Web Appendix

An Analysis of the Memphis Nurse-Family Partnership Program[1]

James J. Heckman
University of Chicago

Margaret L. Holland
Yale University

Kevin K. Makino
University of Hawaii

Rodrigo Pinto
UCLA

Maria Rosales-Rueda
University of California, Irvine

This draft: July 11, 2017

[1]James J. Heckman is the Henry Schultz Distinguished Service Professor of Economics, University of Chicago, Director of the Center for the Economics of Human Development, and Research Associate, American Bar Foundation. David Olds is the founder of the Nurse-Family Partnership program, Professor of Pediatrics, Psychiatry and Preventive Medicine, University of Colorado. Rodrigo Pinto is an Assistant Professor of Economics at the University of California, Los Angeles. Maria Rosales is an Assistant Professor of Education Policy at the University of California, Irvine and Visiting Research Scholar at the Center of Health and Wellbeing, Princeton University. Margaret Holland is a health services researcher at Yale School of Nursing. Kevin Makino is a Pediatric Resident at the John A. Burns School of Medicine, University of Hawaii. We are very grateful to Professor David Olds, founder of the NFP, and his team for generously sharing the data and source materials from the NFP Memphis Randomize Control Trial. Years of collaboration and productive discussions with him have made this study possible. We thank Terrance Oey and Willem van Vliet for superb research assistance. We are grateful to Juan Pantano, Sylvi Kuperman, Jorge Luis García, Maryclare Griffin, Andrés Hojman, Yu Kyung Koh, Cullen Roberts, Karl Schulze, Yu Kyung Koh, Naoko Takeda, and Joyce Zhu for helpful comments. An early version of this paper was presented at seminars at the University of Chicago and the Association for Public Policy Analysis and Management 2012 Fall conference. This research was supported by NICHD R37HD065072 and NICHD R01HD054702, and previous support of the Pritzker Childrens Initiative. The views expressed in this paper are solely those of the authors and do not necessarily represent the official views of the National Institutes of Health. The Web Appendix for this paper may be found at https://cehd.uchicago.edu/nfp-reanalysis.

Contents

A Memphis NFP Randomization Protocol

The NFP Memphis trial recruited pregnant women from June 1, 1990 to August 31, 1991 through the Memphis-Shelby County Tennessee Health Department. Eligible mothers satisfied the following biological criteria: (1) less than 29 weeks of pregnancy; (2) no previous live birth; and (3) no chronic illnesses that could contribute to fetal-growth retardation or preterm delivery. They also satisfied two or more of the following socio-economic criteria: (1) unmarried; (2) less than 12 years of education; (3) unemployed. The randomization protocol was sequential, that is to say, that each participant was randomized according to the order of enrollment. Pregnant women who agreed to enroll were classified in strata defined by 5 characteristics:

1. Maternal race (African American vs non-African American);
2. Maternal age (< 17, $17 - 18$, > 18 years);
3. Gestational age at enrollment (< 20, ≥ 20 weeks);
4. Employment status of the head of household
5. 4 geographic regions of residence.[1]

Within strata, randomization was performed following the Soares and Wu (1983) method as follows:

1. If the participant had a sibling already enrolled in the program, the participant was assigned to the same treatment status as the elder sibling.
2. Else, the participant was randomized into the control (C) or treatment group (T). However, if the sample size difference between treatment and control group was larger than a threshold, the participant was deterministically assigned to the treatment status that had fewer participants.

[1] The regions are: Inner City, Bisson, Cawthon and Hollywood.

3. Next the participant was randomized again into her final treatment status. If in step 2 the participant was assigned to the control, she was next randomized to:

    - *Group 1*: control women who received free transportation to and from their prenatal appointments (sample size: 166).
    - *Group 2*: control women who received developmental screening and referral services at ages 6, 12 and 24 months in addition to the benefits of Group 1 (sample size: 514).

    If she was assigned to the treatment group previously, she is next randomized into:

    - *Group 3*: treated women who received home visits by nurses during pregnancy, one visit in the hospital and one visit at home after childbirth in addition to the benefits of Group 2 (sample size: 230).
    - *Group 4*: treated women that received home visits by nurses during pregnancy until the child's 2^{nd} birthday, in addition to the benefits of Group 2 (sample size: 228)[2]

    As in the previous step, if the absolute difference in group size exceeded some threshold then the participant was deterministically assigned to the group with the lowest number of participants. Otherwise, the pregnant woman was randomly assigned.

Importantly, the randomization method incorporated a trigger mechanism that deterministically assigned a treatment status to participants if the sequence of assignments became too imbalanced due to sampling variation. In this context, imbalance was measured by the difference of persons assigned to T and the persons assigned to C. In practice less than 1% of the women were assigned according to the trigger mechanism. Thus, the NFP Memphis trial can be treated as a non-sequential protocol.

[2]Nurses completed on average 7 visits during pregnancy (range: 0-18) and 26 visits during the first two years of the child's life (range: 0-71) (Olds and Korfmacher, 1998)

B Brief survey of the NFP Literature

The NFP program has been evaluated in three Randomized Control Trials (RCTs): Elmira, NY (1978), Memphis, TN (1990) and Denver, CO (1994), each of them targeting low-income first-time mothers from different racial backgrounds. In Elmira, the sample was mainly low-income white; in Memphis, the majority of the participants were low-income African American; and in Denver, the sample included a significant fraction of low-income Hispanics. Another important difference is that in the Elmira and Memphis trials, the visits were conducted by nurses. In contrast, in the Denver trial, one treatment group received visits by Nurses and another group by Paraprofessionals. Table B.1 describes the main features of the randomization arms of each NFP trial.

- Denver Trial Treatment Groups:

  1. Women in the control group (n = 255) were provided developmental screening and referral services for their children at 6, 12, 15, 21, and 24 months old *(same as Memphis Group 2)*.
  2. Paraprofessional Group: Women assigned to the paraprofessional group (n = 245) were provided the screening and referral services plus paraprofessional home visitation during pregnancy and infancy, that is, the first 2 years of the child's life *(No such group in Memphis Trial)*.
  3. Nurse Group: Women in the nurse group (n = 235) were provided screening and referral plus nurse home visitation during pregnancy and infancy *(Same as Memphis Group 4)*.

- Elmira Trial Treatment Groups

  1. **Group 1.** When the children were 1 and 2 years of age, an infant specialist hired by the research project screened them for sensory and developmental problems and

referred those with suspected problems to other specialists for further evaluation and treatment *(similar to Memphis Group 2)*.

2. **Group 2.** Families were provided free transportation for regular prenatal care at local clinics and physicians' offices through a contract with a local taxicab company, as well as the sensory and developmental screening outlined in treatment 1 *(same as Memphis Group 2)*.

3. **Group 3.** Families were provided a nurse home visitor during pregnancy, in addition to the screening and transportation services. The nurses visited families approximately once every 2 weeks and made an average of nine visits during pregnancy. The average visit lasted 1 hour and 15 minutes *(similar but weaker treatment than Memphis Group 3)*.

4. **Group 4.** Families received the same services as those in treatment 3, but in addition the nurse continued to visit until the children were 2 years of age. For 6 weeks after delivery the nurses visited families every week; from 6 weeks to 4 months, they visited every 2 weeks; from 4 to 14 months, every 3 weeks; from 14 to 20 months, every 4 weeks; and from 20 to 24 months, every 6 weeks. Under predetermined crisis conditions the nurses visited weekly. As with the visits during pregnancy, the average visit lasted approximately 1 hour and 15 minutes, but the mean number of visits completed from birth through the end of the program was 23 *(same as Memphis Group 4)*.

