | 9 | 21 |

In Example 2 we show a contingency table of $M_i$ and $D_i$ for the 21 participants $i \in \mathcal{S}_3$ in cohort 3. The sizes of the larger and smaller undesignated groups would have been $\lceil |\mathcal{S}_3|/2 \rceil = 11$

and $\lfloor |\mathcal{S}_3|/2 \rfloor = 10$, respectively. However, either of these two groups could have been the initial treatment group. Since there are 12 members in the final control group and 9 in the final treatment group, there are 2 possible cases: if the initial treatment group had 10 members, there would have been $10 - 9 = 1$ transfer; but if it had 11 members, there would have been $11 - 9 = 2$ transfers. Since the number of transfers involving children of working mothers is either 1 or 2, the number of possibilities for the initial treatment group is $\binom{5}{1} + \binom{5}{2} = 5 + 10 = 15$, as all the 5 children of working mothers in this cohort are in the control group. Let $\tau^2_{3,1}, \ldots, \tau^2_{3,15}$ be the Hotelling statistics for those 15 possibilities. Then, $\delta_3 \geq \min\{\tau^2_{3,1}, \ldots, \tau^2_{3,15}\}$. In addition, $m_3 \in \{1, 2\}$, since $m_3$ is the sum of the number of transfers (either 1 or 2) and the number of remaining children in the final treatment group (0 in this cohort). As no working mother remained in the treatment group, $\min(\eta_3, m_3) = 0$, implying that $\eta_3 = 0$, which is the only number consistent with this equality. Thus, the Perry staff were unable to provide special home visitation accommodations for any of the participants in this cohort.

**Example 3:** Wave 4 (A case with no transfers in the last stage).

| Wave 4 | $D_i = 0$ | $D_i = 1$ | Total |
|---|---|---|---|
| $M_i = 0$ | 5 | 10 | 15 |
| $M_i = 1$ | 4 | 0 | 4 |
| Total | 9 | 10 | 19 |

In Example 3 we show a contingency table of $M_i$ and $D_i$ for the 19 participants $i \in \mathcal{S}_4$ in cohort 4. The sizes of the larger and smaller undesignated groups would have been $\lceil |\mathcal{S}_3|/2 \rceil = 10$ and $\lfloor |\mathcal{S}_3|/2 \rfloor = 9$. These coincide with the final sizes of the treatment and control groups, respectively. Accordingly, we can conclude that the observed final treatment group was indeed the initial treatment group for this cohort. Otherwise, the control group would have had at least 10 members. Let $\tau^2_{4,1}$ be the Hotelling statistic for the observed partition of $\mathcal{S}_4$ based on the final treatment status. Then, $\delta_4 \geq \tau^2_{4,1}$. In addition, note that there are no children of working mothers in the final treatment group, which was also the initial treatment group, and so $m_4 = 0$. Since $\min(\eta_4, m_4) = \min(\eta_4, 0) = 0$ and there are 4 members with working mothers in total, it follows that the capacity constraint could be any of the numbers from 0 through 4, i.e., $\eta_4 \in \{0, 1, 2, 3, 4\}$, because any of these values satisfies the equality. Thus, the observed data for cohort 4 is not helpful in bounding $\eta_4$.

*Partial identification of the satisficing thresholds and capacity constraints in general.* We now present a general characterization of how to partially identify the satisficing thresholds and capacity constraints on special home visits.

| Wave $c$ | $D_i = 0$ | $D_i = 1$ | Total |
|---|---|---|---|
| $M_i = 0$ | $\omega_{0,0}$ | $\omega_{0,1}$ | $\omega_{0,*}$ |
| $M_i = 1$ | $\omega_{1,0}$ | $\omega_{1,1}$ | $\omega_{1,*}$ |
| Total | $\omega_{*,0}$ | $\omega_{*,1}$ | $\lvert\mathcal{S}_c\rvert$ |

In the above contingency table, there are $\omega_{m,d}$ participants with $(M_i, D_i) = (m, d) \in \{0, 1\}^2$ among the set of participants $\mathcal{S}_c$ in cohort $c$.[25] The total number of children with nonworking

[25] Note that $\omega_{m,d} \equiv \omega_{m,d,c}$ for all $(m, d) \in \{0, 1\}^2$ but we suppress the subscript $c$ for simplicity.

mothers is $\omega_{0,*} = \omega_{0,0} + \omega_{0,1}$ and that of working mothers is $\omega_{1,*} = \omega_{1,0} + \omega_{1,1}$. The total number of participants in the final control group is $\omega_{*,0} = \omega_{0,0} + \omega_{1,0}$ and that in the final treatment group is $\omega_{*,1} = \omega_{0,1} + \omega_{1,1}$. The partial identification of the satisficing thresholds and capacity constraints would vary depending on whether $|\mathcal{S}_c|$ is even or odd and also depending on whether $\omega_{*,1} = \lceil |\mathcal{S}_c|/2 \rceil$ or $\omega_{*,1} < \lceil |\mathcal{S}_c|/2 \rceil$. We discuss each of these cases separately.

First, consider the case where $|\mathcal{S}_c|$ is even or odd and $\omega_{*,1} = \lceil |\mathcal{S}_c|/2 \rceil$. In this case, since the size of the final treatment group remains the same as that of the initial treatment group, there must have been no transfers of children with working mothers from the treatment group to the control group. Since the final treatment group is the same as the initial one, we can bound the satisficing threshold as follows: $\delta_c \geq \tau^2_{c,1}$, where $\tau^2_{c,1}$ is the Hotelling statistic for the partition of $\mathcal{S}_c$ based on the final treatment status. In addition, since there are no transfers, the number of children of working mothers in the initial treatment group $m_c$ equals $\omega_{1,1}$. Since $\min(\eta_c, \omega_{1,1}) = \omega_{1,1}$, it follows that $\eta_c \in \{\omega_{1,1}, \ldots, \omega_{1,*}\}$, i.e., the number of slots available for special home visits must be at least the number $\omega_{1,1}$ observed in the data.

Second, consider the case where $|\mathcal{S}_c|$ is even and $\omega_{*,1} < \lceil |\mathcal{S}_c|/2 \rceil$. As in Example 1, in this case it is clear that the number of transfers in the final stage must have been $\chi_c = \lceil |\mathcal{S}_c|/2 \rceil - \omega_{*,1}$, which is a positive number. The $\chi_c$ transferred children must be among the $\omega_{1,0}$ members with working mothers in the final control group. Thus, there are $\binom{\omega_{1,0}}{\chi_c}$ possibilities for the initial treatment group. Let $\vartheta^{\delta}_c$ be the set containing the Hotelling statistics for those possibilities. Then, $\delta_c \geq \min \vartheta^{\delta}_c$. In addition, there must have been $m_c = \omega_{1,1} + \chi_c$ children with working mothers in the initial treatment group. It remains to determine which values of $\eta_c$ are consistent with the equality $\min(\eta_c, \omega_{1,1} + \chi_c) = \omega_{1,1}$. Since $\chi_c > 0$, it follows that $\eta_c = \omega_{1,1}$.

Third, consider the case where $|\mathcal{S}_c|$ is odd and $\omega_{*,1} < \lceil |\mathcal{S}_c|/2 \rceil$. As in Example 2, in this case there are two possibilities for the number $\chi_c$ of transfers in the final stage. Specifically, $\chi_c \in \{\lfloor |\mathcal{S}_c|/2 \rfloor - \omega_{*,1}, \lceil |\mathcal{S}_c|/2 \rceil - \omega_{*,1}\}$. These $\chi_c$ transferred children must be among the $\omega_{1,0}$ members with working mothers in the final control group. Thus, there are $\binom{\omega_{1,0}}{\lfloor |\mathcal{S}_c|/2 \rfloor - \omega_{*,1}} + \binom{\omega_{1,0}}{\lceil |\mathcal{S}_c|/2 \rceil - \omega_{*,1}}$ possibilities for the initial treatment group. Let $\vartheta^{\delta}_c$ be the set containing the Hotelling statistics for those possibilities. Then, $\delta_c \geq \min \vartheta^{\delta}_c$. The number $m_c$ of children with working mothers initially assigned treatment is either equal to $\omega_{1,1} + \lfloor |\mathcal{S}_c|/2 \rfloor - \omega_{*,1}$ or equal to $\omega_{1,1} + \lceil |\mathcal{S}_c|/2 \rceil - \omega_{*,1}$. Let $\vartheta^{\eta}_c$ be the set of values of $\eta_c$ consistent with the equality $\min(\eta_c, m_c) = \omega_{1,1}$. If $m_c = \omega_{1,1} + \lceil |\mathcal{S}_c|/2 \rceil - \omega_{*,1}$, then $\eta_c = \omega_{1,1}$, since $\lceil |\mathcal{S}_c|/2 \rceil > \omega_{*,1}$. However, if $m_c = \omega_{1,1} + \lfloor |\mathcal{S}_c|/2 \rfloor - \omega_{*,1}$, there are two sub-cases: if $\lfloor |\mathcal{S}_c|/2 \rfloor > \omega_{*,1}$, then $\eta_c = \omega_{1,1}$; but if $\lfloor |\mathcal{S}_c|/2 \rfloor = \omega_{*,1}$, then $\eta_c \in \{\omega_{1,1} \ldots, \omega_{1,*}\}$. Therefore, the special home visiting slots can be partially identified as follows: $\eta_c \in \vartheta^{\eta}_c$, where $\vartheta^{\eta}_c = \{\omega_{1,1} \ldots, \omega_{1,*}\}$ if $\lfloor |\mathcal{S}_c|/2 \rfloor = \omega_{*,1}$, and $\vartheta^{\eta}_c = \{\omega_{1,1}\}$ if $\lfloor |\mathcal{S}_c|/2 \rfloor > \omega_{*,1}$.

This general characterization of the partial identification of satisficing thresholds $\delta_c$ applies to all cohorts $c \in \{0, 1, 2, 3, 4\}$ but that of the special home visiting capacity constraints $\eta_c$ applies only to cohorts $c \in \{2, 3, 4\}$. However, similar reasoning can be used to partially identify the capacity constraint $\eta_{0,1}$ for pooled cohorts 0 and 1.[26]

[26] Specifically, $\eta_{0,1} \in \{\eta \in \{0, \ldots, \sum_{i \in \mathcal{S}_0 \cup \mathcal{S}_1} M_i\} : \min(\eta, \chi_0 + \chi_1 + \omega^{0,1}_{1,1}) = \omega^{0,1}_{1,1}, \chi_0 \in \mathcal{C}_0, \chi_1 \in \mathcal{C}_1\}$, where $\omega^{0,1}_{1,1} = \sum_{i \in \mathcal{S}_0 \cup \mathcal{S}_1} M_i \, D_i$ and $\mathcal{C}_c = \{\lceil |\mathcal{S}_c|/2 \rceil - \omega_{*,1,c}, \max\{0, \lfloor |\mathcal{S}_c|/2 \rfloor - \omega_{*,1,c}\}\}$ for $c \in \{0, 1\}$. In our application, $\eta_{0,1} \in \{3\}$. Since we do not make assumptions on the missing mother's working status at baseline for a subject in wave 0 and the missing gender of another subject in wave 1 (among the five who dropped out of the initial sample of 128 for extraneous reasons), our partial identification of $\delta_0$ and $\delta_1$ depends on the values in the partially identified set for the missing variables. Since we do not make assumptions on the two missing binary variables, this is a strength of our analysis, despite quadrupling the computational cost. We also use known information that there was at least one transfer in wave 0 (Weikart et al., 1964) to narrow the partially identified set for that cohort.

### *3.3. Applicability of our approach to other compromised experiments*

Our approach can be applied to many of the studies that Bruhn and McKenzie (2009) criticize, especially experiments using undocumented rerandomization. All of these experiments have the feature that some criterion determines "satisfactory balance." For example, Bruhn and McKenzie (2009) quote a survey response that says, "[experimenters] regressed variables like education on assignment to treatment, and then re-did the assignment if these coefficients were too big." With appropriate modifications, our model of satisficing thresholds directly applies to experiments conducted in such a subjective and incompletely documented manner. Suitable adjustments include replacing Hotelling's statistic in our model with studentized regression coefficients (selected by pretesting or otherwise) or other metrics actually used to measure covariate imbalance between the treatment and control groups. Our methods for partially identifying the underlying randomization rules can be used when the subjective satisficing thresholds are not documented. Even though we only use one balancing criterion (Hotelling's statistic) for dimensionality reduction in our definition of $\mathbb{U}_c(\cdot)$, it can be trivially modified to accommodate multiple balancing criteria. In addition, if the experiment has strata instead of cohorts, the $c$s in our model would correspond to strata.

If an experiment does not have transfers after forming the intermediate treatment and control groups, then there are no capacity constraints, i.e., the $\eta_c$s play no role. However, in some social experiments, post-randomization transfer of some participants from the control to the treatment group can occur if additional funding for the intervention becomes available. For example, wait-list control groups are used in some clinical studies. While this is the reverse of what occurred in the Perry experiment, our model (with appropriate modifications) can be readily applied. Overall, our approach can be adapted to analyse a variety of compromised experiments across multiple disciplines.

## 4. HYPOTHESES OF INTEREST AND INFERENCE

The conventional way to analyse randomized experiments is to posit a null hypothesis that the average effect of treatment is zero and to proceed in testing it with large-sample methods using asymptotic or bootstrap distributions. Given the relatively small size of many experimental samples, reliance on large-sample methods can be problematic.[27]

In some settings, permutation tests can be used to test the null hypothesis that the outcomes in the control group have the same distribution as those in the treatment group without relying on large-sample theory. Permutation tests exploit the property that treatment and control labels within the same strata are exchangeable under the null hypothesis of a common outcome distribution. If randomization of the treatment status did not involve explicit stratification on baseline covariates, permutation tests need to make restrictive assumptions on the strata within which treatment and control labels are exchangeable. This approach is used by Heckman et al. (2010a).[28] They assume

[27] In a set of 53 studies of randomized controlled trials published in some leading economics journals, Young (2019) also finds that experimental results obtained using asymptotic theory are misleading, relative to results based on randomization tests.

[28] However, unless the permutation method reflects the method used for random assignment of the treatment, permutation tests do not in general allow us to test hypotheses about counterfactual outcomes of the individual Perry participants.

that conditioning on covariates solves the problem of post-random assignment reallocation but without any explicit model for why it is effective in doing so.[29]

This paper uses knowledge of the randomization protocol to draw inferences about treatment effects. Once a precise null hypothesis is specified, we can determine the distribution of estimates generated by the randomization scheme and assess the statistical significance of the observed treatment effects.