B.1 Summary of key Findings of the NFP Literature

We summarize selected papers on NFP below. We also describe their main results. Tables K.1–K.15 provide further information of the published papers of the NFP literature.

1. Memphis Trial

Table B.1: Description of Randomization Groups for Each NFP Trial

Elmira Trial
Services provided (+) in Each of the Four Treatment Groups

	1 (N=90)	2 (N=94)	3 (N=100)	4 (N=116)
Health and developmental screening at the child's 12th and 24th month of life	X	X	X	X
Free transportation to regular prenatal and well-child visits		X	X	X
Nurse home visitation during pregnancy			X	X
Nurse home visitation during the child's first 2 years of life				X

Memphis Trial
Services provided (+) in Each of the Four Treatment Groups

	1 (N=166)	2 N=(514)	3 (N=230)	4 (N=228)
Free transportation to regular prenatal and well-child visits	X	X	X	X
Health and developmental screening at the child's 6th, 12th and 24th month of life		X	X	X
Nurse home visitation during pregnancy			X	X
Nurse home visitation during the child's first 2 years of life				X

Denver Trial
Services provided (+) in Each of the Three Treatment Groups

	1 (N=255)	2 (N=245)	3 (N=235)
Developmental screening and referral services at the child's 6th, 12th , 15th, 21st and 24th month of life	X	X	X
Paraprofessional home visitation during pregnancy and the child's first 2 years of life		X	
Nurse home visitation during pregnancy and the child's first 2 years of life			X

- Kitzman et al. (2000):

  Treatment group had fewer subsequent pregnancies; longer interval between the birth of the first child and the second; fewer months of using welfare. All results were statistically significant.

- Olds et al. (2004):

  Women visited by nurses had fewer subsequent pregnancies and births, longer intervals between births of the first and second children, longer relationships with current partners, fewer months of using welfare. Nurse-visited children were more likely to be enrolled in out-of-home care between 2 and 4.5 years, had higher higher intellectual functioning, higher vocabulary scores, fewer behavior problems and higher arithmetic achievement

- Olds et al. (2007):

  Nurse-visited women had longer intervals between the births of their first and second children, fewer subsequent births, and longer relationships with current partners. From birth through child age 9, nurse-visited women used welfare and food stamps for fewer months. Nurse-visited children born to mothers with low psychological resources had higher achievement test scores in math and reading in grades 1 through 3.

2. Denver Trial

- Olds et al. (2002):

  Nurse-visited mothers reduced smoking during pregnancy, had fewer subsequent pregnancies and births; had delayed subsequent pregnancies; worked more; interacted more with the child. At 6 months of age, nurse-visited infants were less likely to exhibit emotional vulnerability in response to fear stimuli and nurse-visited infants born to women with low psychological resources were less likely to exhibit low emotional vitality in response to joy and anger stimuli. At 21 months, nurse-visited children born to women with low psychological resources were less likely to exhibit language delays. At 24 months, they exhibited superior mental development (Development Index scores). No statistically significant program effects for use of ancillary prenatal services, educational achievement, use of welfare, child temperament or behavior problems. There is a single statistically significant result for paraprofessional-visited mothers when compared to their control group. Mothers with low psychological resources visited by paraprofessional interacted more with their children.

- Olds et al. (2004):

  Women who were visited by paraprofessionals, were less likely to be married, more likely to live with the biological father, were more likely to work, reported a greater sense of mastery and had better mental health. Paraprofessional-visited women, had fewer subsequent miscarriages, gave birth to fewer low birth weight newborns, displayed greater sensitivity and responsiveness toward one another and, for mothers with low psychological resources, home environments were more supportive. Nurse-visited mothers reported greater intervals between the births of their first and second children, experienced less domestic violence and enrolled their children less frequently in preschool, Head Start, or licensed day care than did control subjects. Nurse-visited children whose mothers had low levels of psychological resources had home environments that were more supportive of childrens early learning, more advanced language, and superior executive functioning. There were no statistically significant effects of either nurse or paraprofessional visits on the number of subsequent pregnancies, mother's educational achievement, use of substances, use of welfare, or children's externalizing behavior problems.

3. Elmira Trial

- Olds et al. (1997):

Women who were visited by nurses during pregnancy and infancy were less likely to be perpetrators of child abuse and neglect compared to the control group. Nurse-visited women from low socioeconomic status households who were unmarried had 1.3 vs 1.6 subsequent births and delayed the second birth. Nurse-visited women were less likely to use welfare, had fewer behavioral impairments due to use of alcohol and other drugs and had fewer arrests.

- Eckenrode et al. (2010):

  Nurse-visited mothers had fewer lifetime arrests and convictions. Nurse-visited mothers of low-income had fewer children and were less likely to use Medicaid.

We present a more detailed analysis of these trials in Appendix K.

C Assessment Instruments

C.1 HOME

The Home Observation for Measurement of the Environment (HOME) was first developed in the 1960s by Caldwell. It measures the quality and quantity of stimulation and support available to a child at home (Caldwell and Bradley, 1984). The assessment documents the child's home environment and is observed during a visit of 45 to 90 minutes. A more in depth explanation of the HOME inventory is in Caldwell and Bradley (1984).

There are several versions of the inventory. The initial version, Infant/Toddler HOME, is for children aged 0 to 3 years old. It consists of 45 binary-choice items grouped into 6 subscales. The Early Childhood HOME is for children aged 3 to 6 years old. It consists of 55 binary-choice items clustered into 8 subscales. Finally, the Middle Childhood HOME is used for children aged 6 to 10 years old. It consists of 59 items in 8 subscales. The NFP uses the first version of the Inventory, the Infant/Toddler HOME.

The 45 items of the HOME inventory contain the six following subscales:

1. *Emotional and Verbal Responsiveness of the Mother (11 items): measures the mother's ability to communicate with the child.*

2. *Avoidance of Restriction and Punishment (8 items): measures the mother's ability to discipline the child.*

3. *Organization of the Environment (6 items): measures the daily changes in the child's environment.*

4. *Provision of Appropriate Play Material (9 items): measures the types of toys and their contributions to the child's motor skills.*

5. *Maternal Involvement with Child (6 items): measures the aspects in which the mother is involved in the child's daily life.*

6. *Opportunities for Variety in Daily Stimulation (5 items): measures the levels of interaction the mother and other family members have with the child.*

The NFP measured the HOME Inventory when the child was 12 and 24 months old.

C.2 KABC

The Kaufman Assessment Battery for Children (K-ABC) was developed by Alan S. Kaufman and Nadeen L. Kaufman in 1983 with a later revision in 2004. The KABC focuses on processes required to solve problems compared to psychological instruments that focus on measuring raw cognitive skills. In broad terms, KABC focuses on the process of acquiring and manipulating information according to a determined protocol. The KABC contains 16 subtests (10 mental processing and 6 achievement), which can be grouped into 3 scales. Due to the nature of the subtests, 13 subtests can be taken at once, with the mandatory age range to be between 7 to 12.5 years old. The NFP used the following 11 subtests:

1. *Sequential Processing Scale (Hand Movements, Number Recall, Word Order): measures short-term memory and problem-solving skills. It emphasizes how children are able to follow ordered sequences.*

2. *Simultaneous Processing Scale (Gestalt Closure, Triangles, Matrix Analogies, Spatial Memory, Photo Series): measures problem-solving skills. It involves several processes at once such as scenes in a partially completed picture.*

3. *Achievement Scale (Arithmetic, Riddles, Reading/Decoding): measures achievement and focuses on applied skills and facts learned through the home/school environment.*

The NFP Program used these three scales when the child was 6 years old.

C.3 PPVT

The Peabody Picture Vocabulary Test (PPVT) is an individual verbal intelligence test that measures receptive vocabulary, developed by Llyod M. Dunn and Leota M. Dunn in 1959. It is a verbal test that lasts between 20 and 30 minutes. The child is presented a series of pictures. There are four pictures in a page. The examiner states a word and asks the child to associate it with a picture. The diffusion of the figures increases over time. The exam stops when the child answers six out of eight questions incorrectly. After completion, the raw score is given, normalized to a mean of 100 and standard deviation of 15. The NFP Program used PPVT when the child was 6 years old.

C.4 WISC-III

The Wechsler Intelligence Scale for Children — Third Edition (WISC-III) was created in 1949. The third edition was published in 1991 (Wechsler, 1991). WISC is an intelligence test for children between the ages 6 and 16 years old. It can be completed without reading or writing. The exam takes between 65 and 80 minutes. There are two subscales: verbal and performance, which provide a Verbal IQ (VIQ), a Performance IQ (PIQ), and a Full Scale IQ (FSIQ). The NFP only used the coding part of the Processing Speed Index:

1. *Coding: the child marks rows of shapes with different lines to transcribe a digit-symbol code. It measures visual or motor integration and visual scanning.*

The NFP Program used WISC-III when the child was 6 years old.

C.5 CBCL

The Child Behavior Checklist (CBCL) is a parent-report questionnaire developed by Thomas M. Achenbach. In it, the child is rated on several behavioral and emotional problems. The goal of the inventory is to assess internalizing and externalizing behaviors. The responses are recorded using a Likert scale: 0 = Not True, 1 = Sometimes True, 2 = Very True. The preschool checklist (18 months to 5 years) contains 100 questions and the school-age checklist (6 to 18 years) contains 120 questions. The preschool checklist questions can be broken down into the following subscales: anxious/depressed, withdrawn, sleeping problems, somatic problems, aggressive behavior, and destructive behavior. The school-age checklist questions can be broken down into the following subscales: withdrawn, somatic complaints, anxious/depressed, social problems, thought problems, attention problems, delinquent behavior, aggressive behavior, and other problems. The NFP Program used the CBCL when the child was 2 and 6 years old.

C.6 MacArthur

The MacArthur Story Stem Battery (MSSB) was created by the MacArthur Narrative Working Group that included Bretherton, Buchsbaum, and several other collaborators. The story stem method is a procedure in which the examiner presents a story to the child that culminates at a high point, at which the child is then asked to complete the story; this type of method allows insight into the inner workings of the child's mind. The MSSB uses 15 stories and measures: dysregulated aggression, empathy/warmth, emotional integration, and performance anxiety.

1. *Dysregulated Aggression Dimension: aggression, injury, danger, destruction, dishonesty, escalation of conflict, negative story endings, inappropriate child power, controlling toward examiner.*

2. *Empathy/Warmth Dimension: empathy-helping, affiliation, affection, reparation or guilt, parental warmth.*

3. *Emotional Integration Construct: ability to maintain story coherence with the inclusion of emotional expression. The affects included are joy, anger, distress, concern, sadness.*

4. *Avoidance or Withdrawal Dimension: characters leaving the scene repetition of previous story fragments, denial of central conflict or challenge, family characters leave, avoiding separation from parents, dissociative behaviors.*

5. *Performance Anxiety Dimension: unwillingness to verbalize, unresponsiveness to examiner, anxious behaviors.*

The NFP Program used the MSSB when the child was 6 years old.

D Permutation-based Inference and Multiple Hypothesis Testing

The standard model of program evaluation describes the observed outcome Y_i of participant $i \in J$ by

$$Y_i = D_i Y_{i,1} + (1 - D_i) Y_{i,0}, \tag{1}$$

where $J = \{1, \ldots, N\}$ denotes the sample space indexing set, D_i denotes the treatment assignment for participant $i \in J$, ($D_i = 1$ if treatment occurs, $D_i = 0$ otherwise) and $(Y_{i,0}, Y_{i,1})$ are potential outcomes for participant i when treatment is *fixed* at control and treatment status respectively.

Randomized experiments solve potential problems of selection bias by inducing independence between counterfactual outcomes $(Y_{i,0}, Y_{i,1})$ and treatment status D_i when conditioned on the pre-program variables $\boldsymbol{X}$ used in the randomization protocol. All variables are defined in the common probability space $(\Omega, \mathscr{F}, \mathrm{P})$. In our notation, a randomized experiment must satisfy the following assumption:

Assumption A-1. $Y(d) \perp\!\!\!\perp D \mid \boldsymbol{X};\ d \in \mathrm{supp}(D),$

where variables $\boldsymbol{X} = (\boldsymbol{X}_i;\ i \in J), D = (D_i;\ i \in J)$ are N-dimensional vectors of treatment assignments and pre-program variables, and $Y(d) = (Y_{i,d_i}; i \in J,\ d_i \in \{0,1\})$ and $d \in \text{supp}(D) = \{0,1\}^{|J|}$ denotes the vector of counterfactual outcomes. In the same fashion, we represent the vector of observed outcomes of Equation (1) by $Y = (Y_i; i \in \mathcal{I})$. The no-treatment hypothesis is equivalent to the statement that the conditional counterfactual outcome vectors share the same distribution:

Hypothesis H-1. $Y(d) \stackrel{d}{=} Y(d') \mid \boldsymbol{X}\ ;\ d, d' \in \text{supp}(D),$

Hypothesis **H-1** can be restated in more tractable form:

Hypothesis H-1′. Under Assumption **A-1** and Hypothesis **H-1**, we have that $Y \perp\!\!\!\perp D \mid \boldsymbol{X}$.

Testing Hypothesis **H-1′** poses some statistical challenges. First, small sample sizes cast doubt on inference that relies on the asymptotic behavior of test statistics. We address the problem of small sample size by generating the exact test statistic conditioned on data. Second, the presence of multiple outcomes allows for the arbitrary selection of statistically significant outcomes. Selectively reporting statistically significant outcomes is often termed *cherry picking* and generates downward-biased p-values. We solve the problem of cherry picking by implementing multiple-hypothesis testing based on the stepdown procedure of (Romano and Wolf, 2005). They explain that the stepdown procedure strongly controls for family-wise error rate (FWER), while classical tests do not. Also, Romano and Wolf (2005) show that the strong FWER control can be obtained by imposing a certain monotonicity condition on the test statistics. This requirement is weaker than the assumption of subset pivotality, used in various methods of resampling outcomes presented in Westfall and Young (1993).

To summarize, our method is based on three steps. First, we seek to characterize the exact conditional distribution of $D|\boldsymbol{X}$. Specifically we characterize the set $D_x(d)$, defined by:

$$D_x(d) = \{d' \in \{0,1\}^{|J|}\,; \text{P}(D = d|\boldsymbol{X} = x) = \text{P}(D = d'|\boldsymbol{X} = x)\},$$

such that the distribution of D conditioned on realized data is uniform among elements of $D_x(d)$. Next we use the assumption of the null hypothesis of no-treatment effects, i.e. $H_0 : Y \perp\!\!\!\perp D|\boldsymbol{X}$, to generate the exact conditional distribution of a test statistic $T(Y,D)|\boldsymbol{X}$. Under H_0, we can construct an inference that controls for the probability of falsely rejecting the null hypothesis. We control for this probability in two ways: (1) in the case of single (joint) null hypothesis, we control for the standard Type-I error; (2) in the case of multiple hypothesis inference, we control for the family-wise error rate.