In this section, we first formulate our hypotheses of interest. We then discuss conventional inferential procedures. Finally, we introduce worst-case (least favourable) randomization tests and discuss how to conduct them using stochastic approximations, and then we compare our methods with alternative approaches for inference with imperfect randomization.

### *4.1. Hypotheses of interest*

Let $Y^1$ be the treated outcome, $Y^0$ be the untreated outcome, $Z$ represent background variables, and $F$ be their joint distribution at the population level. The conventional approach tests the hypothesis $\mathcal{H}_{\mathcal{C}}$ of equality of means, i.e.,

$$\mathcal{H}_{\mathcal{C}} : \mathbb{E}_F[Y^1 - Y^0] = 0, \tag{4.1}$$

assuming that the corresponding versions $(Y_i^1, Y_i^0, Z_i)$ of those variables for individual $i$ are distributed according to $F$ for all $i \in \mathcal{P}$, where $\mathcal{P}$ is the set of experimental subjects. Because each participant in our sample is assigned to either the treatment group or the control group, we only observe either $Y_i^1$ or $Y_i^0$ for each $i \in \mathcal{P}$. The hypothesis $\mathcal{H}_{\mathcal{C}}$ is traditionally tested by applying large-sample methods to the observed data $(Y_i, D_i, Z_i)_{i\in\mathcal{P}}$, where $D_i$ is the treatment status, $Y_i = D_i\, Y_i^1 + (1 - D_i)\, Y_i^0$, and $Z_i$ is the vector of pre-programme covariates.

Instead of appealing to some hypothetical large-sample experiment to conduct inference, we seek knowledge of the small sample *in hand*. One hypothesis of interest is whether the average treatment effect *within* the sample is zero, i.e.,

$$\mathcal{H}_{\mathcal{N}} : \frac{1}{N_{\mathcal{P}}} \sum_{i\in\mathcal{P}} \left(Y_i^1 - Y_i^0\right) = 0, \tag{4.2}$$

where $N_{\mathcal{P}} = |\mathcal{P}|$.[30] A special case of $\mathcal{H}_{\mathcal{N}}$ is the *sharp* null hypothesis of no treatment effects whatsoever for *each* participant:

$$\mathcal{H}_{\mathcal{F}} : \tau_i \equiv Y_i^1 - Y_i^0 = 0, \tag{4.3}$$

for all $i \in \mathcal{P}$,[31] Fisher's (1925; 1935) null hypothesis. It involves a joint test of zero individual treatment effects and is trivially equivalent to $\mathcal{H}_{\mathcal{N}}$ if there is no heterogeneity in the treatment effects. The advantage of Fisher's hypothesis $\mathcal{H}_{\mathcal{F}}$ is that it provides a testable model in which all the counterfactual outcomes are specified.[32] Such hypothesis testing can be conducted using our

[29] In practice, their approach relies on large-sample methods in using regression analysis to condition on covariates.

[30] This is attributed to Neyman (1923).

[31] While this formulation states that each individual treatment effect $\tau_i$ is zero, the analyst may fix each $\tau_i$ at a desired value for hypothesis testing. Such a hypothesis is often called *sharp* because it specifies one set of counterfactual outcomes for the participants.

[32] Note that we observe either $Y_i^1$ or $Y_i^0$ for each participant $i \in \mathcal{P}$. Thus, under the null model (4.3), the other counterfactual outcome can be imputed according to the fact that $Y_i^1 = Y_i^0$. In general, if $\tau_i$ is hypothesized to be equal to a number $\tau_i^\circ$, the counterfactual outcomes $(Y_i^1, Y_i^0)$ under the null model are equal to $(Y_i + \tau_i^\circ, Y_i)$ if $D_i = 0$ and is equal to $(Y_i, Y_i - \tau_i^\circ)$ if $D_i = 1$ for all $i \in \mathcal{P}$.

knowledge of the randomization protocol without relying on large-sample theory. With all the counterfactual outcomes specified, we can learn about the randomization distribution of treatment effects, and we can gauge the extent to which the observed data can be rationalized using the specified null model.[33]

Hypothesis $\mathcal{H}_{\mathcal{N}}$ nests the sharp null hypothesis $\mathcal{H}_{\mathcal{F}}$. In general there are many configurations of the individual treatment effects that are all consistent with $\mathcal{H}_{\mathcal{N}}$. Thus, to test $\mathcal{H}_{\mathcal{N}}$ using only limited knowledge of the randomization protocol, we would need to test each one of all the sharp null hypotheses like $\mathcal{H}_{\mathcal{F}}$ that imply $\mathcal{H}_{\mathcal{N}}$.[34] However, a nonrejection of $\mathcal{H}_{\mathcal{F}}$ implies nonrejection of $\mathcal{H}_{\mathcal{N}}$, and so testing other sharp null hypotheses may not be necessary if we are unable to reject $\mathcal{H}_{\mathcal{F}}$. Of course, a rejection of $\mathcal{H}_{\mathcal{F}}$ would not imply a rejection of $\mathcal{H}_{\mathcal{N}}$. The latter is a very conservative criterion. We next discuss conventional hypothesis testing procedures.

### *4.2. Conventional hypothesis testing procedures*

For tests of population-level parameters such as $\mathcal{H}_{\mathcal{C}}$ in equation (4.1), the most commonly reported measure of statistical significance is the asymptotic $p$-value. For completely randomized experiments, it can be interpreted as the $p$-value based on a large-sample approximation of the distribution of an estimator, say difference-in-means, over all possible randomizations under the null hypothesis $\mathcal{H}_{\mathcal{N}}$ (Neyman, 1923). Li et al. (2018) derive an asymptotic theory of the difference-in-means estimator in experiments involving rerandomization with a pre-specified balancing rule using the Mahalanobis distance, for which the asymptotic distribution of the estimator is a linear combination of normal and truncated normal variables. Resampling methods are also widely used to quantify statistical uncertainty. For example, the bootstrap standard error is reported in many research papers with an associated bootstrap $p$-value.

Permutation tests are often used when researchers are interested in testing whether treatment and control groups have a common outcome distribution without relying on large-sample theory. Such tests rely on the property that the treatment and control labels are exchangable within each stratum of the experiment under the null hypothesis of a common distribution. In their permutation tests, Heckman et al. (2010a) use strata defined by wave, gender, and indicator for above-median socioeconomic status, assuming that experimental labels within each stratum are exchangeable. To compare their permutation procedures with the methods developed in this paper, we use a simplified version of their permutation tests using block permutations within cohorts of eldest participant-siblings (whose treatment statuses determine that of their younger participant-siblings).

In the Perry context, Heckman et al. (2020) develop an extension of permutation tests to account for imperfect randomization. In this paper, we offer an alternative design-based approach to conduct inference for a broader class of compromised experiments. We first present our approach and then compare it with theirs.

[33] See Athey and Imbens (2017) and Abadie et al. (2020) for background on this topic. Also, note that our randomization tests are conditional tests that exploit random variation in the treatment status but fix the other observed data. See Lehmann (1993).

[34] When the outcomes under consideration are binary and the experiment involves a completely randomized design, there are strategies to test the weak null hypothesis in a computationally feasible way (see, e.g., Li and Ding, 2016; Rigdon and Hudgens, 2015).

### 4.3. Worst-case randomization tests

This paper advocates and uses worst-case approximate randomization tests to analyse the Perry data. Fisher's sharp null hypothesis $\mathcal{H}_{\mathcal{F}}$ specifies all the counterfactual outcomes, which are imputed according to the hypothesis using the observed data. If we knew the exact randomization protocol of the Perry experiment, we could measure where the observed test statistic falls along its exact randomization distribution, i.e., the distribution of the test statistic over all possible treatment status vectors that could have been hypothetically generated by the randomization protocol. The more extreme the observed test statistic falls along the null distribution, the more incompatible the observed data would be with the sharp null hypothesis. However, for Perry and many other social experiments, the exact randomization protocol is unknown: even in our stylized model of the randomization protocol, the satisficing thresholds and capacity constraints are only partially identified. To account for this ambiguity, we could in theory conduct randomization tests[35] over the set of all possible randomization protocols. Thus, we could conduct the worst-case randomization test, conditional on the observed outcomes and baseline covariates,[36] using the least favourable distribution among all the possible randomization distributions. This results in the following worst-case $p$-value that serves as an upper bound for the true randomization $p$-value:

$$p_w(D) = \sup_{\gamma^* \in \Xi} \mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_{\gamma^*}) \geq T(D)\}, \tag{4.4}$$

where $\Xi$ is the partially identified set[37] for $\gamma = (\delta_0, \ldots, \delta_4, \eta_{0,1}, \eta_2, \eta_3, \eta_4)$, the vector of true values of parameters (satisficing thresholds and capacity constraints), $\mathbb{P}_{\Lambda_{\gamma^*}}$ represents probability (conditional on the observed outcomes and pre-programme covariates) under the probability space $\Lambda_{\gamma^*}$ of randomizations generated by the protocol parameterized by $\gamma^*$, $\tilde{D}_{\gamma^*}$ represents a random treatment status vector defined on the probability space $\Lambda_{\gamma^*}$, $D$ denotes the observed treatment status vector, and $T(\cdot)$ is the chosen test statistic such that $T(\cdot)$ maps a treatment status vector to a real number measuring the magnitude of the outcome difference between the treatment and control groups. Since the sharp null hypothesis specifies counterfactual outcomes, the data $(Y_i^0, Y_i^1, Z_i)_{i \in \mathcal{P}}$ are fixed according to $\mathcal{H}_{\mathcal{F}}$, and the only random variation in the above construction comes from the randomization protocol. The sample space $\Omega_{\gamma^*}$ of the uniform probability space $\Lambda_{\gamma^*}$, on which the random treatment status vector $\tilde{D}_{\gamma^*}$ is defined, is given by

$$\Omega_{\gamma^*} = \left( \bigtimes_{c=0}^{4} \mathbb{U}_c(\delta_c^*) \right) \times \Omega_{Q, V_{\gamma^*}}, \tag{4.5}$$

where $\mathbb{U}(\delta_0^*, \ldots, \delta_4^*) \equiv \times_{c=0}^{4} \mathbb{U}_c(\delta_c^*)$ is the Cartesian product of the sets of admissible partitions of $\mathcal{S}_c$ (in the initial step of the protocol) across all cohorts $c \in \{0, \ldots, 4\}$, and $\Omega_{Q, V_{\gamma^*}}$ is the Cartesian product of the sample spaces for all other random variables, characterizing the outcomes $Q_c$ of the coin flips and vectors of variables $V_{i,c}$ used for determining which children of working mothers are transferred from the treatment to control group in the last step across all cohorts $c$

[35] These tests are not strictly exact because our model simplifies the actual randomization procedure and can at best be considered a useful approximation of the true model of the protocol.

[36] Since our randomization tests follow the standard Fisherian framework, they are conditional tests that exploit random variation in the treatment status but fix the other observed data. See Lehmann (1993).

[37] Note that $\Xi$ is a sharp identified set because we follow the Fisherian framework where the observed outcomes and baseline covariates in our sample are treated as fixed.

$\in \{0, \ldots, 4\}$, used in the randomization protocol parameterized by $\gamma^*$.[38] Using this notation we establish the following proposition:

PROPOSITION 4.1. *Under the model of the randomization protocol in Section 3, the hypothesis test that rejects the sharp null hypothesis whenever $p_w(D) \leq \alpha$ controls the Type I error rate at level $\alpha$ for any $\alpha \in (0, 1)$.*

**Proof.** Let $p_{\gamma^*}(D) \equiv \mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_{\gamma^*}) \geq T(D)\}$ for all $\gamma^* \in \Xi$, let $p_w(D) \equiv \sup_{\gamma^* \in \Xi} p_{\gamma^*}(D)$ represent the worst-case $p$-value, and let $\psi_\alpha(D) \equiv \mathbb{I}\{p_w(D) \leq \alpha\}$ be the test for a given $D$, a realization of the random treatment status vector defined on the probability space $\Lambda_\gamma$, where $\gamma$ is the true value of the model parameter. Since $p_\gamma(D) \leq p_w(D)$ by construction, it follows that $\mathbb{E}_{\Lambda_\gamma}[\psi_\alpha(D)] = \mathbb{E}_{\Lambda_\gamma}[\mathbb{I}\{p_w(D) \leq \alpha\}] \leq \mathbb{E}_{\Lambda_\gamma}[\mathbb{I}\{p_\gamma(D) \leq \alpha\}] = \mathbb{P}_{\Lambda_\gamma}\{p_\gamma(D) \leq \alpha\} \leq \alpha$ under $\mathcal{H}_\mathcal{F}$ for any $\alpha \in (0, 1)$. □

This proof is an extension of the simple standard argument used to show the finite-sample validity of randomization tests (see Lehmann and Romano, 2005). The above proposition can be equivalently stated in terms of a critical value for the test statistic, as in Heckman et al. (2020).

Although it would be ideal to compute the exact value of $p_w(D)$, it is computationally not feasible. As is common practice in computing permutation and randomization $p$-values (see Lehmann and Romano, 2005), we resort to stochastic approximations. Even so, there are two challenges in estimating the worst-case $p$-value. First, approximating the probability $\mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_{\gamma^*}) \geq T(D)\}$ for a given value $\gamma^* \in \Xi$ is computationally demanding. Second, estimating $p_w(D)$ based on such tail probability estimates for a finite number of points on $\Xi$ is also challenging. We tackle these two challenges sequentially and discuss how we handle some forms of stochastic approximation errors.

*4.3.1. Approximating tail probabilities of randomization distributions.* The first challenge is to approximate $\mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_{\gamma^*}) \geq T(D)\}$ for a given value $\gamma^*$ in the partially identified set, i.e., for $\gamma^* = (\delta_0^*, \ldots, \delta_4^*, \eta_{0,1}^*, \eta_2^*, \eta_3^*, \eta_4^*) \in \Xi$. Our approach is to break up the sample space of $\Lambda_{\gamma^*}$ into two parts, compute the tail probability (measuring how extreme the observed test statistic is in its randomization null distribution) for each of these two parts, and then use the law of total probability and Monte Carlo methods to get the desired final result. To do so, we introduce additional notation. Let $\delta_c^\dagger$ be the lower bound of the partially identified set for the true value of the satisficing threshold $\delta_c$ for $c \in \{0, \ldots, 4\}$. Then, for any given value $\delta_c^* \geq \delta_c^\dagger$, observe that

$$\mathbb{U}_c(\delta_c^*) = \mathcal{X}_c \cup \mathcal{Y}_c(\delta_c^*), \tag{4.6}$$

where

$$\mathcal{X}_c = \{(\mathcal{A}, \mathcal{B}) \in \mathbb{U}_c(\infty) : \tau_c^2(\mathcal{A}, \mathcal{B}) \leq \delta_c^\dagger\}, \tag{4.7}$$

and

$$\mathcal{Y}_c(\delta_c^*) = \{(\mathcal{A}, \mathcal{B}) \in \mathbb{U}_c(\infty) : \delta_c^\dagger < \tau_c^2(\mathcal{A}, \mathcal{B}) \leq \delta_c^*\}, \tag{4.8}$$

for all $c \in \{0, \ldots, 4\}$. Thus, we can use $\mathbb{U}_c(\infty)$, which is the set with an infinite satisficing threshold such that all allowed partitions of $\mathcal{S}_c$ are satisfactory, to construct $\mathcal{X}_c$, $\mathcal{Y}_c(\delta_c^*)$, and $\mathbb{U}_c(\delta_c^*)$. The set $\mathcal{X}_c$ has elements with Hotelling statistics below the lower bound $\delta_c^\dagger$ of the partially

[38] Specifically, $\Omega_{Q,V_{\gamma^*}} = \{0, 1\}^5 \times \left(\times_{c \in \{(0,1),2,3,4\}} \times_{m=1}^{M_c} \{v \in \{0, 1\}^m : ||v||_1 = \min(\eta_c^*, m)\}\right)$, where $M_{0,1} = \sum_{i \in \mathcal{S}_0 \cup \mathcal{S}_1} M_i$ and $M_c = \sum_{i \in \mathcal{S}_c} M_i$ for all $c \in \{2, 3, 4\}$.

identified set for the satisficing threshold. The other set $\mathcal{Y}_c(\delta_c^*) = \mathbb{U}_c(\delta_c^*) \setminus \mathcal{X}_c$ has elements with Hotelling statistics between $\delta_c^\dagger$ and $\delta_c^*$. Let $\widetilde{\Omega}^{\mathcal{X}} = \times_{c=0}^4 \mathcal{X}_c$ be the Cartesian product of the sets $\mathcal{X}_c$ across cohorts, and let $\widetilde{\Omega}^{\mathcal{Y}}_{\gamma^*} = \mathbb{U}(\delta_0^*, \ldots, \delta_4^*) \setminus \widetilde{\Omega}^{\mathcal{X}} = \{\times_{c=0}^4 \mathbb{U}_c(\delta_c^*)\} \setminus \widetilde{\Omega}^{\mathcal{X}}$. Both $\widetilde{\Omega}^{\mathcal{X}}$ and $\widetilde{\Omega}^{\mathcal{Y}}_{\gamma^*}$ can be constructed using $\mathbb{U}(\infty, \ldots, \infty)$ by discarding elements in their respective complements. Since the sets $\mathcal{X}_c$ do not depend on the values $\delta_c^*$, the set $\widetilde{\Omega}^{\mathcal{X}}$ remains constant. Notice that

$$\Omega_{\gamma^*} = (\widetilde{\Omega}^{\mathcal{X}} \cup \widetilde{\Omega}^{\mathcal{Y}}_{\gamma^*}) \times \Omega_{Q,V_{\gamma^*}} = (\widetilde{\Omega}^{\mathcal{X}} \times \Omega_{Q,V_{\gamma^*}}) \cup (\widetilde{\Omega}^{\mathcal{Y}}_{\gamma^*} \times \Omega_{Q,V_{\gamma^*}}). \tag{4.9}$$

Let $\Lambda^{\mathcal{X}}_{\gamma^*}$ and $\Lambda^{\mathcal{Y}}_{\gamma^*}$ be the uniform probability spaces over the sample spaces $\Omega^{\mathcal{X}}_{\gamma^*} \equiv \widetilde{\Omega}^{\mathcal{X}} \times \Omega_{Q,V_{\gamma^*}}$ and $\Omega^{\mathcal{Y}}_{\gamma^*} \equiv \widetilde{\Omega}^{\mathcal{Y}}_{\gamma^*} \times \Omega_{Q,V_{\gamma^*}}$, respectively. In addition, let

$$x(\gamma^*) \equiv \frac{|\,\Omega^{\mathcal{X}}_{\gamma^*}\,|}{|\,\Omega_{\gamma^*}\,|} = \frac{|\,\widetilde{\Omega}^{\mathcal{X}} \times \Omega_{Q,V_{\gamma^*}}\,|}{|\,\Omega_{\gamma^*}\,|} = \frac{|\,\widetilde{\Omega}^{\mathcal{X}}\,| \cdot |\,\Omega_{Q,V_{\gamma^*}}\,|}{|\,\widetilde{\Omega}^{\mathcal{X}} \cup \widetilde{\Omega}^{\mathcal{Y}}_{\gamma^*}\,| \cdot |\,\Omega_{Q,V_{\gamma^*}}\,|} = \frac{|\,\widetilde{\Omega}^{\mathcal{X}}\,|}{|\,\widetilde{\Omega}^{\mathcal{X}} \cup \widetilde{\Omega}^{\mathcal{Y}}_{\gamma^*}\,|}, \tag{4.10}$$

which is the proportion of elements in the sample space $\Omega_{\gamma^*}$ belonging to $\Omega^{\mathcal{X}}_{\gamma^*}$. Note that the last equality above implies that $x(\gamma^*)$ can be simply computed with the sets $\widetilde{\Omega}^{\mathcal{X}}$ and $\widetilde{\Omega}^{\mathcal{Y}}_{\gamma^*}$ constructed using $\mathbb{U}(\infty, \ldots, \infty)$.[39] Then, by the law of total probability, we have that

$$\begin{aligned}\mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_{\gamma^*}) \geq T(D)\} = {} & x(\gamma^*) \cdot \mathbb{P}_{\Lambda^{\mathcal{X}}_{\gamma^*}}\{T(\tilde{D}^{\mathcal{X}}_{\gamma^*}) \geq T(D)\} \\ & + y(\gamma^*) \cdot \mathbb{P}_{\Lambda^{\mathcal{Y}}_{\gamma^*}}\{T(\tilde{D}^{\mathcal{Y}}_{\gamma^*}) \geq T(D)\},\end{aligned} \tag{4.11}$$

where $\tilde{D}^{\mathcal{X}}_{\gamma^*}$ and $\tilde{D}^{\mathcal{Y}}_{\gamma^*}$ represent random treatment status vectors defined on the probability spaces $\Lambda^{\mathcal{X}}_{\gamma^*}$ and $\Lambda^{\mathcal{Y}}_{\gamma^*}$, respectively, and $y(\gamma^*) = 1 - x(\gamma^*)$. Since the sample spaces in the model are large, we use Monte Carlo draws from the probability spaces through rejection sampling to stochastically approximate the tail probability $\mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_{\gamma^*}) \geq T(D)\}$.[40,41] Our approach provides a feasible way to estimate $\mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_{\gamma^*}) \geq T(D)\}$ for points $\gamma^*$ in $\Xi$ efficiently using rejection sampling.

*4.3.2. Estimating and bounding the worst-case tail probability.* The second challenge is to estimate or bound the worst-case tail probability. Taking the supremum of tail probabilities over points in the set $\Xi$ may seem intractable, since $\Xi$ is the Cartesian product of a finite set and a noncompact set.[42] However, we exploit the fact that $\mathbb{U}_c(\nu) = \mathbb{U}_c(\infty)$ for all $\nu \geq \Delta_c$, where $\Delta_c = \max\{\tau_c^2(\mathcal{A}, \mathcal{B}) : (\mathcal{A}, \mathcal{B}) \in \mathbb{U}_c(\infty)\}$, since $\mathbb{U}_c(\infty)$ is a finite set, for all $c \in \{0, \cdots, 4\}$. Let $\Xi^\circ$

[39] We use 500,000 Monte Carlo draws from $\mathbb{U}(\infty, \ldots, \infty) = \times_{c=0}^4 \mathbb{U}_c(\infty)$, a very large set, to approximate $x(\gamma^*)$.

[40] We use 400 Monte Carlo draws from $\Lambda^{\mathcal{X}}_{\gamma^*}$ to approximate $\mathbb{P}_{\Lambda^{\mathcal{X}}_{\gamma^*}}\{T(\tilde{D}^{\mathcal{X}}_{\gamma^*}) \geq T(D)\}$. This is effectively importance sampling. In addition, we use 2,600 Monte Carlo draws from $\Lambda^{\mathcal{Y}}_{\gamma^\infty}$, where $\gamma^\infty = (\infty, \ldots, \infty, \eta^*_{0,1}, \eta^*_2, \eta^*_3, \eta^*_4)$, and use rejection sampling to draw random samples from $\Lambda^{\mathcal{Y}}_{\gamma^*}$ for approximating $\mathbb{P}_{\Lambda^{\mathcal{Y}}_{\gamma^*}}\{T(\tilde{D}^{\mathcal{Y}}_{\gamma^*}) \geq T(D)\}$. It takes much longer to compute these tail probabilities than to compute $x(\gamma^*)$. Limited computational power restricted the number of Monte Carlo draws.

[41] Since the randomly sampled treatment status vectors are i.i.d. and uniformly distributed on corresponding sample spaces, for a given $\gamma^*$ the associated $p$-value stochastic approximations can be used to construct valid tests. For details, see section 4 of Romano (1989), section 3.2 of Romano and Wolf (2005), or section 15.2.1 of Lehmann and Romano (2005). Although this holds when $\gamma^*$ is taken as given, our main object of interest is the worst-case $p$-value in equation (4.4). Since it is infeasible to compute a $p$-value for each $\gamma^* \in \Xi$, we also resort to stochastic approximations of the supremum in equation (4.4). In Section 4.3.2, we discuss how we account for uncertainty in the stochastic approximation of the worst-case $p$-value.

[42] Specifically, $\Xi = \times_{c=0}^4 [\delta_c^\dagger, \infty) \times \vartheta_{0,1}^\eta \times \times_{c=2}^4 \vartheta_c^\eta$, where $\delta_c^\dagger$ is the lower bound for the satisficing threshold $\delta_c$, and $\vartheta_c^\eta$ is the finite partially identified set for the capacity constraint $\eta_c$.

be the compact subset of $\Xi$ given by

$$\Xi^{\circ} = \{\tilde{\gamma} \equiv (\tilde{\delta}_0, \ldots, \tilde{\delta}_4, \tilde{\eta}_{0,1}, \tilde{\eta}_2, \tilde{\eta}_3, \tilde{\eta}_4) \in \Xi : \delta_c^{\dagger} \leq \tilde{\delta}_c \leq \Delta_c \, \forall c\}. \tag{4.12}$$

It then follows that

$$p_w(D) \equiv \sup_{\gamma^* \in \Xi} \mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_{\gamma^*}) \geq T(D)\} = \max_{\gamma^* \in \Xi^{\circ}} \mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_{\gamma^*}) \geq T(D)\}. \tag{4.13}$$

Thus, it suffices to estimate the worst-case tail probability over the set $\Xi^{\circ}$, which is compact.[43] We use stochastic approximations for this purpose as well. It is computationally infeasible to compute a $p$-value for each of the points in the set $\Xi^{\circ}$ and take the maximum of those $p$-values. To deal with this challenge, we first write $\Xi^{\circ} = \bigcup_{l=1}^{L} \Xi_l^{\circ}$, where $\Xi_1^{\circ}, \ldots, \Xi_L^{\circ}$ are disjoint hyper-rectangles that form a partition of the set $\Xi^{\circ}$. In our application, $L = 20$, and each hyper-rectangle represents the partially identified set for $(\delta_0, \ldots, \delta_4)$ at fixed values of $(\eta_{0,1}, \eta_2, \eta_3, \eta_4)$.[44] Then, note that

$$p_w(D) = \max\{p_w^1(D), \ldots, p_w^L(D)\}, \tag{4.14}$$

where

$$p_w^l(D) = \max_{\gamma^* \in \Xi_l^{\circ}} \mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_{\gamma^*}) \geq T(D)\}, \tag{4.15}$$

for $l \in \{1, \ldots, L\}$. We approximate $p_w^l(D)$ for each $l \in \{1, \ldots, L\}$ using the $p$-values $p_{(1)}^l, \ldots, p_{(S)}^l$ arranged in descending order for $S = 900$ uniformly sampled random points on the set $\Xi_l^{\circ}$.[45]

We estimate $p_w^l(D)$ for each $l \in \{1, \ldots, L\}$ using the maximum order statistic $\tilde{p}_M^l$:

$$\tilde{p}_M^l = \max_{1 \leq s \leq S} \; p_{(s)}^l = p_{(1)}^l, \tag{4.16}$$

which converges almost surely to $p_w^l(D)$ as $S \to \infty$. However, this estimate may have stochastic approximation error. One way to deal with stochastic approximation-related uncertainty in $\tilde{p}_M^l$ is by constructing a confidence band for $p_w^l(D)$. To do so, we construct an upper bound based on de Haan's (1981) 90% asymptotic confidence band for the true maximum using the $S$ randomly sampled $p$-values. The upper confidence bound $\tilde{p}_{dH}^l$ is given by

$$\tilde{p}_{dH}^l = p_{(1)}^l + (p_{(1)}^l - p_{(2)}^l) \cdot K_{dH}^l, \tag{4.17}$$

[43] In fact, we can further simplify the worst-case tail probability. Let $\Gamma_c = \{\tau_c^2(\mathcal{A}, \mathcal{B}) : (\mathcal{A}, \mathcal{B}) \in \mathbb{U}_c(\infty)\}$, which is a finite set, for all $c \in \{0, \ldots, 4\}$, and let $\Xi^{\Gamma} = \{\tilde{\gamma} \equiv (\tilde{\delta}_0, \ldots, \tilde{\delta}_4, \tilde{\eta}_{0,1}, \tilde{\eta}_2, \tilde{\eta}_3, \tilde{\eta}_4) \in \Xi^{\circ} : \tilde{\delta}_c \in \Gamma_c \, \forall c\}$, which is also a finite set. Then, we have that $p_w(D) = \max_{\gamma^* \in \Xi^{\Gamma}} \mathbb{P}_{\Lambda_{\gamma^*}}\{T(\tilde{D}_{\gamma^*}) \geq T(D)\}$. However, even though the set $\Xi^{\Gamma}$ is finite, its size is too large in practice, making stochastic approximations still necessary.

[44] Note that in our application, $\eta_{0,1}$, $\eta_2$, and $\eta_3$ are point-identified while $\eta_4$ is partially identified to be in the set $\{0, \ldots, 4\}$. Thus, $(\eta_{0,1}, \eta_2, \eta_3, \eta_4)$ has 5 possible values. In addition, since we do not know the mother's working status at baseline for a subject in wave 0 and the gender of a subject in wave 1 (both of whom are among the 5 participants who dropped out of the study for extraneous reasons), there are 4 possible configurations of the two missing binary variables. Thus, in total there are $L = 5 \times 4 = 20$ hyper-rectangles that make up $\Xi^{\circ}$.