More notation is helpful for describing the method. Let K represent the indexing set for all available outcomes $Y_k; k \in K$. We represent the single (joint) null hypothesis that a set $L \subset K$ of outcomes $Y_k; k \in L$ are jointly independent of treatment status D conditional on pre-program variables $\boldsymbol{X}$ by

Hypothesis H-1″. $H_L : Y_L \perp\!\!\!\perp D|\boldsymbol{X},$ where $Y_L = (Y_k : k \in L)$.

When L is a singleton, say $L = \{k\}$, then the null hypothesis is given by $H_{\{k\}} : Y_k \perp\!\!\!\perp D|\boldsymbol{X}$. In this notation, we can write the joint Hypothesis H-1″ as $H_L = \cap_{k \in L} H_{\{k\}}$.

Our goal is to test single (or joint) null hypotheses controlling for the probability of Type I error at level α, that is, P(reject $H_L|H_L$ is true) $\leq \alpha$. To do so, we rely on the fact that, under H_L,

$$(Y_L, D)|\boldsymbol{X} \overset{d}{=} (Y_L, gD)|\boldsymbol{X} \;\forall\; g \in \mathscr{G}_X, \tag{2}$$

where $\mathscr{G}_X$ comprises all the permutations within strata of $\boldsymbol{X}$, that is,

$$\mathscr{G}_X = \{g\,;\, g : J \to J \text{ is a bijection and } g(j) = j' \Rightarrow (\boldsymbol{X}_j) = (\boldsymbol{X}_{j'})\},$$

and gD is a vector defined by:

$$gD = (\tilde{D}_i \in \text{supp}(D); i \in J \text{ and } \tilde{D}_i = D_{g(i)}).$$

We use Relation (2) to generate a statistical test where the exact distribution of the test statistic $T_L(Y_L, gD)$ is obtained by re-evaluating $T_L(Y_L, gD)$ as g varies in $\mathscr{G}_X$. Note that the inference for Hypothesis **H-1″** depends on the choice of statistics. That is to say that even though any statistic $T_L(Y_L, D)$ whose value provides evidence against the null hypothesis can be used, the inference is dependent on this choice of statistic. An example of such statistic is the maximum of the t-statistic associated with the difference in means between treated and control groups over outcomes Y_k such that $k \in L$. Formally,

$$T_L(Y_L, D) = \max_{k \in L} T_k(Y_k, D), \tag{3}$$

where $T_k(Y_k, D)$ is the t-statistic for outcome Y_k. Relationship (2) implies that $T_L(Y_L, D)|\boldsymbol{X} \overset{d}{=} T_L(Y_L, gD)|\boldsymbol{X}$ for any $g \in \mathscr{G}_X$. Moreover, let $d \in \{0,1\}^{|J|}$ such that $\mathrm{P}(D = d|\boldsymbol{X} = x) > 0$, then the distribution of D conditioned on $\boldsymbol{X} = x$ is uniform across elements of $D_x(d)$ (see Lehmann and Romano (2005), Chapter 15). Thus, a critical value $c_{L,x}(Y_L, d, \alpha)$ such that $\mathrm{P}(T_L(Y_L, D) > c_{L,x}(Y_L, d, \alpha)|\boldsymbol{X} = x, H_L \text{ is true}) \leq \alpha$ can be computed as:

$$c_{L,x}(Y_L, d, \alpha) = \inf_{t \in \mathbf{R}} \left\{ \sum_{d' \in D_x(d)} I\{T_L(Y_L, d') \leq t\} \geq (1-\alpha)|D_x| \right\} ,$$

where $I\{\cdot\}$ is the indicator function. The following notation is useful to further characterize $c_{L,x}(Y_L, d, \alpha)$. Let $T_{L,x}^{(1)}, \ldots, T_{L,x}^{(|D_x(d)|)}$ be the sequence of increasing ordered statistics $T_L(Y_L, d')$ as d' varies in $D_x(d)$. In this notation we can write the critical value as

$$c_{L,x}(Y_L, d, \alpha) = T_{L,x}^{(\lceil(1-\alpha)|D_x|\rceil)} \tag{4}$$

where $\lceil a \rceil$ stands for the smallest integer bigger or equal than a.

Under the null hypothesis H_L, the probability of a test statistic be bigger or equal than the statistic $T_L(Y_L, d)$ actually observed, i.e. the p-value, is given by:

$$p_{L,x}(d) = \inf_{\alpha \in [0,1]} \Big\{ c_{L,x}(Y_L, d, \alpha) \leq T_L(Y_L, d) \Big\}. \tag{5}$$

Now let $r_{L,x} \in \{1, \ldots, |D_x(d)|\}$ be the lowest rank that the value of the observed test statistic $T_L(Y_L, d)$ takes in the sequence $T_{L,x}^{(1)}, \ldots, T_{L,x}^{(|D_x(d)|)}$, that is to say:

$$r_{L,x} = 1 + \sum_{d' \in D_x(\mathbf{d})} I\{T_L(Y_L, d') < T_L(Y_L, d)\}.$$

Thus:

$$T_{L,x}^{(r_{L,x})} = T_L(Y_L, d). \tag{6}$$

Then, by the ordered property of $T_{L,x}^{(r)}; r \in \{1, \ldots, |D_x(d)|\}$ and the definition of $r_{L,x}$, we have that:

$$p_{L,x}(d) = 1 - \frac{r_{L,x}}{|D_x(d)|}. \tag{7}$$

Moreover, p-value $p_{L,x}(d)$ complies with the following property:

$$\mathrm{P}(p_{L,x}(d) \leq \phi | \boldsymbol{X} = \boldsymbol{x}) \leq \phi \, \forall \phi \in [0, 1].$$

We implement a method of inference that tests the multiple null hypotheses that each outcome $Y_k; k \in L$ is independent of treatment status D conditional on pre-program variables $\boldsymbol{X}$. The representation of these multiple hypothesis is in the same fashion as the single (joint) null hypothesis, namely, $H_L = \cap_{k \in L} H_{\{k\}}$; $H_{\{k\}} : Y_k \perp\!\!\!\perp D | \boldsymbol{X}$. The multiple hypothesis testing differs from the single (joint) hypothesis testing in the way it controls for the probability of false rejection. Specifically, let the subset L_0 be the set of true Hypothesis $H_{\{k\}}$ such that $k \in L_0 \subset L$. Our multiple hypothesis testing controls for the family-wise error rate (FWER), that is, the probability of even one false rejection among the set of true hypothesis L_0. Formally, we control for:

$$\mathrm{P}(\text{ reject at least one } H_{\{k\}}; k \in L_0 | H_{L_0} \text{ is true }) \leq \alpha,$$

while single (joint) hypothesis testing controls for $\mathrm{P}(\text{ reject } H_L | H_L \text{ is true }) \leq \alpha$.

Bonferroni or Holm are examples of inference methods that test multiple hypothesis controlling for FWER. These methods rely upon a "least favorable" dependence structure among the p-values. The stepdown procedure of Romano and Wolf (2005) is less conservative as it accounts for the dependence structure of p-values. The method is based on a monotonicity assumption, which, in our case, can be stated as:

$$c_{K,x}(Y_K, d, \alpha) \geq c_{L_0,x}(Y_{L_0}, d, \alpha) \text{ for any subset } K \text{ of } L \text{ containing } L_0 \text{ i.e. } L_0 \subset K \subset L. \quad (8)$$

Equation (8) is satisfied by our choice of test statistic (3) and the fact that $L_0 \subset K$.

The stepdown procedure given in Romano and Wolf (2005) is a stepwise method summarized in the following algorithm:

Algorithm 1.

Step 1: Set $L_1 = L$. If

$$\max_{k \in L_1} T_k(Y_k, d) \leq c_{L,x}(Y_{L_1}, d, \alpha) \ , \quad (9)$$

then stop and reject no null hypotheses; otherwise, reject any $H_{\{k\}}$ with

$$T_k(Y_k, d) > c_{L,x}(Y_{L_1}, d, \alpha)$$

and go to Step 2.

$\vdots$

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

$$\max_{k \in L_j} T_k(Y_k, d) \leq c_{L,x}(Y_{L_j}, d, \alpha), \quad (10)$$

then stop and reject no further null hypotheses; otherwise, reject any $H_{\{k\}}$ with

$$T_k(Y_k, d) > c_{L,x}(Y_{L_j}, \mathbf{d}, \alpha)$$

and go to Step $j+1$.

$\vdots$

We can compute the multiplicity-adjusted p-values of Equations (9)–(10) in the same fashion described by Equations (5)–(7).

D.1 Conditioning and Linearity

A typical problem in small sample randomized trials is sampling variation, where pre-program variables differ across treatment groups by chance. One can increase the power of any statistical inference by conditioning on those pre-program variables. Let $\boldsymbol{Z}$ be the pre-program variables that were not used in the randomization protocol that we seek to control for.

Variables $\boldsymbol{Z}$ precede the treatment intervention and therefore $\boldsymbol{Z} \perp\!\!\!\perp D \mid \boldsymbol{X}$ holds due to randomization. Under the hypothesis of no-treatment, $\boldsymbol{Y} \perp\!\!\!\perp D \mid \boldsymbol{X}$ also holds. These two relations imply that $\boldsymbol{Y} \perp\!\!\!\perp D \mid (\boldsymbol{X}, \boldsymbol{Z})$. We can use this relationship to generate a permutation test that considers the strata formed by values of covariates $\boldsymbol{X}$ and $\boldsymbol{Z}$. This way we can generate an inference method that non-parametrically conditions on variables $\boldsymbol{X}$ and $\boldsymbol{Z}$.

Non-parametric conditioning through block permutation comes at a cost. A fine conditioning set decreases the share of available data that can be permuted and a sufficiently large conditioning set prohibits the implementation of a permutation-based test. We solve this problem by evoking linearity. That is to say, we condition variables through a linear regression instead of a non-parametric block permutation. Anderson and Legendre (1999) test a range of permutation methods for linear models. They find that the Freedman and

Lane (1983) procedure generates the most consistent and reliable results among the available models in this literature.

We non-parametrically condition on variables used in the randomization protocol to achieve valid exchangeable properties (i.e. we use permutations in $\mathscr{G}_X$); We linearly condition on additional pre-program variables $\boldsymbol{Z}$ not used in the randomization protocol. Following the Freedman and Lane (1983) method, our approach can be summarized by the following steps: (1) compute the residuals $\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}$ such that $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'Y$; (2) permute these residuals according to permutations $g \in \mathscr{G}_X$. (3) add these permuted residuals to $\boldsymbol{Z}\hat{\boldsymbol{\beta}}$, call it $\tilde{\boldsymbol{Y}}$; (4) regress $\tilde{\boldsymbol{Y}}$ on $\boldsymbol{Z}$ and the vector treatment statuses $\boldsymbol{D}$. (5) we then use the t-statistic associated with covariate $\boldsymbol{D}$ of the last regression as test statistic.

Beaton (1978) and Freedman and Lane (1983) suggest permutation inference based on Shuffle Residuals. By this, we mean regressing Y on $\boldsymbol{X}$, shuffling the residuals from this regression, and adding them to the predicted $\boldsymbol{Y}$, say $\hat{\boldsymbol{Y}}$, to form a new variable, say $\tilde{\boldsymbol{Y}}$, which is then regressed on $\boldsymbol{Z}$ and $\boldsymbol{D}$. Formally, let the regression:

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{\delta} + \epsilon,$$

where $\boldsymbol{Z}$ stands for the pre-program variables we wish to control for and includes a vector of elements ones that play the role of a contant term for the regression. Error term ϵ is a mean-zero exogenous random variable independent of $\boldsymbol{Z}$ and $\boldsymbol{D}$.

Now let $\boldsymbol{B}_g; g \in \mathscr{G}_X$ be a permutation matrix associated with a permutation g in $\mathscr{G}_X$. Let the operator that projects a vector in the orthogonal space generated by columns of $\boldsymbol{Z}$ be $\boldsymbol{M_Z} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$, where I denotes the identity matrix. As properties of Matrix $\boldsymbol{M_Z}$, we can say that $\boldsymbol{M_Z}$ is symmetric and idempotent, that is:

$$\boldsymbol{M_Z} = \boldsymbol{M_Z'} = \boldsymbol{M_Z}\boldsymbol{M_Z} = \boldsymbol{M_Z'}\boldsymbol{M_Z}. \tag{11}$$

The estimated residuals of Y generated by the the regression

$$\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\epsilon}$$

is given by $\hat{\boldsymbol{e}} = \boldsymbol{M_Z Y}$. The predicted outcome based on this regression is given by: $\hat{\boldsymbol{Y}} = \boldsymbol{Z}(\boldsymbol{Z'Z})^{-1}\boldsymbol{X'Y}$.

We define the new outcome based on the sum of the predicted outcome $\hat{\boldsymbol{Y}}$ with permuted errors $\hat{e}$ according to permutation $g \in \mathscr{G}_X$ as

$$\tilde{\boldsymbol{Y}} = \hat{\boldsymbol{Y}} + \boldsymbol{B}_g \hat{\boldsymbol{e}}. \tag{12}$$

We then use the newly computed outcome in the following regression:

$$\tilde{\boldsymbol{Y}} = \boldsymbol{Z\beta} + \boldsymbol{D\delta} + \tilde{\boldsymbol{\epsilon}}. \tag{13}$$

We now examine the δ estimate on Equation (13). This estimate ia actually the same as the one computed in the following regression:

$$\boldsymbol{M_Z}\tilde{\boldsymbol{Y}} = \boldsymbol{M_Z D\delta} + \tilde{\boldsymbol{\epsilon}}. \tag{14}$$

Thus, by applying the Ordinary Least Square formula, we obtain:

$$\hat{\boldsymbol{\delta}}_{\boldsymbol{g}} = (\boldsymbol{D'M'_Z M_Z D})^{-1}\boldsymbol{D'M'_Z M_Z}\tilde{\boldsymbol{Y}}. \tag{15}$$

We now use previous equations to transform Equation (15) into a more general formula:

$$
\begin{aligned}
\hat{\boldsymbol{\delta}}_{\boldsymbol{g}} &= (\boldsymbol{D}'\boldsymbol{M}_{\boldsymbol{Z}}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{M}_{\boldsymbol{Z}}'\boldsymbol{M}_{\boldsymbol{Z}}\tilde{\boldsymbol{Y}}, \text{ by (15)}, \\
&= (\boldsymbol{D}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{M}_{\boldsymbol{Z}}\tilde{\boldsymbol{Y}}, \text{ by (11)}, \\
&= (\boldsymbol{D}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{M}_{\boldsymbol{Z}}(\boldsymbol{Y} + \boldsymbol{B}_g\hat{\boldsymbol{e}}), \text{ by (12) }, \\
&= (\boldsymbol{D}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{M}_{\boldsymbol{Z}}((I - \boldsymbol{M}_{\boldsymbol{Z}})\boldsymbol{Y} + \boldsymbol{B}_g\hat{\boldsymbol{e}}), \text{ because } M_Z = I - Z(Z'Z)^{-1}Z', \\
&= (\boldsymbol{D}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{D})^{-1}\boldsymbol{D}'((\boldsymbol{M}_{\boldsymbol{Z}} - \boldsymbol{M}_{\boldsymbol{Z}})\boldsymbol{Y} + \boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{B}_g\hat{\boldsymbol{e}}), \\
&= (\boldsymbol{D}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{D})^{-1}\boldsymbol{D}'(\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{B}_g\hat{\boldsymbol{e}}), \\
&= (\boldsymbol{D}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{D})^{-1}\boldsymbol{D}'(\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{B}_g\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{Y}), \text{ because } \hat{\boldsymbol{e}} = \boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{Y}. \qquad (16)
\end{aligned}
$$

Kennedy (1995) points out that the Freedman and Lane (1983) algorithm is summarized by Equation (16). Notationally, we can use $T_{\boldsymbol{Z}}(\boldsymbol{Y}, g\boldsymbol{D}); g \in \mathscr{G}_X$ (instead of $T(\boldsymbol{Y}, g\boldsymbol{D}); g \in \mathscr{G}_X$) to represent the distribution of the test statistic associated with the t-statistic of the $\boldsymbol{D}$ covariate in the Freedman and Lane (1983) regression just described. Using this notation, the analysis of the previous sections holds unaltered.