[45] To ensure that we are covering $\Xi^{\circ}$ and its edges well when sampling the random points, we use a normalization. We use the distribution $F_{\tau_c^2}$ of Hotelling statistics on $\mathbb{U}_c(\infty)$ to normalize $\delta_c$ so that $F_{\tau_c^2}(\delta_c) \in [F_{\tau_c^2}(\delta_c^{\dagger}), 1]$, a compact set, for all $c \in \{0, \ldots, 4\}$. Thus, $\gamma$ and $\Xi_l^{\circ}$ are monotonically transformed accordingly in practice. We can do this because $\mathbb{U}_c(\infty)$ is a finite set and $\mathbb{U}_c(\delta_c) \equiv \{(\mathcal{A}, \mathcal{B}) : \mathcal{A} \subset \mathcal{S}_c, \; \mathcal{B} = \mathcal{S}_c \setminus \mathcal{A}, \; |\mathcal{A}| = \lceil |\mathcal{S}_c|/2 \rceil, \; \tau_c^2(\mathcal{A}, \mathcal{B}) \leq \delta_c\}$ is equivalent to the set $\{(\mathcal{A}, \mathcal{B}) : \mathcal{A} \subset \mathcal{S}_c, \; \mathcal{B} = \mathcal{S}_c \setminus \mathcal{A}, \; |\mathcal{A}| = \lceil |\mathcal{S}_c|/2 \rceil, \; F_{\tau_c^2}(\tau_c^2(\mathcal{A}, \mathcal{B})) \leq F_{\tau_c^2}(\delta_c)\}$.

where $K^l_{dH}$ is a factor provided by de Haan (1981) for the 90% asymptotic confidence bound.[46] Thus, the 90% confidence interval for $p^l_w(D)$ is given by $[\tilde{p}^l_M, \tilde{p}^l_{dH}]$. Finally, the true worst-case $p$-value $p_w(D)$ can be approximated by the *worst-case maximum (max.) p-value* $\tilde{p}_M$ given by

$$\tilde{p}_M = \max\{\tilde{p}^1_M, \ldots, \tilde{p}^L_M\}, \tag{4.18}$$

and its upper confidence bound is given by the *worst-case de Haan p-value* $\tilde{p}_{dH}$ as follows:

$$\tilde{p}_{dH} = \max\{\tilde{p}^1_{dH}, \ldots, \tilde{p}^L_{dH}\}, \tag{4.19}$$

which provides at least 90% coverage as $S \to \infty$. Of course, these stochastic approximations affect the exact finite-sample validity of the resulting hypothesis tests, but the validity of these approximations can be arbitrarily increased with adequate additional computational power. This is an issue common to most resampling methods in statistics (see Lehmann and Romano, 2005).

In the previous discussion, the test statistic $T(\cdot)$ used to compute the worst-case tail probability is left general. There is reason to suspect that the choice of the test statistic matters, as shown for permutation tests by Chung and Romano (2013; 2016). Wu and Ding (2020) show that using studentized test statistics in certain randomization tests can control type I error asymptotically under certain weak null hypotheses while preserving finite-sample validity under sharp null hypotheses. Their theory ignores covariates and is limited to completely randomized factorial experiments and stratified or clustered experiments. However, they conjecture that "the strategy [of using studentized test statistics to make randomization tests asymptotically robust under weak null hypotheses while retaining their finite-sample validity under sharp null hypotheses may also be] applicable for experiments with general treatment assignment mechanisms" (Wu and Ding, 2020). While we do not attempt to prove or disprove their conjecture in the Perry experimental setting, we take it seriously given their results for certain randomization tests along with Chung and Romano's (2013; 2016) results for permutation tests. Thus, we provide worst-case $p$-values using both the nonstudentized and studentized test statistics.

*4.3.3. Multiple testing.* Since $\mathbb{P}_{\Lambda_\gamma}\{p_w(D) \leq \alpha\} \leq \alpha$ under $\mathcal{H}_\mathcal{F}$ for any $\alpha \in (0, 1)$ by Proposition 4.1, Holm (1979) tests of multiple hypotheses based on the worst-case $p$-values would also have finite-sample validity. Multiplicity-adjusted $p$-values can be computed as follows. Let $\rho_{(1)}, \ldots, \rho_{(K)}$ be the associated single worst-case $p$-values arranged in ascending order. Then, the Holm stepdown $p$-values adjusted for multiple testing are given by $\varrho_{(k)} = \max_{j \leq k} \min(1, (K - j + 1)\rho_{(j)})$ for $k \in \{1, \ldots, K\}$. However, these adjusted $p$-values can be even more conservative because they assume least favourable dependence structure between the single worst-case $p$-values (Romano et al., 2010), making this the "worst-case" of the "worst-case." However, slightly less conservative multiple hypothesis tests are available in the literature (see Romano and Wolf, 2005; Romano and Shaikh, 2010). Since it is unclear how much improvement in terms of power they provide relative to Holm tests in our context, we do not discuss the more computationally involved stepdown procedures in this paper.

[46] Specifically, $K^l_{dH} = \left[0.9^{\upsilon^l_{dH}} - 1\right]^{-1}$, where $\upsilon^l_{dH} = -\ln[(p^l_{(3)} - p^l_{(\sqrt{S})})/(p^l_{(2)} - p^l_{(3)})]/\ln(\sqrt{S})$, based on de Haan's (1981) result. In the context of estimating the minimum of a function over a compact set using order statistics, de Haan (1981) proposes construction of a confidence band for the minimum. We apply this result without loss of generality in our context (estimation of the maximum rather than the minimum).

### *4.4. Comparing methods for inference with imperfect randomization*

Our approach complements that of Heckman et al. (2020), who improve on the methodology of Heckman et al. (2010a) by (i) exploiting a symmetry generated by the Perry randomization protocol, (ii) using finite-sample inference that accounts for imperfect randomization, and (iii) making transfers in the fifth step of the randomization protocol depend on a binary variable[47] indicating whether the mother is available for home visits, assuming programme infrastructure is available to support it. We also exploit the symmetry: $Q_c$ represents the result of a fair coin flip to determine which of the two initially undesignated groups becomes the intended treatment group. However, we model other features of the protocol differently.

Heckman et al. (2020) model the reassignment of children of some working women by introducing a partially observed binary variable $U_i$ that equals 1 if the mother of participant $i$ was unavailable for home visits and 0 otherwise. It is known only for children of nonworking mothers, for whom $U_i = 0$, and for the children of working mothers in the final treatment group, who also have $U_i = 0$. For children of working mothers in the control group, $U_i$ is not known and could be either 0 or 1. To deal with this difficulty, Heckman et al. (2020) construct two permutation tests. The first test sets $U_i$ to 0 for all children of working mothers in the final control group and conducts a generalized permutation test accordingly. The second test: (i) samples a vector of $U_i$ from the space of possibilities for $U_i$; (ii) conducts a generalized permutation test given the sampled vector of $U_i$ and obtains the corresponding permutation $p$-value; and (iii) repeats steps (i) and (ii) until the space of possibilities is exhausted. It then takes the maximum $p$-value among the computed $p$-values. Our worst-case inferential methods are similar in spirit. However, there are three key differences between our approach and theirs.

First, Heckman et al. (2020) interpret $U_i$ as a fixed *trait* of mothers regardless of the (random) circumstances facing programme administrators. However, whether or not a working mother and her child are visited at home (through special arrangements, e.g., on a weekend) depends, at least in part, on the availability and capacity constraints of the Perry staff. While $U_i = 0$ for nonworking mothers in both papers, we do not view $U_i$ as a fixed binary trait of working mothers. Consistent with our review of the randomization protocol, we assume that children of working mothers are able to participate in the programme if special arrangements, such as weekend home visits, are made for them. In our model, there are capacity constraints for making special arrangements, so only a limited number of slots are available.[48] In their model, if $U_i = 1$ for a working mother, her child would always be placed in the control group, because she would not accept any special accommodations even if provided by the Perry staff. Unlike the $V_{i,c}$ variable that determines post-randomization transfers in our model, the $U_i$ characteristic in their model is allowed to be related to potential outcomes, but this is a consequence of its interpretation as a fixed trait of mothers independent of the capacities of programme administrators.

Second, their procedure assumes that "some participants were exchanged between the treatment and control groups in order to balance gender and socioeconomic status score while keeping Stanford–Binet IQ score roughly constant."[49] However, as shown in Appendix B, Perry data from wave 4 reveal that the exchanges were not necessarily between consecutively ranked IQ

[47] It is only partially observed in their model.

[48] Our model is limited in the sense that it does not allow for heterogeneity among working mothers in their availability for special arrangements. We assume that the Perry administrators choose with equal probability which working mothers get special arrangements.

[49] This is Step 4′ in their paper. Accordingly, their tests involve "permuting treatment status among those families with the same observed and unobserved characteristics (defined by the characteristics of the eldest child in the case of families with multiple children)." In practice, they discretize SES into a binary indicator of above-median SES.

pairs. Our approach accommodates this feature while also making more explicit the notion of balance.

Third, on a more minor note, we incorporate the five children (out of the original 128) who dropped out of the study due to extraneous reasons, since those five children were also a part of the initial randomization protocol. Our approach can also more readily be applied than that of Heckman et al. (2020) to a variety of compromised experiments, including many discussed by Bruhn and McKenzie (2009). We next demonstrate that there are important differences between inferences obtained from our procedure and theirs.

## 5. REANALYSIS OF HECKMAN ET AL. (2020)

This section uses the methods developed in this paper to reconsider the conclusions reached by Heckman et al. (2020) on the Perry participants through to age 40. We first list our estimators of treatment effects. Using the corresponding test statistics, we then apply our worst-case inferential methods to reanalyse the results in Heckman et al. (2020).

### *5.1. Estimators and test statistics for hypothesis testing*

A variety of test statistics and estimators can be used in our approach and that of Heckman et al. (2020). Our empirical work focuses conventional ones often used in practice. Let $D_i$ represent the treatment status of participant $i$, and let $Z_i$ be the vector of baseline variables.[50] In addition, let $Y_i$ denote the observed outcome of interest of participant $i$ in a relevant subsample $\mathcal{P}$ containing $N_{\mathcal{P}} = |\mathcal{P}|$ participants, and let $Y_i^d$ be the counterfactual outcome of participant $i$ when his or her treatment status $D_i$ is fixed at $d \in \{0, 1\}$. In switching regression notation (Quandt, 1958; 1972),

$$Y_i = D_i\, Y_i^1 + (1 - D_i)\, Y_i^0. \tag{5.1}$$

The average treatment effect $\bar{\tau}$ in the subsample $\mathcal{P}$ is given by

$$\bar{\tau} = \frac{1}{N_{\mathcal{P}}} \sum_{i \in \mathcal{P}} \left( Y_i^1 - Y_i^0 \right), \tag{5.2}$$

and is conventionally estimated by a difference-in-means (DIM) estimator that takes raw mean differences between nonattrited treated and control observations. However, the randomization procedures used in Perry and other similar experiments only justify conditional independence: $(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i \mid Z_i$. Exploiting this property and controlling for $Z_i$ in a regression of $Y_i$ on $D_i$ and $Z_i$ using complete case observations, we obtain the ordinary least squares (OLS) estimator.[51] It would be desirable to relax linearity, but the Perry sample size makes this impractical.

All of these estimators assume that nonresponse is determined at random or at random conditional on observed covariates. Let $R_i$ be an indicator of whether $Y_i$ is missing. It could depend on the treatment status $D_i$ and the pre-programme covariates $Z_i$. The augmented inverse probability weighting (AIPW) estimator allows for this possibility by using the weaker assumption that

[50] In the Perry context, it consists of the four pre-programme covariates used during the randomization phase, i.e., Stanford–Binet IQ, index of SES, gender, and mother's working status.

[51] Both OLS and DIM estimators can be studentized using their cluster-robust asymptotic standard errors, allowing for correlation between error terms of the participant-siblings in the Perry experiment.

$Y_i \perp\!\!\!\perp R_i \mid D_i,\ Z_i$. The AIPW estimator of the treatment effect is

$$\hat{\Pi}_{\text{AIPW}} = \frac{1}{N_{\mathcal{P}}} \sum_{i \in \mathcal{P}} \left( \hat{\pi}_i^1 - \hat{\pi}_i^0 \right), \tag{5.3}$$

where

$$\hat{\pi}_i^d = \hat{Y}_i^d + \frac{\mathbb{I}\{R_i = 1,\ D_i = d\}}{\hat{\lambda}_i^d\, \hat{\phi}_i^d} \left( Y_i^d - \hat{Y}_i^d \right). \tag{5.4}$$

In this expression, $\hat{Y}_i^d$ is the OLS (projection) estimator of the conditional expectation $\mathbb{E}[Y_i \mid Z_i, D_i = d, R_i = 1]$ for $d \in \{0, 1\}$, $\hat{\phi}_i^d$ is an estimator of $\Pr(D_i = d \mid Z_i)$, the $i$-th subject's propensity of being in experimental state $d$, and $\hat{\lambda}_i^d$ is an estimator of $\Pr(R_i = 1 \mid Z_i, D_i = d)$, which is the propensity of having a nonmissing outcome after fixing the treatment status $D_i$, for $d \in \{0, 1\}$. Propensity scores are often estimated using logits.[52] The AIPW estimator adjusts the outcome based on pre-programme covariates and corrects for covariate imbalance and various forms of nonresponse.[53] It has a double robustness property: the estimator is robust to misspecification of either the propensity score models or the models for counterfactual outcomes, but not both.[54] For this reason, the AIPW estimator is sometimes preferred over the DIM and OLS estimators.[55] We use the studentized version of the AIPW estimate as our main test statistic in our empirical analysis.[56]

We could use a local average treatment effect (LATE) estimator, and other standard estimation methods dealing with imperfect compliance, if we knew each observation's initial treatment status. However, in the Perry example, we do not know which members were transferred from the

[52] We estimate the propensity scores using a logit specification and the penalized maximum likelihood method of Greenland and Mansournia (2015), which circumvents the issue of separation in small samples.