E Additional Baseline Tables

Table E.1 presents the statistical description of retention levels by gender and time of survey. Table E.2 presents the statistical description of selected pre-program variables after 6 years of the program. Table E.3 presents the statistical description of selected pre-program variables after 12 years of the program.

Table E.1: Retention Rates by Gender

	All Males Groups 2 and 4		Control Males Group 2		Treated Males Group 4		Difference Groups 2 and 4
	Sample	Attrition	Sample	Attrition	Sample	Attrition	p-value
Month 6	333	0.93	232	0.93	101	0.92	0.65
Month 12	338	0.94	234	0.94	104	0.95	0.83
Year 2	339	0.94	235	0.94	104	0.95	0.95
Year 4.5	324	0.90	223	0.90	101	0.92	0.51
Year 6	323	0.90	224	0.90	99	0.90	0.99
Year 9	315	0.88	218	0.88	97	0.88	0.87
Year 12	300	0.84	202	0.81	98	0.89	0.06
	All Females Groups 2 and 4		Control Females Group 2		Treated Females Group 4		Difference Groups 2 and 4
	Sample	Attrition	Sample	Attrition	Sample	Attrition	p-value
Month 6	338	0.94	237	0.95	101	0.90	0.072
Month 12	347	0.96	239	0.96	108	0.96	0.840
Year 2	340	0.94	235	0.94	105	0.94	0.814
Year 4.5	322	0.89	220	0.88	102	0.91	0.443
Year 6	318	0.88	220	0.88	98	0.88	0.817
Year 9	312	0.86	218	0.88	94	0.84	0.354
Year 12	294	0.81	205	0.82	89	0.79	0.519

Notes: The table presents sample attrition over time. The table displays two panes, the top one focuses on data for males, the bottom one describes data for females. The first column in each panel re are The first column of each panel gives the time of survey. Each panel displays four blocks of data description. The first block on the sub-sample consisting of all data for each gender (Groups 2 and 4), the second block assess the control group (Group 2) and the third block focus on the treatment group (Group 4). First column of each of these three blocks provides the sample size and the second block presents the percentages of non-missing data. The last block presents the double-sided p-value for testing whether the the difference of non-missing percentage values between treated and control groups is statistically different than zero.

Table E.2: Descriptive Statistic of Baseline Characteristics (Year 6)

	Whole Sample					Female Sample					Male Sample				
	C Mean	C SD	T Mean	T SD	Pval	C Mean	C SD	T Mean	T SD	Pval	C Mean	C SD	T Mean	T SD	Pval
Background Characteristics															
Maternal Race (Black)	0.060	0.238	0.100	0.301	**0.100**	0.067	0.251	0.081	0.274	0.668	0.054	0.226	0.119	0.325	**0.070**
Marital Status (Married)	0.016	0.124	0.015	0.122	0.952	0.009	0.094	0.010	0.101	0.922	0.022	0.148	0.020	0.140	0.883
Maternal Age	18.060	3.220	18.060	3.294	0.999	18.219	3.299	18.152	3.607	0.874	17.902	3.138	17.970	2.971	0.850
Years of Education	10.263	1.881	10.120	2.024	0.395	10.313	1.841	10.081	2.069	0.339	10.214	1.922	10.158	1.989	0.813
Mother in School	0.609	0.489	0.580	0.495	0.497	0.570	0.496	0.616	0.489	0.432	0.647	0.479	0.545	0.500	**0.084**
Head of Household is Employed	0.562	0.497	0.492	0.501	0.106	0.605	0.490	0.475	0.502	**0.031**	0.518	0.501	0.510	0.502	0.897
% of Census Tract Below Poverty	34.812	21.371	35.518	20.221	0.687	33.195	20.304	36.724	22.248	0.179	36.428	22.316	34.336	18.049	0.371
Household Density	0.940	0.497	1.027	0.569	**0.064**	0.961	0.499	1.070	0.669	0.151	0.920	0.495	0.986	0.451	0.236
Total Household Income (Past 6 Months)															
Less than $3000	0.283	0.451	0.365	0.483	**0.044**	0.290	0.455	0.364	0.483	0.202	0.277	0.448	0.366	0.484	0.116
$3000 - $6999	0.237	0.425	0.225	0.419	0.746	0.219	0.414	0.222	0.418	0.945	0.254	0.437	0.228	0.421	0.601
$7000 - $10999	0.228	0.420	0.205	0.405	0.515	0.219	0.414	0.222	0.418	0.945	0.237	0.426	0.188	0.393	0.317
Greater than $11000	0.161	0.368	0.125	0.332	0.222	0.188	0.391	0.081	0.274	**0.005**	0.134	0.341	0.168	0.376	0.434
Income, No Response	0.092	0.289	0.080	0.272	0.625	0.085	0.279	0.111	0.316	0.476	0.098	0.298	0.050	0.218	**0.099**
Region of Residence															
Inner City	0.295	0.456	0.290	0.455	0.905	0.286	0.453	0.303	0.462	0.755	0.304	0.461	0.277	0.450	0.628
Bisson	0.192	0.394	0.215	0.412	0.506	0.179	0.384	0.232	0.424	0.282	0.205	0.405	0.198	0.400	0.879
Cawthon	0.194	0.396	0.190	0.393	0.900	0.210	0.408	0.162	0.370	0.297	0.179	0.384	0.218	0.415	0.420
Hollywood	0.319	0.467	0.305	0.462	0.719	0.326	0.470	0.303	0.462	0.684	0.313	0.465	0.307	0.464	0.920
Maternal Mental Health															
Maternal IQ (Shipley)	96.270	10.287	96.440	10.360	0.847	96.223	10.279	96.061	10.618	0.898	96.317	10.317	96.812	10.140	0.686
Maternal Bavolek Score	99.794	7.657	101.133	8.502	**0.057**	100.091	7.411	101.431	8.727	0.186	99.499	7.899	100.842	8.309	0.173
Maternal Mental Health	100.184	9.979	99.447	10.352	0.398	99.717	9.777	99.741	10.172	0.984	100.649	10.178	99.158	10.568	0.235
Self-Efficacy	100.083	10.017	99.788	9.866	0.727	100.862	9.778	100.583	9.253	0.806	99.307	10.212	99.008	10.419	0.810
Maternal Mastery	100.065	10.213	99.535	9.992	0.537	99.879	10.155	99.173	10.246	0.568	100.250	10.290	99.891	9.774	0.764
Maternal Psychological Resources	100.060	10.045	99.533	10.649	0.554	100.030	9.711	99.619	10.634	0.743	100.090	10.390	99.448	10.715	0.615
Maternal Health Characteristics															
Maternal Height	164.557	7.253	164.064	6.569	0.397	164.331	7.404	164.472	6.546	0.865	164.781	7.108	163.651	6.601	0.170
Pre-Pregnancy Weight	62.097	14.866	62.339	13.588	0.839	62.828	13.775	61.394	12.375	0.355	61.362	15.885	63.264	14.683	0.294
Gestational Age (Intake)	16.560	5.794	16.630	5.728	0.887	16.402	5.746	16.364	5.596	0.955	16.719	5.850	16.891	5.870	0.807
Maternal Social Support															
Grandmother Social Support	100.197	9.474	101.517	8.566	**0.081**	99.357	10.486	101.434	9.100	**0.073**	101.034	8.285	101.599	8.054	0.563
Husband/Boyfriend Social Support	100.030	9.994	100.704	9.754	0.421	99.892	10.057	99.907	9.524	0.990	100.169	9.952	101.484	9.960	0.272
Maternal Risky Behaviors															
Alcohol Consumption (Past 2 wks)	0.043	0.202	0.050	0.218	0.680	0.036	0.186	0.071	0.258	0.228	0.049	0.217	0.030	0.171	0.385
Smoking (Past 3 days)	0.085	0.279	0.110	0.314	0.334	0.081	0.273	0.121	0.328	0.284	0.089	0.286	0.099	0.300	0.784
Used Marijuana (Past 2 wks)	0.034	0.309	0.070	0.860	0.560	0.027	0.283	0.020	0.201	0.809	0.040	0.332	0.119	1.194	0.51'
Used Cocaine (Past 2 wks)	0.007	0.142	0.000	0.000	0.318	0.000	0.000	0.000	0.000	.	0.013	0.200	0.000	0.000	0.318
Sexually Transmitted Diseases	0.333	0.472	0.375	0.485	0.301	0.330	0.471	0.354	0.480	0.688	0.335	0.473	0.396	0.492	0.294