[53] The AIPW estimator also assumes conditional independence of the counterfactual outcomes and the treatment status, i.e., $(Y_i^1, Y_i^0) \perp\!\!\!\perp D_i \mid Z_i$, which is valid because of the random assignment of the treatment status conditional on pre-programme variables. Note that the propensity score model used in the AIPW estimator is a direct consequence of the law of conditional probability: $\Pr(R_i = 1,\ D_i = d \mid Z_i) = \Pr(R_i = 1 \mid Z_i,\ D_i = d) \Pr(D_i = d \mid Z_i)$ for $d \in \{0, 1\}$. In the econometrics literature, the AIPW estimator is better known as a type of efficient influence function (EIF) estimator (Cattaneo, 2010). The estimator given by equation (5.3) can be studentized using the empirical sandwich standard error (Lunceford and Davidian, 2004). For studentization, we use a cluster-robust version of this asymptotic standard error, given by the following formula: $\frac{1}{N_{\mathcal{P}}}[\sum_{j \in J}(\sum_{i \in \mathcal{F}_j} \hat{\pi}_i^1 - \hat{\pi}_i^0 - \hat{\Pi}_{\text{AIPW}})^2]^{1/2}[|J|/(|J| - 1)]^{1/2}$, where $\mathcal{F}_j$ represents a cluster of participant-siblings in the set $J$ of clusters. Our studentized test statistics are based on the asymptotic standard error mainly for computational ease, but studentization based on the bootstrap standard error would be superior in theory.

[54] See Robins et al. (1994), Lunceford and Davidian (2004), and Kang and Schafer (2007). The double robustness property (consistency despite certain forms of misspecification) is easier to understand by rewriting equation (5.4) as follows: $\hat{\pi}_i^d = Y_i^d + (\hat{\lambda}_i^d\, \hat{\phi}_i^d)^{-1}(\mathbb{I}\{R_i = 1,\ D_i = d\} - \hat{\lambda}_i^d\, \hat{\phi}_i^d)(Y_i^d - \hat{Y}_i^d)$ for $d \in \{0, 1\}$. If the propensity score models or the the counterfactual outcome model are correctly specified, sample average of the whole second term (in the rewritten expression for $\hat{\pi}_i^d$) converges in probability to zero. Thus, the AIPW estimator remains consistent for the average treatment effect even if either the propensity score models or the counterfactual outcome models are misspecified.

[55] However, we present estimates from all of these procedures in the online appendices as a form of sensitivity analysis. The AIPW estimator can become unstable if both the propensity score models and the counterfactual outcome models are misspecified (Kang and Schafer, 2007). Thus, we do not solely rely on the AIPW estimator but use it in conjunction with the DIM and OLS estimators.

[56] Since AIPW clearly has an asymptotic justification, it is not strictly a small-sample procedure from an estimation perspective. Nevertheless, we can conduct inference using its finite-sample worst-case randomization null distribution using our design-based methods.

initial treatment group to the control group in the last step of the randomization protocol. Given this problem, we do not present LATE estimates.[57]

### 5.2. Empirical analysis

We first conduct a head-to-head comparison of Heckman et al.'s (2020) methods and ours using the same outcomes they analyse. Additionally, to compare the impact of using mean differences versus AIPW test statistics in the conventional inferential approaches and our design-based worst-case inference, we extend the outcomes they study and analyse data on violent crime.

Tables 2 and 3 report our reanalyses of Heckman et al. (2020), analysing each outcome one at a time using the doubly robust attrition-adjusted AIPW estimator. Tables 4 and 5 provide stepdown $p$-values for the outcomes based on multiple testing. Extended versions of these tables are presented in Online Appendices S3 to S9 using alternative test statistics for inference.[58]

In Tables 6 and 7, we reproduce Heckman et al.'s (2020) results and provide a side-by-side comparison of their inferences with our own. The most stringent (max-$U$) single $p$-values they report for the effects on the California Achievement Test (CAT) reading, arithmetic, language, mechanics, and spelling scores at age 14 in the male sample using the studentized DIM test statistic are 0.036, 0.086, 0.012, 0.023, and 0.012, respectively, which are lower than the asymptotic $p$-values we report in Table 2. After adjusting for multiple testing, their adjusted max-$U$ $p$-values are no more than 0.086, based on which they conclude that these effects are statistically significant. In contrast, using our approach, the worst-case maximum (single) $p$-values using studentized DIM test statistic are 0.144, 0.119, 0.069, 0.046, and 0.114, respectively. As shown in our Table 2, using the studentized AIPW test statistic, our worst-case maximum $p$-values are 0.325, 0.272, 0.176, 0.123, and 0.274, respectively,[59] implying that the effects on the CAT scores for males are not statistically significant. Of course, the stepdown $p$-values for these outcomes shown in Table 4 are also insignificant. Our inference for the female sample is qualitatively similar to theirs. As shown in Table 3, most of the block related to CAT scores for females is statistically significant at the 10% level. However, the multiplicity-adjusted stepdown worst-case de Haan $p$-values in Table 5 are 0.13 or larger.

Table 4 reports stepdown $p$-values for male outcomes. No estimated effect (after age 5) remains statistically significant at the 10% level after adjusting for multiple hypothesis testing using the worst-case maximum or worst-case de Haan $p$-values. However, in Table 5, which presents stepdown analysis of female outcomes, the treatment effects on post-programme outcomes (related to some CAT scores and educational outcomes) are statistically significant at the 10% level using our worst-case maximum $p$-value. Nevertheless, all of these effects on female outcomes, except for two (high school graduation and grade point average), disappear when worst-case de Haan $p$-values are used.

Tables 2 to 5 use the studentized AIPW test statistic for inference. Heckman et al. (2020) use the studentized DIM test statistic instead. Tables 6 and 7 compare their inferences with ours using

[57] In theory, we could bound the LATE estimate by considering all possible values for each observation's initial treatment status, and then we could use the LATE bound as a test statistic for inference. However, this is very demanding computationally and thus not feasible in practice.

[58] In these online appendices, for each outcome we include the conventional $p$-values (i.e., asymptotic, bootstrap, and permutation $p$-values) and design-based $p$-values (i.e., worst-case maximum and worst-case de Haan $p$-values) associated with each of the DIM, OLS, and AIPW estimators of treatment effects. We also include permutation and worst-case $p$-values based on both nonstudentized and studentized test statistics. In addition, we include stepdown versions of the worst-case $p$-values.

[59] The corresponding worst-case de Haan (single) $p$-values are 0.427, 0.343, 0.348, 0.236, and 0.459, respectively.

**Table 2.** Reanalysis of **Male Outcomes** in Heckman et al. (2020) using **Single Tests**.

| *Variable* | *Age* | *Untreated mean* | *Treated mean* | *AIPW estimate* | *Asymptotic p-value* | *Bootstrap p-value* | *Permutation p-value* | *Worst-case max. p* | *Worst-case de Haan p* |
|---|---|---|---|---|---|---|---|---|---|
| Stanford–Binet IQ | 4 | 83.077 | 94.909 | 8.988 | **0.0000** | **0.0000** | **0.0004** | **0.0052** | **0.0060** |
| Stanford–Binet IQ | 5 | 84.793 | 95.400 | 9.167 | **0.0000** | **0.0002** | **0.0004** | **0.0050** | **0.0059** |
| Stanford–Binet IQ | 6 | 85.821 | 91.485 | 3.056 | **0.0557** | **0.0512** | **0.0712** | **0.0838** | 0.2362 |
| Stanford–Binet IQ | 7 | 87.711 | 91.121 | 1.576 | 0.2040 | 0.2143 | 0.2104 | 0.1905 | 0.3701 |
| Stanford–Binet IQ | 8 | 89.054 | 88.333 | −3.829 | **0.0512** | **0.0719** | **0.0556** | 0.1519 | 0.1894 |
| Stanford–Binet IQ | 9 | 89.026 | 88.394 | −4.167 | **0.0398** | **0.0577** | **0.0472** | 0.1147 | 0.3800 |
| Stanford–Binet IQ | 10 | 86.026 | 83.697 | −4.722 | **0.0225** | **0.0412** | **0.0292** | **0.0678** | 0.1012 |
| CAT reading score | 14 | 9.000 | 13.926 | 1.815 | 0.2957 | 0.3221 | 0.3112 | 0.3253 | 0.4273 |
| CAT arithmetic score | 14 | 8.107 | 16.000 | 3.095 | 0.2410 | 0.2629 | 0.2608 | 0.2722 | 0.3434 |
| CAT language score | 14 | 6.536 | 14.333 | 5.029 | **0.0815** | **0.0995** | 0.1076 | 0.1764 | 0.3482 |
| CAT mechanics score | 14 | 6.964 | 15.556 | 5.979 | **0.0538** | **0.0638** | **0.0712** | 0.1234 | 0.2364 |
| CAT spelling score | 14 | 11.536 | 18.519 | 3.171 | 0.2652 | 0.2865 | 0.2600 | 0.2741 | 0.4587 |
| High school graduate | 19 | 0.513 | 0.485 | 0.015 | 0.4550 | 0.4540 | 0.4868 | 0.5651 | 0.7190 |
| Vocational training | 40 | 0.333 | 0.394 | 0.071 | 0.2762 | 0.2886 | 0.2932 | 0.3619 | 0.4378 |
| Highest grade completed | 19 | 11.282 | 11.364 | 0.087 | 0.3902 | 0.3901 | 0.4240 | 0.4583 | 0.6282 |
| Grade point average | 19 | 1.794 | 1.814 | −0.035 | 0.4366 | 0.4336 | 0.4328 | 0.5267 | 0.6983 |
| Total nonjuvenile arrests | 40 | 11.718 | 7.455 | −3.895 | **0.0461** | **0.0368** | **0.0668** | **0.0951** | 1.0000 |
| Total crime cost | 40 | 775.901 | 424.679 | −313.263 | 0.1376 | 0.1361 | 0.1764 | 0.2024 | 0.3880 |
| Total charges | 40 | 13.385 | 9.000 | −4.132 | **0.0678** | **0.0579** | **0.0920** | 0.1216 | 0.2598 |
| Nonvictimless charges | 40 | 3.077 | 1.485 | −1.444 | **0.0274** | **0.0238** | **0.0372** | **0.0856** | 0.2792 |
| Currently employed | 19 | 0.410 | 0.545 | 0.147 | 0.1263 | 0.1315 | 0.1292 | 0.2999 | 0.4666 |
| Unemployed last year | 19 | 0.128 | 0.242 | 0.102 | 0.1817 | 0.1827 | 0.2148 | 0.2861 | 0.6064 |
| Jobless months (past 2 yrs) | 19 | 3.816 | 5.281 | 1.367 | 0.2572 | 0.2501 | 0.2928 | 0.3243 | 1.0000 |
| Currently employed | 27 | 0.564 | 0.600 | 0.089 | 0.2156 | 0.2259 | 0.2452 | 0.3335 | 0.8446 |
| Unemployed last year | 27 | 0.308 | 0.242 | −0.081 | 0.2238 | 0.2190 | 0.2388 | 0.3488 | 0.5882 |
| Jobless months (past 2 yrs) | 27 | 8.795 | 5.133 | −3.868 | **0.0438** | **0.0430** | **0.0588** | 0.1115 | 0.3548 |
| Currently employed | 40 | 0.500 | 0.700 | 0.266 | **0.0089** | **0.0075** | **0.0204** | **0.0484** | **0.0971** |
| Unemployed last year | 40 | 0.462 | 0.364 | −0.143 | **0.0843** | **0.0957** | **0.0912** | 0.1695 | 0.5219 |
| Jobless months (past 2 yrs) | 40 | 10.750 | 7.233 | −4.758 | **0.0154** | **0.0200** | **0.0188** | **0.0650** | 0.1341 |

*Note:* This table reports *p*-values for single hypothesis tests of treatment effects on various outcomes of male participants at the given ages. The inferences are based on the studentized AIPW test statistic.

**Table 3.** Reanalysis of **Female Outcomes** in Heckman et al. (2020) using **Single Tests**.

| *Variable* | *Age* | *Untreated mean* | *Treated mean* | *AIPW estimate* | *Asymptotic p-value* | *Bootstrap p-value* | *Permutation p-value* | *Worst-case max. p* | *Worst-case de Haan p* |
|---|---|---|---|---|---|---|---|---|---|
| Stanford–Binet IQ | 4 | 83.692 | 96.360 | 13.425 | **0.0000** | **0.0000** | **0.0004** | **0.0053** | **0.0055** |
| Stanford–Binet IQ | 5 | 81.650 | 94.316 | 14.157 | **0.0008** | **0.0006** | **0.0064** | **0.0258** | **0.0503** |
| Stanford–Binet IQ | 6 | 87.160 | 90.913 | 5.271 | **0.0365** | **0.0281** | **0.0636** | **0.0799** | 0.4790 |
| Stanford–Binet IQ | 7 | 86.000 | 92.520 | 7.347 | **0.0313** | **0.0154** | **0.0564** | **0.0952** | 0.1950 |
| Stanford–Binet IQ | 8 | 83.600 | 87.840 | 4.669 | 0.1144 | **0.0896** | 0.1704 | 0.2080 | 0.2832 |
| Stanford–Binet IQ | 9 | 83.043 | 86.739 | 4.809 | **0.0633** | **0.0679** | 0.1128 | 0.1578 | 0.2873 |
| Stanford–Binet IQ | 10 | 81.789 | 86.750 | 6.480 | **0.0277** | **0.0323** | **0.0596** | 0.1840 | 0.4267 |
| CAT reading score | 14 | 8.444 | 16.500 | 7.345 | **0.0130** | **0.0128** | **0.0268** | **0.0561** | 0.1125 |
| CAT arithmetic score | 14 | 6.889 | 11.818 | 6.227 | **0.0102** | **0.0138** | **0.0284** | **0.0624** | **0.0731** |
| CAT language score | 14 | 7.833 | 19.455 | 11.923 | **0.0009** | **0.0013** | **0.0044** | **0.0168** | **0.0524** |
| CAT mechanics score | 14 | 8.833 | 20.636 | 12.425 | **0.0014** | **0.0015** | **0.0072** | **0.0211** | **0.0606** |
| CAT spelling score | 14 | 10.722 | 29.500 | 18.270 | **0.0017** | **0.0042** | **0.0064** | **0.0176** | **0.0254** |
| High school graduate | 19 | 0.231 | 0.840 | 0.570 | **0.0000** | **0.0000** | **0.0004** | **0.0054** | **0.0058** |
| Vocational training | 40 | 0.077 | 0.240 | 0.183 | **0.0286** | **0.0494** | **0.0420** | 0.1056 | 0.2630 |
| Highest grade completed | 19 | 10.750 | 11.760 | 1.202 | **0.0023** | **0.0106** | **0.0120** | **0.0345** | **0.0935** |
| Grade point average | 19 | 1.527 | 2.415 | 0.958 | **0.0000** | **0.0155** | **0.0004** | **0.0119** | **0.0164** |
| Total nonjuvenile arrests | 40 | 4.423 | 2.160 | −1.938 | **0.0657** | **0.0795** | **0.0880** | 0.1695 | 0.4424 |
| Total crime cost | 40 | 293.497 | 22.165 | −246.242 | 0.1475 | 0.1227 | 0.2436 | 0.2524 | 0.5508 |
| Total charges | 40 | 4.923 | 2.240 | −2.309 | **0.0439** | **0.0528** | **0.0580** | 0.1526 | 0.2963 |
| Nonvictimless charges | 40 | 0.308 | 0.040 | −0.249 | **0.0365** | **0.0263** | **0.0612** | **0.0906** | 0.2574 |
| Currently employed | 19 | 0.154 | 0.440 | 0.297 | **0.0054** | **0.0048** | **0.0152** | **0.0578** | 0.2187 |
| Unemployed last year | 19 | 0.577 | 0.240 | −0.354 | **0.0029** | **0.0033** | **0.0104** | **0.0341** | 0.1878 |
| Jobless months (past 2 yrs) | 19 | 10.421 | 5.217 | −4.197 | **0.0723** | 0.1386 | 0.1140 | 0.1886 | 0.4619 |
| Currently employed | 27 | 0.545 | 0.800 | 0.215 | **0.0471** | **0.0604** | **0.0648** | 0.1048 | 0.2521 |
| Unemployed last year | 27 | 0.542 | 0.250 | −0.269 | **0.0523** | **0.0457** | **0.0728** | 0.1721 | 1.0000 |
| Jobless months (past 2 yrs) | 27 | 10.455 | 6.240 | −1.298 | 0.3328 | 0.3449 | 0.2916 | 0.4821 | 0.7544 |
| Currently employed | 40 | 0.818 | 0.833 | −0.016 | 0.4536 | 0.4586 | 0.4912 | 0.6385 | 1.0000 |
| Unemployed last year | 40 | 0.409 | 0.160 | −0.194 | **0.0807** | 0.1079 | 0.1324 | 0.1892 | 0.3070 |
| Jobless months (past 2 yrs) | 40 | 5.045 | 4.000 | 0.057 | 0.4910 | 0.4927 | 0.4700 | 0.6326 | 1.0000 |

*Note:* This table reports *p*-values for single hypothesis tests of treatment effects on various outcomes of female participants at the given ages. The inferences are based on the studentized AIPW test statistic.

**Table 4.** Reanalysis of **Male Outcomes** in Heckman et al. (2020) using **Stepdown Tests**.

| *Variable* | *Age* | *Untreated mean* | *Treated mean* | *AIPW estimate* | *Asymptotic p-value* | *Bootstrap p-value* | *Permutation p-value* | *Worst-case max. p* | *Worst-case de Haan p* |
|---|---|---|---|---|---|---|---|---|---|
| Stanford–Binet IQ | 4 | 83.077 | 94.909 | 8.988 | **0.0001** | **0.0002** | **0.0028** | **0.0348** | **0.0415** |
| Stanford–Binet IQ | 5 | 84.793 | 95.400 | 9.167 | **0.0003** | **0.0012** | **0.0028** | **0.0348** | **0.0415** |
| Stanford–Binet IQ | 6 | 85.821 | 91.485 | 3.056 | 0.1593 | 0.2058 | 0.1888 | 0.3391 | 0.7574 |
| Stanford–Binet IQ | 7 | 87.711 | 91.121 | 1.576 | 0.2040 | 0.2143 | 0.2104 | 0.3440 | 0.7574 |
| Stanford–Binet IQ | 8 | 89.054 | 88.333 | −3.829 | 0.1593 | 0.2058 | 0.1888 | 0.3440 | 0.7574 |
| Stanford–Binet IQ | 9 | 89.026 | 88.394 | −4.167 | 0.1593 | 0.2058 | 0.1888 | 0.3440 | 0.7574 |
| Stanford–Binet IQ | 10 | 86.026 | 83.697 | −4.722 | 0.1126 | 0.2058 | 0.1460 | 0.3391 | 0.5062 |
| CAT reading score | 14 | 9.000 | 13.926 | 1.815 | 0.7229 | 0.7886 | 0.7800 | 0.8167 | 1.0000 |
| CAT arithmetic score | 14 | 8.107 | 16.000 | 3.095 | 0.7229 | 0.7886 | 0.7800 | 0.8167 | 1.0000 |
| CAT language score | 14 | 6.536 | 14.333 | 5.029 | 0.3260 | 0.3980 | 0.4304 | 0.7058 | 1.0000 |
| CAT mechanics score | 14 | 6.964 | 15.556 | 5.979 | 0.2690 | 0.3190 | 0.3560 | 0.6171 | 1.0000 |
| CAT spelling score | 14 | 11.536 | 18.519 | 3.171 | 0.7229 | 0.7886 | 0.7800 | 0.8167 | 1.0000 |
| High school graduate | 19 | 0.513 | 0.485 | 0.015 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Vocational training | 40 | 0.333 | 0.394 | 0.071 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Highest grade completed | 19 | 11.282 | 11.364 | 0.087 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Grade point average | 19 | 1.794 | 1.814 | −0.035 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Total nonjuvenile arrests | 40 | 11.718 | 7.455 | −3.895 | 0.1384 | 0.1103 | 0.2004 | 0.3424 | 1.0000 |
| Total crime cost | 40 | 775.901 | 424.679 | −313.263 | 0.1384 | 0.1361 | 0.2004 | 0.3424 | 1.0000 |
| Total charges | 40 | 13.385 | 9.000 | −4.132 | 0.1384 | 0.1158 | 0.2004 | 0.3424 | 1.0000 |
| Nonvictimless charges | 40 | 3.077 | 1.485 | −1.444 | 0.1096 | **0.0952** | 0.1488 | 0.3424 | 1.0000 |
| Currently employed | 19 | 0.410 | 0.545 | 0.147 | 0.3790 | 0.3946 | 0.3876 | 0.8583 | 1.0000 |
| Unemployed last year | 19 | 0.128 | 0.242 | 0.102 | 0.3790 | 0.3946 | 0.4296 | 0.8583 | 1.0000 |
| Jobless months (past 2 yrs) | 19 | 3.816 | 5.281 | 1.367 | 0.3790 | 0.3946 | 0.4296 | 0.8583 | 1.0000 |
| Currently employed | 27 | 0.564 | 0.600 | 0.089 | 0.4313 | 0.4380 | 0.4776 | 0.6670 | 1.0000 |
| Unemployed last year | 27 | 0.308 | 0.242 | −0.081 | 0.4313 | 0.4380 | 0.4776 | 0.6670 | 1.0000 |
| Jobless months (past 2 yrs) | 27 | 8.795 | 5.133 | −3.868 | 0.1313 | 0.1290 | 0.1764 | 0.3344 | 1.0000 |
| Currently employed | 40 | 0.500 | 0.700 | 0.266 | **0.0268** | **0.0225** | **0.0564** | 0.1451 | 0.2912 |
| Unemployed last year | 40 | 0.462 | 0.364 | −0.143 | **0.0843** | **0.0957** | **0.0912** | 0.1695 | 0.5219 |
| Jobless months (past 2 yrs) | 40 | 10.750 | 7.233 | −4.758 | **0.0309** | **0.0399** | **0.0564** | 0.1451 | 0.2912 |

*Note:* This table reports Holm stepdown *p*-values for multiple hypothesis tests of treatment effects on various outcomes of male participants at the given ages. The inferences are based on the studentized AIPW test statistic. The blocks used for multiple testing are indicated above using divider lines.

**Table 5.** Reanalysis of **Female Outcomes** in Heckman et al. (2020) using **Stepdown Tests**.

| *Variable* | *Age* | *Untreated mean* | *Treated mean* | *AIPW estimate* | *Asymptotic p-value* | *Bootstrap p-value* | *Permutation p-value* | *Worst-case max. p* | *Worst-case de Haan p* |
|---|---|---|---|---|---|---|---|---|---|
| Stanford–Binet IQ | 4 | 83.692 | 96.360 | 13.425 | **0.0000** | **0.0000** | **0.0028** | **0.0369** | **0.0387** |
| Stanford–Binet IQ | 5 | 81.650 | 94.316 | 14.157 | **0.0046** | **0.0035** | **0.0384** | 0.1550 | 0.3020 |
| Stanford–Binet IQ | 6 | 87.160 | 90.913 | 5.271 | 0.1387 | 0.1125 | 0.2820 | 0.3997 | 1.0000 |
| Stanford–Binet IQ | 7 | 86.000 | 92.520 | 7.347 | 0.1387 | **0.0771** | 0.2820 | 0.3997 | 0.9749 |
| Stanford–Binet IQ | 8 | 83.600 | 87.840 | 4.669 | 0.1387 | 0.1359 | 0.2820 | 0.4734 | 1.0000 |
| Stanford–Binet IQ | 9 | 83.043 | 86.739 | 4.809 | 0.1387 | 0.1359 | 0.2820 | 0.4734 | 1.0000 |
| Stanford–Binet IQ | 10 | 81.789 | 86.750 | 6.480 | 0.1387 | 0.1125 | 0.2820 | 0.4734 | 1.0000 |
| CAT reading score | 14 | 8.444 | 16.500 | 7.345 | **0.0205** | **0.0255** | **0.0536** | 0.1122 | 0.2097 |
| CAT arithmetic score | 14 | 6.889 | 11.818 | 6.227 | **0.0205** | **0.0255** | **0.0536** | 0.1122 | 0.2097 |
| CAT language score | 14 | 7.833 | 19.455 | 11.923 | **0.0043** | **0.0064** | **0.0220** | **0.0842** | 0.2097 |
| CAT mechanics score | 14 | 8.833 | 20.636 | 12.425 | **0.0056** | **0.0064** | **0.0256** | **0.0842** | 0.2097 |
| CAT spelling score | 14 | 10.722 | 29.500 | 18.270 | **0.0056** | **0.0127** | **0.0256** | **0.0842** | 0.1272 |
| High school graduate | 19 | 0.231 | 0.840 | 0.570 | **0.0000** | **0.0000** | **0.0016** | **0.0218** | **0.0232** |
| Vocational training | 40 | 0.077 | 0.240 | 0.183 | **0.0286** | **0.0494** | **0.0420** | 0.1056 | 0.2630 |
| Highest grade completed | 19 | 10.750 | 11.760 | 1.202 | **0.0046** | **0.0318** | **0.0240** | **0.0690** | 0.1871 |
| Grade point average | 19 | 1.527 | 2.415 | 0.958 | **0.0000** | **0.0318** | **0.0016** | **0.0357** | **0.0493** |
| Total nonjuvenile arrests | 40 | 4.423 | 2.160 | −1.938 | 0.1461 | 0.1589 | 0.2320 | 0.4578 | 1.0000 |
| Total crime cost | 40 | 293.497 | 22.165 | −246.242 | 0.1475 | 0.1589 | 0.2436 | 0.4578 | 1.0000 |
| Total charges | 40 | 4.923 | 2.240 | −2.309 | 0.1461 | 0.1585 | 0.2320 | 0.4578 | 1.0000 |
| Nonvictimless charges | 40 | 0.308 | 0.040 | −0.249 | 0.1461 | 0.1051 | 0.2320 | 0.3625 | 1.0000 |
| Currently employed | 19 | 0.154 | 0.440 | 0.297 | **0.0107** | **0.0099** | **0.0312** | 0.1155 | 0.5635 |
| Unemployed last year | 19 | 0.577 | 0.240 | −0.354 | **0.0088** | **0.0099** | **0.0312** | 0.1022 | 0.5635 |
| Jobless months (past 2 yrs) | 19 | 10.421 | 5.217 | −4.197 | **0.0723** | 0.1386 | 0.1140 | 0.1886 | 0.5635 |
| Currently employed | 27 | 0.545 | 0.800 | 0.215 | 0.1412 | 0.1371 | 0.1944 | 0.3143 | 0.7562 |
| Unemployed last year | 27 | 0.542 | 0.250 | −0.269 | 0.1412 | 0.1371 | 0.1944 | 0.3443 | 1.0000 |
| Jobless months (past 2 yrs) | 27 | 10.455 | 6.240 | −1.298 | 0.3328 | 0.3449 | 0.2916 | 0.4821 | 1.0000 |
| Currently employed | 40 | 0.818 | 0.833 | −0.016 | 0.9072 | 0.9173 | 0.9400 | 1.0000 | 1.0000 |
| Unemployed last year | 40 | 0.409 | 0.160 | −0.194 | 0.2421 | 0.3237 | 0.3972 | 0.5675 | 0.9211 |
| Jobless months (past 2 yrs) | 40 | 5.045 | 4.000 | 0.057 | 0.9072 | 0.9173 | 0.9400 | 1.0000 | 1.0000 |

*Note:* This table reports Holm stepdown *p*-values for multiple hypothesis tests of treatment effects on various outcomes of female participants at the given ages. The inferences are based on the studentized AIPW test statistic. The blocks used for multiple testing are indicated above using divider lines.