Notes: This table presents the statistical description of selected pre-program variables after 6 years of the program. The first column of the table gives the variable description. Variables are divided into groups that share similar meanings. The remainder of the table consists of the description of the blocks of variables associated with the whole sample, the female sample and the male sample. Each block has 6 columns: (1) Control mean (C Mean), (2) Control standard deviation (C SD), (3) Treatment mean (T Mean), (4) Treatment standard deviation (T SD), and (5) Asymptotic p-value associated with the difference in means. Bold p-values indicate that the t-statistic between the control and the treatment means is significant at the 10% level.

Table E.3: Descriptive Statistic of Baseline Characteristics (Year 12)

	Whole Sample					Female Sample					Male Sample				
	C Mean	C SD	T Mean	T SD	Pval	C Mean	C SD	T Mean	T SD	Pval	C Mean	C SD	T Mean	T SD	Pval
Background Characteristics															
Maternal Race (Black)	0.057	0.232	0.084	0.278	0.244	0.056	0.231	0.065	0.248	0.770	0.057	0.233	0.101	0.303	0.208
Marital Status (Married)	0.014	0.119	0.010	0.102	0.690	0.005	0.069	0.011	0.104	0.603	0.024	0.153	0.010	0.101	0.346
Maternal Age	18.052	3.215	18.047	3.268	0.986	18.258	3.324	18.174	3.581	0.847	17.842	3.093	17.929	2.960	0.812
Years of Education	10.254	1.860	10.073	2.025	0.296	10.324	1.828	10.043	2.080	0.265	10.182	1.893	10.101	1.982	0.735
Mother in School	0.599	0.491	0.565	0.497	0.444	0.557	0.498	0.598	0.493	0.505	0.641	0.481	0.535	0.501	**0.081**
Head of Household is Employed	0.556	0.497	0.495	0.501	0.163	0.585	0.494	0.478	0.502	**0.089**	0.526	0.501	0.510	0.502	0.793
% of Census Tract Below Poverty	34.800	21.380	35.727	20.185	0.606	33.632	20.150	37.208	22.390	0.189	35.990	22.550	34.351	17.900	0.492
Household Density	0.940	0.486	1.023	0.559	**0.081**	0.969	0.483	1.049	0.662	0.299	0.911	0.488	0.998	0.445	0.123
Total Household Income (Past 6 Months)															
Less than $3000	0.280	0.449	0.361	0.482	**0.048**	0.277	0.449	0.337	0.475	0.305	0.282	0.451	0.384	0.489	**0.083**
$3000 - $6999	0.242	0.429	0.236	0.425	0.870	0.239	0.428	0.239	0.429	0.995	0.244	0.431	0.232	0.424	0.822
$7000 - $10999	0.230	0.421	0.188	0.392	0.238	0.230	0.422	0.217	0.415	0.808	0.230	0.422	0.162	0.370	0.151
Greater than $11000	0.159	0.366	0.126	0.332	0.269	0.178	0.384	0.087	0.283	**0.022**	0.139	0.347	0.162	0.370	0.606
Income, No Response	0.090	0.287	0.089	0.285	0.967	0.075	0.264	0.120	0.326	0.251	0.105	0.308	0.061	0.240	0.166
Region of Residence															
Inner City	0.291	0.455	0.283	0.452	0.825	0.282	0.451	0.293	0.458	0.836	0.301	0.460	0.273	0.448	0.603
Bisson	0.194	0.396	0.225	0.419	0.391	0.169	0.376	0.239	0.429	0.176	0.220	0.415	0.212	0.411	0.874
Cawthon	0.204	0.403	0.188	0.392	0.657	0.221	0.416	0.152	0.361	0.148	0.187	0.391	0.222	0.418	0.477
Hollywood	0.310	0.463	0.304	0.461	0.867	0.329	0.471	0.315	0.467	0.819	0.292	0.456	0.293	0.457	0.985
Maternal Mental Health															
Maternal IQ (Shipley)	96.066	9.987	96.759	10.181	0.433	96.075	10.002	96.011	10.789	0.961	96.057	9.997	97.455	9.585	0.240
Maternal Bavolek Score	99.947	7.604	101.078	8.568	0.118	100.190	7.489	101.427	8.718	0.238	99.701	7.729	100.754	8.459	0.296
Maternal Mental Health	100.106	9.744	99.550	10.612	0.538	99.766	9.529	99.909	10.429	0.911	100.451	9.968	99.216	10.821	0.339
Self-Efficacy	99.813	9.995	99.671	9.912	0.870	100.640	9.746	100.192	9.339	0.705	98.973	10.197	99.186	10.440	0.866
Maternal Mastery	100.059	10.236	99.446	10.098	0.489	99.954	10.301	99.085	10.533	0.507	100.165	10.193	99.781	9.718	0.751
Maternal Psychological Resources	99.857	9.652	99.664	10.914	0.834	99.947	9.401	99.537	10.896	0.754	99.765	9.923	99.782	10.984	0.990
Maternal Health Characteristics															
Maternal Height	164.595	7.349	164.303	6.680	0.630	164.297	7.483	164.904	6.664	0.485	164.896	7.217	163.732	6.680	0.170
Pre-Pregnancy Weight	62.398	15.149	62.735	13.786	0.786	63.078	14.076	61.880	12.369	0.458	61.701	16.178	63.530	15.003	0.332
Gestational Age (Intake)	16.474	5.830	16.607	5.639	0.789	16.235	5.791	16.228	5.489	0.993	16.718	5.873	16.960	5.780	0.733
Maternal Social Support															
Grandmother Social Support	100.407	9.331	101.623	8.406	0.110	99.624	10.361	101.370	9.306	0.148	101.200	8.102	101.858	7.514	0.485
Husband/Boyfriend Social Support	100.266	9.951	100.299	9.986	0.969	100.023	9.949	99.833	9.700	0.877	100.511	9.971	100.731	10.275	0.859
Maternal Risky Behaviors															
Alcohol Consumption (Past 2 wks)	0.040	0.197	0.047	0.212	0.710	0.038	0.191	0.065	0.248	0.345	0.043	0.203	0.030	0.172	0.568
Smoking (Past 3 days)	0.081	0.273	0.105	0.307	0.356	0.075	0.265	0.120	0.326	0.255	0.086	0.281	0.091	0.289	0.891
Used Marijuana (Past 2 wks)	0.036	0.318	0.073	0.880	0.566	0.028	0.291	0.022	0.209	0.824	0.043	0.344	0.121	1.206	0.528
Used Cocaine (Past 2 wks)	0.007	0.146	0.000	0.000	0.318	0.000	0.000	0.000	0.000	.	0.014	0.208	0.000	0.000	0.318
Sexually Transmitted Diseases	0.347	0.477	0.372	0.485	0.554	0.335	0.473	0.337	0.475	0.972	0.359	0.481	0.404	0.493	0.450