**Table 6.** Comparing Heckman et al.'s (2020) DIM-based inference with ours for **Male Sample**.

| | | *Heckman et al.'s (2020) p-values* | | | | *Worst-case p-values using our method* | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Variable* | *Age* | *$U = 0$ p-value (unadj.)* | *$U = 0$ p-value (adjusted)* | *Max-U p-value (unadj.)* | *Max-U p-value (adjusted)* | *Worst-case max. p (unadjusted)* | *Worst-case max. p (adjusted)* | *Worst-case de Haan p (unadjusted)* | *Worst-case de Haan p (adjusted)* |
| Stanford–Binet IQ | 4 | **0.001** | **0.001** | **0.008** | **0.008** | **0.0035** | **0.0246** | **0.0051** | **0.0358** |
| Stanford–Binet IQ | 5 | **0.022** | 0.691 | **0.077** | 0.800 | **0.0053** | **0.0319** | 0.1314 | 0.6571 |
| Stanford–Binet IQ | 6 | **0.033** | **0.034** | **0.094** | 0.102 | **0.0289** | 0.1447 | **0.0975** | 0.5848 |
| Stanford–Binet IQ | 7 | 0.103 | 0.172 | 0.247 | 0.374 | **0.0858** | 0.3433 | 0.2259 | 0.9034 |
| Stanford–Binet IQ | 8 | 0.599 | 0.691 | 0.733 | 0.800 | 0.5501 | 1.0000 | 0.7234 | 1.0000 |
| Stanford–Binet IQ | 9 | 0.450 | 0.548 | 0.631 | 0.680 | 0.5635 | 1.0000 | 0.9429 | 1.0000 |
| Stanford–Binet IQ | 10 | 0.684 | 0.691 | 0.790 | 0.800 | 0.2529 | 0.7588 | 0.3615 | 1.0000 |
| CAT reading score | 14 | **0.017** | **0.035** | **0.036** | **0.086** | 0.1444 | 0.3410 | 0.1975 | 0.6537 |
| CAT arithmetic score | 14 | **0.032** | **0.035** | **0.086** | **0.086** | 0.1185 | 0.3410 | 0.3746 | 0.6537 |
| CAT language score | 14 | **0.001** | **0.004** | **0.012** | **0.027** | **0.0686** | 0.2743 | 0.1592 | 0.6537 |
| CAT mechanics score | 14 | **0.006** | **0.007** | **0.023** | **0.035** | **0.0464** | 0.2320 | 0.1307 | 0.6537 |
| CAT spelling score | 14 | **0.003** | **0.035** | **0.012** | **0.086** | 0.1137 | 0.3410 | 0.2247 | 0.6537 |
| High school graduate | 19 | 0.614 | 0.674 | 0.704 | 0.716 | 0.6373 | 1.0000 | 0.8990 | 1.0000 |
| Vocational training | 40 | 0.341 | 0.567 | 0.547 | 0.608 | 0.3582 | 1.0000 | 0.4612 | 1.0000 |
| Highest grade completed | 19 | 0.383 | 0.622 | 0.410 | 0.669 | 0.3526 | 1.0000 | 0.5578 | 1.0000 |
| Grade point average | 19 | 0.457 | 0.674 | 0.567 | 0.716 | 0.5132 | 1.0000 | 0.8153 | 1.0000 |

**Table 6.** Continued

| | | *Heckman et al.'s (2020) p-values* | | | | *Worst-case p-values using our method* | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Variable* | *Age* | *$U = 0$ p-value (unadj.)* | *$U = 0$ p-value (adjusted)* | *Max-U p-value (unadj.)* | *Max-U p-value (adjusted)* | *Worst-case max. p (unadjusted)* | *Worst-case max. p (adjusted)* | *Worst-case de Haan p (unadjusted)* | *Worst-case de Haan p (adjusted)* |
| Total nonjuvenile arrests | 40 | **0.036** | **0.038** | **0.100** | 0.115 | **0.0713** | 0.2752 | 0.3823 | 1.0000 |
| Total crime cost | 40 | **0.037** | **0.049** | **0.042** | 0.143 | 0.1746 | 0.2752 | 0.3494 | 1.0000 |
| Total charges | 40 | **0.049** | **0.049** | 0.143 | 0.143 | 0.1136 | 0.2752 | 0.2610 | 1.0000 |
| Nonvictimless charges | 40 | **0.025** | **0.037** | **0.063** | **0.091** | **0.0688** | 0.2752 | 0.3433 | 1.0000 |
| Currently employed | 19 | **0.050** | 0.164 | 0.224 | 0.290 | 0.2763 | 0.7413 | 0.7335 | 1.0000 |
| Unemployed last year | 19 | 0.901 | 0.901 | 0.922 | 0.922 | 0.2471 | 0.7413 | 0.5214 | 1.0000 |
| Jobless months (past 2 yrs) | 19 | 0.821 | 0.849 | 0.873 | 0.890 | 0.3161 | 0.7413 | 0.5369 | 1.0000 |
| Currently employed | 27 | 0.268 | 0.295 | 0.485 | 0.512 | 0.3304 | 0.6608 | 0.6281 | 0.9799 |
| Unemployed last year | 27 | 0.235 | 0.295 | 0.360 | 0.512 | 0.3393 | 0.6608 | 0.4793 | 0.9799 |
| Jobless months (past 2 yrs) | 27 | **0.020** | **0.020** | **0.036** | **0.051** | **0.0866** | 0.2599 | 0.3266 | 0.9799 |
| Currently employed | 40 | 0.103 | 0.116 | 0.130 | 0.146 | **0.0595** | 0.1784 | 0.1149 | 0.3446 |
| Unemployed last year | 40 | 0.154 | 0.154 | 0.216 | 0.216 | 0.1694 | 0.1784 | 1.0000 | 1.0000 |
| Jobless months (past 2 yrs) | 40 | **0.064** | 0.116 | **0.070** | 0.146 | **0.0779** | 0.1784 | 0.1220 | 0.3446 |

*Note:* This table compares inferences reported by Heckman et al. (2020) with the inferences obtained using our worst-case tests. The first two columns list the blocks of outcomes analysed by Heckman et al. (2020). The next four columns reproduce their zero-$U$ ($U = 0$) $p$-values and max-$U$ $p$-values before and after adjusting for multiplicity of hypotheses. Since all of their tests are based on studentized DIM estimate, we report our inferences (using the studentized DIM test statistic) side by side for comparison. The last four columns report our worst-case maximum $p$-values and worst-case de Haan $p$-values before and after adjusting for multiplicity of hypotheses. The unadjusted $p$-values refer to single $p$-values that are unadjusted for multiplicity of hypotheses. The adjusted $p$-values refer to stepdown $p$-values after adjusting for multiple testing.

**Table 7.** Comparing Heckman et al.'s (2020) DIM-based inference with ours for **Female Sample**.

| | | *Heckman et al.'s (2020) p-values* | | | | *Worst-case p-values using our method* | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Variable* | *Age* | *$U = 0$ p-value (unadj.)* | *$U = 0$ p-value (adjusted)* | *Max-U p-value (unadj.)* | *Max-U p-value (adjusted)* | *Worst-case max. p (unadjusted)* | *Worst-case max. p (adjusted)* | *Worst-case de Haan p (unadjusted)* | *Worst-case de Haan p (adjusted)* |
| Stanford–Binet IQ | 4 | **0.008** | **0.008** | **0.020** | **0.020** | **0.0025** | **0.0174** | **0.0026** | **0.0181** |
| Stanford–Binet IQ | 5 | **0.012** | 0.203 | **0.014** | 0.354 | **0.0208** | 0.1246 | **0.0635** | 0.3810 |
| Stanford–Binet IQ | 6 | **0.094** | 0.164 | 0.160 | 0.346 | 0.1285 | 0.5141 | 0.3200 | 1.0000 |
| Stanford–Binet IQ | 7 | 0.133 | 0.137 | 0.191 | 0.222 | **0.0796** | 0.3982 | 0.3847 | 1.0000 |
| Stanford–Binet IQ | 8 | 0.152 | 0.164 | 0.339 | 0.346 | 0.1514 | 0.5141 | 0.6181 | 1.0000 |
| Stanford–Binet IQ | 9 | 0.203 | 0.203 | 0.354 | 0.354 | 0.2102 | 0.5141 | 0.3197 | 1.0000 |
| Stanford–Binet IQ | 10 | 0.203 | 0.203 | 0.267 | 0.354 | 0.1301 | 0.5141 | 0.7997 | 1.0000 |
| CAT reading score | 14 | **0.078** | **0.082** | 0.136 | 0.167 | **0.0358** | **0.0715** | 0.1081 | 0.3244 |
| CAT arithmetic score | 14 | **0.035** | **0.082** | **0.074** | 0.167 | 0.1046 | 0.1046 | 0.1704 | 0.3407 |
| CAT language score | 14 | **0.008** | **0.070** | **0.020** | 0.144 | **0.0113** | **0.0566** | **0.0328** | 0.1640 |
| CAT mechanics score | 14 | **0.047** | **0.082** | **0.097** | 0.167 | **0.0137** | **0.0566** | 0.1974 | 0.3407 |
| CAT spelling score | 14 | **0.043** | **0.082** | **0.082** | 0.167 | **0.0115** | **0.0566** | **0.0434** | 0.1736 |
| High school graduate | 19 | **0.008** | **0.008** | **0.020** | **0.020** | **0.0037** | **0.0148** | **0.0236** | **0.0709** |
| Vocational training | 40 | **0.078** | **0.078** | 0.144 | 0.144 | 0.1085 | 0.1085 | 0.1872 | 0.1872 |
| Highest grade completed | 19 | **0.070** | **0.070** | 0.113 | 0.113 | **0.0297** | **0.0593** | **0.0585** | 0.1169 |
| Grade point average | 19 | **0.039** | **0.039** | **0.082** | **0.082** | **0.0086** | **0.0259** | **0.0151** | **0.0603** |

**Table 7.** Continued

| | | *Heckman et al.'s (2020) p-values* | | | | *Worst-case p-values using our method* | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Variable* | *Age* | *$U = 0$ p-value (unadj.)* | *$U = 0$ p-value (adjusted)* | *Max-U p-value (unadj.)* | *Max-U p-value (adjusted)* | *Worst-case max. p (unadjusted)* | *Worst-case max. p (adjusted)* | *Worst-case de Haan p (unadjusted)* | *Worst-case de Haan p (adjusted)* |
| Total nonjuvenile arrests | 40 | **0.020** | 0.133 | 0.121 | 0.158 | 0.1245 | 0.2403 | 0.1625 | 0.4874 |
| Total crime cost | 40 | **0.024** | 0.133 | **0.082** | 0.158 | **0.0601** | 0.2403 | **0.0983** | 0.3932 |
| Total charges | 40 | **0.020** | **0.067** | **0.043** | **0.090** | 0.1005 | 0.2403 | 0.1661 | 0.4874 |
| Nonvictimless charges | 40 | 0.125 | 0.133 | 0.158 | 0.158 | **0.0677** | 0.2403 | 0.2141 | 0.4874 |
| Currently employed | 19 | **0.008** | **0.031** | **0.035** | **0.090** | **0.0562** | 0.1323 | **0.0899** | 0.2697 |
| Unemployed last year | 19 | **0.024** | **0.031** | **0.074** | **0.090** | **0.0441** | 0.1323 | 0.1037 | 0.2697 |
| Jobless months (past 2 yrs) | 19 | 0.125 | 0.125 | 0.206 | 0.206 | **0.0858** | 0.1323 | 0.2354 | 0.2697 |
| Currently employed | 27 | 0.110 | 0.149 | 0.175 | 0.198 | **0.0760** | 0.2281 | 0.1810 | 0.3969 |
| Unemployed last year | 27 | **0.078** | 0.149 | 0.128 | 0.175 | **0.0962** | 0.2281 | 0.1323 | 0.3969 |
| Jobless months (past 2 yrs) | 27 | 0.110 | 0.149 | 0.166 | 0.198 | 0.1970 | 0.2281 | 0.2889 | 0.3969 |
| Currently employed | 40 | 0.442 | 0.442 | 0.567 | 0.567 | 0.4818 | 0.8816 | 1.0000 | 1.0000 |
| Unemployed last year | 40 | **0.047** | **0.070** | 0.113 | 0.160 | 0.1000 | 0.3001 | 0.2254 | 0.6761 |
| Jobless months (past 2 yrs) | 40 | 0.352 | 0.367 | 0.540 | 0.540 | 0.4408 | 0.8816 | 0.5519 | 1.0000 |

*Note:* This table compares inferences reported by Heckman et al. (2020) with the inferences obtained using our worst-case tests. The first two columns list the blocks of outcomes analysed by Heckman et al. (2020). The next four columns reproduce their zero-$U$ ($U = 0$) $p$-values and max-$U$ $p$-values before and after adjusting for multiplicity of hypotheses. Since all of their tests are based on studentized DIM estimate, we report our inferences (using the studentized DIM test statistic) side by side for comparison. The last four columns report our worst-case maximum $p$-values and worst-case de Haan $p$-values before and after adjusting for multiplicity of hypotheses. The unadjusted $p$-values refer to single $p$-values that are unadjusted for multiplicity of hypotheses. The adjusted $p$-values refer to stepdown $p$-values after adjusting for multiple testing.

the same test statistic. The effects for males on post-programme outcomes remain statistically insignificant at the 10% level using stepdown worst-case de Haan $p$-values, whereas treatment effects on CAT scores are statistically significant in Heckman et al.'s (2020) analysis.

Heckman et al. (2020) do not analyse the Perry treatment effects on convictions for violent crime, which are substantial and play an important role in cost–benefit analyses of early childhood programmes (see Heckman et al., 2010b). Using administrative data on the criminal activity of participants, we illustrate their importance and, at the same time, the importance of long-term follow-up. Tables 8, 9, and 10 provide estimates and measures of statistical significance of treatments effects in the pooled sample (of all participants) on cumulative convictions for violent misdemeanors and felonies at various ages. Online Appendix S2 presents the expanded versions of these tables reporting inference for various estimators and test statistics for the pooled sample as well as the male and female subsamples. As shown in Table 9, the AIPW estimates of the treatment effect on cumulative violent misdemeanor convictions are below $-0.5$ at ages 30 and 40. These estimates of treatment effects on violent misdemeanor convictions are statistically significant at the 5% and 10% levels before and after multiple hypothesis testing, respectively.

The choice of inferential method becomes more important in analysing treatment effects on cumulative convictions for felonies. At age 30, there are no statistically significant treatment effects. At age 40, as shown in Table 9, the magnitude of the treatment effect is higher at about $-0.21$, which represents more than a four-tenths reduction in the control mean. However, using simple difference-in-means estimates and conventional $p$-values can be misleading. Using conventional $p$-values, the effect at age 40 appears to be statistically significant at the 10% level, as shown in Table 8. However, the design-based worst-case $p$-values, especially those associated with the AIPW estimate, are much higher. The worst-case de Haan $p$-values for the studentized DIM and AIPW estimates are about 0.136 and 0.241, respectively.

The four variables at ages 30 and 40 considered in Tables 8 and 9 are conceptually related, since they are cumulative crime outcomes measured at different ages. To account for this, we treat these outcomes as a single block of variables and conduct multiple hypothesis testing using the more conservative Holm stepdown procedure, producing results in Table 10. After multiple testing, the effects on cumulative convictions for violent misdemeanors remain statistically significant at the 10% level at both ages 30 and 40, whereas the the effects on violent felonies are insignificant at both ages. These analyses show that use of small-sample inference and the method used to account for compromised randomization matter in analysing the data. Failure to account for either can give a very positive spin to the Perry programme. Accounting for them qualifies such conclusions. We have not, however, established the superiority of our approach. We have established that a very cautious design-based approach produces conservative inference, which by itself is not surprising. Our reanalysis of Heckman et al. (2020) is very conservative. Nonetheless, a few conclusions survive. We test Fisher's sharp null hypothesis $\mathcal{H}_{\mathcal{F}}$ of no treatment effect for *each* participant. It may in fact be the case that there are treatment effects for many participants and yet we do not reject the sharp null hypothesis because of our worst-case approach.

## 6. CONCLUSION

In this paper, we develop and apply a design-based finite-sample inferential method for analysing social experiments with compromised randomization. Compromises come in many forms. They include incompletely documented rerandomization procedures used to improve baseline covariate

**Table 8.** DIM-based single hypothesis tests on cumulative convictions for violent crime.

| *Type* | *Age* | *Untreated mean* | *Treated mean* | *DIM estimate* | *Asymptotic p-value* | *Bootstrap p-value* | *Permutation p-value* | *Worst-case max. p* | *Worst-case de Haan p* |
|---|---|---|---|---|---|---|---|---|---|
| Misdemeanor | 30 | 0.5231 | 0.0517 | −0.4714 | **0.0109** | **0.0021** | **0.0036** | **0.0135** | 0.1002 |
| Misdemeanor | 40 | 0.6825 | 0.0877 | −0.5948 | **0.0033** | **0.0005** | **0.0004** | **0.0054** | **0.0092** |
| Felony | 30 | 0.2846 | 0.1897 | −0.0950 | 0.2301 | 0.2263 | 0.2624 | 0.3867 | 0.6691 |
| Felony | 40 | 0.4762 | 0.1930 | −0.2832 | **0.0333** | **0.0332** | **0.0384** | **0.0792** | 0.1362 |

*Note:* This table reports *p*-values for single hypothesis tests of treatment effects on cumulative misdemeanor and felony convictions for violent crime at ages 30 and 40, using the pooled sample of participants. The inferences are based on the studentized DIM (difference-in-means) test statistic.

**Table 9.** AIPW-based single hypothesis tests on cumulative convictions for violent crime.

| *Type* | *Age* | *Untreated mean* | *Treated mean* | *AIPW estimate* | *Asymptotic p-value* | *Bootstrap p-value* | *Permutation p-value* | *Worst-case max. p* | *Worst-case de Haan p* |
|---|---|---|---|---|---|---|---|---|---|
| Misdemeanor | 30 | 0.5231 | 0.0517 | −0.5300 | **0.0064** | **0.0020** | **0.0024** | **0.0102** | **0.0267** |
| Misdemeanor | 40 | 0.6825 | 0.0877 | −0.6491 | **0.0021** | **0.0010** | **0.0008** | **0.0051** | **0.0052** |
| Felony | 30 | 0.2846 | 0.1897 | −0.0561 | 0.3174 | 0.3217 | 0.3488 | 0.4809 | 0.7310 |
| Felony | 40 | 0.4762 | 0.1930 | −0.2052 | **0.0664** | **0.0778** | **0.0708** | 0.1376 | 0.2412 |

*Note:* This table reports *p*-values for single hypothesis tests of treatment effects on cumulative misdemeanor and felony convictions for violent crime at ages 30 and 40, using the pooled sample of participants. The inferences are based on the studentized AIPW test statistic.

**Table 10.** AIPW-based multiple hypothesis tests on cumulative convictions for violent crime.

| *Type* | *Age* | *Untreated mean* | *Treated mean* | *AIPW estimate* | *Asymptotic p-value* | *Bootstrap p-value* | *Permutation p-value* | *Worst-case max. p* | *Worst-case de Haan p* |
|---|---|---|---|---|---|---|---|---|---|
| Misdemeanor | 30 | 0.5231 | 0.0517 | −0.5300 | **0.0192** | **0.0059** | **0.0072** | **0.0306** | **0.0800** |
| Misdemeanor | 40 | 0.6825 | 0.0877 | −0.6491 | **0.0085** | **0.0039** | **0.0032** | **0.0204** | **0.0208** |
| Felony | 30 | 0.2846 | 0.1897 | −0.0561 | 0.3174 | 0.3217 | 0.3488 | 0.4809 | 0.7310 |
| Felony | 40 | 0.4762 | 0.1930 | −0.2052 | 0.1327 | 0.1556 | 0.1416 | 0.2752 | 0.4824 |

*Note:* This table reports Holm stepdown *p*-values for multiple hypothesis tests of treatment effects on cumulative misdemeanor and felony convictions for violent crime at ages 30 and 40, using the pooled sample of participants. The inferences are based on the studentized AIPW test statistic. All the above four variables, which represent cumulative crime outcomes at different ages, are treated as a block for multiple testing.

balance between treatment and control groups. They also include reassignment of treatment status due to administrative constraints.

We build a behavioural model of satisficing experimenters who seek balance in baseline covariates across treatments and controls and who provide readers of their reports qualitative, and sometimes conflicting, summaries of the actual experimental protocols used. We model the randomization protocol as only partially known to the user of experimental data. The empirical researcher recognizes and tries to account for the guiding principles experimenters used in the reassignment of treatment status for balancing baseline covariates while operating under administrative constraints. We show how to partially identify model parameters and construct worst-case (least favourable) randomization tests over a set of possibilities for the actual treatment assignment mechanism.

Our analysis of the Perry programme serves as a proof-of-concept of the usefulness of our worst-case finite-sample testing approaches, which are applicable to other compromised experiments, such as those discussed by Bruhn and McKenzie (2009). Our approach is more portable than that of Heckman et al. (2020), which utilizes very specific features of the Perry randomization protocol. Application of our procedures result in conservative finite-sample inferences.

## ACKNOWLEDGEMENTS

This paper was delivered as the 2019 Sargan Lecture at the Royal Economic Society Annual Conference at the University of Warwick, England. It has been subject to the usual refereeing standards of this journal. We thank the editor and anonymous referees for useful comments. We also thank Juan Pantano and Azeem Shaikh for comments on early drafts of this paper. We are grateful to the HighScope Educational Research Foundation for access to study data and source materials. This research was supported in part by: the Buffett Early Childhood Fund; NIH Grants R01AG042390, R01AG05334301, and R37HD065072. The views expressed in this paper are solely those of the authors and do not necessarily represent those of the funders or the official views of the National Institutes of Health.

## REFERENCES

Abadie, A., S. Athey, G. W. Imbens and J. M. Wooldridge (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica 88*(1), 265–96.

Athey, S. and G. W. Imbens (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments Volume 1*, 73–140. Amsterdam: Elsevier.

Bruhn, M. and D. McKenzie (2009). In pursuit of balance: randomization in practice in development field experiments. *American Economic Journal: Applied Economics 1*(4), 200–32.

Campbell, F. A., C. T. Ramey, E. Pungello, J. Sparling and S. Miller-Johnson (2002). Early childhood education: young adult outcomes from the Abecedarian Project. *Applied Developmental Science 6*(1), 42–57.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics 155*(2), 138–54.

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

Chung, E. and J. P. Romano (2016). Multivariate and multiple permutation tests. *Journal of Econometrics 193*(1), 76–91.

Cunha, F., J. J. Heckman, L. Lochner and D. V. Masterov (2006). Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education 1*, 697–812.

de Haan, L. (1981). Estimation of the minimum of a function using order statistics. *Journal of the American Statistical Association 76*(374), 467–69. [Corrigendum, *JASA 78*(384), 1008.]

Elango, S., J. L. García, J. J. Heckman and A. Hojman (2015). Early childhood education. In *Economics of Means-Tested Transfer Programs in the United States Volume 2*, 235–97. Chicago: University of Chicago Press.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Greenland, S. and M. A. Mansournia (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine 34*(23), 3133–43.

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

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev and A. Yavitz (2010b). The rate of return to the Highscope Perry Preschool Program. *Journal of Public Economics 94*(1–2), 114–28.

Heckman, J. J., R. Pinto and A. M. Shaikh (2020). *Inference with imperfect randomization: the case of the Perry Preschool Program*. Unpublished manuscript, The University of Chicago.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Kang, J. D. Y. and J. L. Schafer (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–39.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association 88*(424), 1242–49.

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

Li, X. and P. Ding (2016). Exact confidence intervals for the average causal effect on a binary outcome. *Statistics in Medicine 35*(6), 957–60.

Li, X., P. Ding and D. B. Rubin (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences 115*(37), 9157–62.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine 23*(19), 2937–60.

Morgan, K. L. and D. B. Rubin (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics 40*(2), 1263–82.

Morgan, K. L. and D. B. Rubin (2015). Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association 110*(512), 1412–21.

Neyman, J. S. (1923). Próba uzasadnienia zastosowań rachunku prawdopodobieństwa do doswiadczeń polowych (On the application of probability theory to agricultural experiments: essay on principles). *Roczniki Nauk Rolniczych (Annals of Agricultural Sciences) 10*, 1–51. Reprinted in *Statistical Science 5*(4), 465–72, as a translation by D. M. Dabrowska and T. P. Speed (1990) from section 9 (29–42) of the original Polish article.

Obama, B. (2013). *The 2013 State of the Union Address*. Washington, DC: The White House Office of the Press Secretary.

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association 53*(284), 873–80.

Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association 67*(338), 306–10.

Rigdon, J. and M. G. Hudgens (2015). Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine 34*(6), 924–35.

Robins, J. M., A. Rotnitzky and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*(427), 846–66.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics 17*(1), 141–59.

Romano, J. P. and A. M. Shaikh (2010). Inference for the identified set in partially identified econometric models. *Econometrica 78*(1), 169–211.

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

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

Schweinhart, L. J. (2013). Long-term follow-up of a preschool experiment. *Journal of Experimental Criminology 9*(4), 389–409.

Schweinhart, L. J. and D. P. Weikart (1980). *Young Children Grow Up: The Effects of the Perry Preschool Program on Youths Through Age 15*. Ypsilanti, MI: HighScope Educational Research Foundation.

Schweinhart, L. J., H. V. Barnes, D. P. Weikart, W. Barnett and A. Epstein (1993). *Significant Benefits: The High/Scope Perry Preschool Study Through Age 27 (Monographs of the High/Scope Educational Research Foundation, 10)*. Ypsilanti, MI: HighScope Educational Research Foundation.

Schweinhart, L. J., J. R. Berrueta-Clement, W. S. Barnett, A. S. Epstein and D. P. Weikart (1985). The promise of early childhood education. *The Phi Delta Kappan 66*(8), 548–53.

Schweinhart, L. J., J. Montie, Z. Xiang, W. S. Barnett, C. R. Belfield and M. Nores (2005). *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40 (Monographs of the High/Scope Educational Research Foundation, 14)*. Ypsilanti, MI: HighScope Educational Research Foundation.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics 69*(1), 99–118.

Weikart, D. P., J. T. Bond and J. T. McNeil (1978). *The Ypsilanti Perry Preschool Project: Preschool Years and Longitudinal Results Through Fourth Grade*. Number 3. Ypsilanti, MI: HighScope Educational Research Foundation.

Weikart, D. P., C. K. Kamii and N. L. Radin (1964). Perry Preschool Project progress report. Technical report. Ypsilanti, MI: Ypsilanti Public Schools.

Wu, J. and P. Ding (2020). Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association*, in press.

Young, A. (2019). Channeling Fisher: randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics 134*(2), 557–98.

Zigler, E. and D. P. Weikart (1993). Reply to Spitz's comments. *American Psychologist 48*(8), 915–16.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendix
Replication Package

*Managing editor Jaap Abbring handled this manuscript.*

## APPENDIX A: BACKGROUND AND ELIGIBILITY CRITERIA OF PERRY PROGRAMME

The Perry Preschool Project was carried out in five waves between autumn 1962 and autumn 1965 near a public school—the Perry Elementary School in Ypsilanti, a small city near Detroit in Michigan. Data collection took place at the baseline age of 3 years and through surveys that were administered annually until age 15. The participants were additionally followed up around ages 19, 27, 40, and 55. Various measures were obtained over the years, including information on education, crime, and other economic outcomes.

Intensity of the programme was low relative to several later early education programmes.[60] Starting at age 3, treatment in the following two years included preschool for 2.5 hours per day on weekdays during the academic year. Another major component of the programme consisted of 1.5-hour weekly home visits by the Perry teachers to promote parental engagement with the child.[61] The Perry curriculum fostered active child-centered learning through intensive interactions between the children and programme teachers (Weikart et al., 1978; Schweinhart et al., 1993).

Door-to-door canvassing and referrals were used to survey and identify disadvantaged families among those of the Perry Elementary School students. To be eligible for participation in the Perry Preschool Project, the children had to: (i) be African American; (ii) have low Stanford–Binet IQ scores at baseline;[62] and (iii) be socioeconomically disadvantaged according to an index of socioeconomic status based on employment and education levels of the parents as well as the number of persons per room at home. The Perry families were more disadvantaged relative to a majority of African American families at that time in the United States. However, the Perry families were, by and large, representative of a substantial fraction of the underprivileged African American population (Heckman et al., 2010a).

Even when compared with the children living in the area surrounding the Perry Elementary School, the Perry participants were especially disadvantaged (Heckman et al., 2010a). Since the parents of all children eligible for the programme participated in the study (Weikart et al., 1978), issues of noncompliance are not a concern. As there were no substitutes to the Perry programme, such as Head Start, available when the Perry experiment was implemented, control group contamination is also not a problem in our experimental setting.

## APPENDIX B: EXCHANGES WERE NOT BASED ON CONSECUTIVE IQ SCORES

We use Perry data from wave 4 as an example to conclude that the exchanges were not necessarily between consecutively ranked pairs. In wave 4, there were 19 participants, excluding any younger siblings in the programme. The IQs of these 19 people were: 61, 71, 75, 76, 76, 76, 78, 78, 79, 79, 80, 80, 81, 82, 83, 83, 83, 85, 88, involving many ties. Regardless of which method was used to break the ties, from a pure ranking procedure the staff would have obtained two initial groups: one with IQs $\{61, 75, 76, 78, 79, 80, 81, 83, 83, 88\}$ and another group with IQs $\{71, 76, 76, 78, 79, 80, 82, 83, 85\}$. The final observed treatment group has IQs in the set: $\{61, 75, 76, 78, 80, 81, 83, 83, 83, 88\}$. Note that the person with IQ 79 is replaced by

[60] An example is the Carolina Abecedarian Project (see, e.g., Campbell et al., 2002). For a discussion and comparison of the intensity of several such programmes, see Cunha et al. (2006) and Elango et al. (2015).

[61] Those in the treatment group of the first entry cohort (wave 0) were provided with the intervention for only one year, starting at age 4, and thus were an exception. In our estimation of treatment effects, we pool all five cohorts, even though the lower programme intensity in the first cohort might in principle attenuate the magnitudes of the effects downward.

[62] The initial eligibility criteria specified that the IQs, as measured by the Stanford–Binet IQ test according to 1960's norming, be between 70 and 85, which was one standard deviation below the population average. However, in practice, the IQ range was 61 to 88. Only about two-thirds of the participants had IQs in the range specified initially.

a person with IQ 83. The final observed control group has IQs in the set: {71, 76, 76, 78, 79, 79, 80, 82, 85}. Note that the person with IQ 83 is replaced by a person with IQ 79. These are the same as the initial treatment and control groups, since there were no transfers in the fifth step of the protocol, as explained in Example 3 of the paper. Thus, we can conclude that an exchange happened between participants with IQs 79 and 83, who do not comprise a consecutively ranked pair. Thus, after the IQ rank ordering, the exchanges between the two initial groups were not always between consecutively ranked IQ pairs. Thus, the Perry staff did not strictly implement a matched pair design.