Notes: This table presents the statistical description of selected pre-program variables after 6 years of the program. The first column of the table gives the variable description. Variables are divided into groups that share similar meanings. The remainder of the table consists of the description of the blocks of variables associated with the whole sample, the female sample and the male sample. Each block has 6 columns: (1) Control mean (C Mean), (2) Control standard deviation (C SD), (3) Treatment mean (T Mean), (4) Treatment standard deviation (T SD), and (5) Asymptotic p-value associated with the difference in means. Bold p-values indicate that the t-statistic between the control and the treatment means is significant at the 10% level.

F Additional Inference Results: Unconditional Analysis and Addressing Attrition using Inverse Propensity Weights

Tables F.1–F.5 present the unconditional analysis of the treatment effects presented in Tables 6–10 of the main paper.

One aspect of the NFP that may cause concern is attrition. In order to address this issue, we use statistical models that account for missing data by reweighting observations according to the inverse probability of retention, which is usually termed Inverse Probability Weighting (IPW). The probabilities of attrition at each wave are estimated by gender using logit models. To select the covariates in the model, we choose the set of pre-program covariates that minimize the Akaike Information Criterion (AIC). Then, we use the estimated probabilities to reweight the observations and compute the treatment effects. The results do not change much after this correction. Tables F.7–F.10 show these results. The tables can be read in the same way as Tables 6–10 in the paper.

Table F.1: Child Health Outcomes (Unconditional Effects)

Outcome Description	Female Sample						Male Sample					
	Basic Statistics				Unrestricted Permutation		Basic Statistics				Unrestricted Permutation	
	Control Mean	Difference in Means	Effect Size	Asymp. p-value	Single p-value	Stepdown p-value	Control Mean	Difference in Means	Effect Size	Asymp. p-value	Single p-value	Stepdown p-value
Birth Outcomes for Child												
Placenta Weight	682.995	-2.290	-0.014	0.544	0.541	0.782	663.819	21.423	0.113	0.168	0.149	0.149
Birth Weight	3055.224	-121.268	-0.219	0.961	0.952	0.952	2997.486	199.869	0.275	**0.006**	**0.006**	**0.025**
Head Circumference	33.262	0.023	0.013	0.454	0.452	0.771	33.511	0.352	0.151	**0.088**	**0.080**	0.142
Length	49.665	0.208	0.076	0.259	0.253	0.595	49.918	0.567	0.150	**0.083**	**0.061**	0.145
Gestational Age at Delivery	39.119	-0.415	-0.186	0.899	0.870	0.935	38.544	0.783	0.220	**0.019**	**0.019**	**0.062**
Child Health Outcomes (Year 12)												
Any Injuries Since Last Interview	-0.164	0.065	0.176	**0.069**	**0.064**	0.162	-0.224	0.061	0.146	0.109	0.109	0.385
# Hospitalizations for Injuries Since Last Interview	-0.009	0.009	0.097	0.178	0.404	0.404	-0.010	0.010	0.099	0.164	**0.023**	0.141
Total # Injuries Since Last Interview	-0.197	0.098	0.199	**0.039**	**0.029**	**0.089**	-0.268	0.054	0.099	0.209	0.222	0.610
Hospitalized Since Last Interview	-0.042	0.042	0.210	**0.023**	**0.027**	0.111	-0.039	-0.033	-0.171	0.892	0.871	0.995
Have Chronic Condition/Health Problem	-0.197	0.012	0.031	0.401	0.404	0.639	-0.361	-0.072	-0.150	0.886	0.886	0.986
Standardized Child BMI	-1.121	0.276	0.308	**0.007**	**0.007**	**0.034**	-0.797	-0.163	-0.179	0.921	0.917	0.917

Note: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Control Mean) of each result set shows the unconditional mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Difference in Means) gives the unconditional difference in means between the treatment group and the control group. As mentioned in Section 2, the control group stands for the 2 of the NFP experiment and the treatment group stands for the original group 4. The third column (Effect Size) presents the unconditional effect size for the respective group. The fourth column (Asymp. p-value) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the unconditional difference in means between treatment and control groups. The fifth column (Unrestricted Permutation – Single p-value) presents the one-sided unrestricted permutation p-values for the single-hypothesis testing based on the t-statistic associated with the treatment indicator. Finally, the last column (Unrestricted Permutation – Stepdown) provides p-values that account for multiple-hypothesis testing based on the Stepdown procedure of Romano and Wolf (2005). Blocks of outcomes that are tested jointly are separated by lines. The selection of blocks of outcomes is done on the basis of their meaning. Outcomes that share similar meaning are grouped together. Female maternal outcomes allude to mothers whose first child is a girl. Likewise, male maternal outcomes alludes to mothers whose first child is a boy.

Table F.2: Family Environment (Unconditional Effects)

Outcome Description	Female Sample						Male Sample					
	Basic Statistics				Unrestricted Permutation		Basic Statistics				Unrestricted Permutation	
	Control Mean	Difference in Means	Effect Size	Asymp. p-value	Single p-value	Stepdown p-value	Control Mean	Difference in Means	Effect Size	Asymp. p-value	Single p-value	Stepdown p-value
Home Environment, Parenting (Year 1)												
Family Environment (HOME Score)	0.000	0.354	0.354	**0.002**	**0.002**	**0.004**	0.000	0.207	0.207	**0.044**	**0.044**	**0.044**
Non-Abusive Parenting Attitudes (Bavolek)	0.000	0.289	0.289	**0.009**	**0.012**	**0.012**	0.000	0.274	0.274	**0.012**	**0.012**	**0.024**
Home Environment, Parenting (Year 2)												
Family Environment (HOME Score)	0.000	0.302	0.302	**0.007**	**0.008**	**0.008**	0.000	0.170	0.170	**0.082**	**0.087**	**0.087**
Non-Abusive Parenting Attitudes (Bavolek)	0.000	0.371	0.371	**0.002**	**0.004**	**0.008**	0.000	0.316	0.316	**0.004**	**0.004**	**0.009**
Maternal Mental Health (Year 2) - Factor Scores												
Anxiety	0.000	0.247	0.247	**0.024**	**0.030**	**0.072**	0.000	0.038	0.038	0.376	0.382	0.561
Depression	0.000	0.129	0.129	0.139	0.137	0.221	0.000	0.062	0.062	0.302	0.306	0.516
Positive Well-Being	0.000	0.101	0.101	0.197	0.199	0.199	0.000	-0.136	-0.136	0.865	0.859	0.859
Emotional Stability	0.000	0.207	0.207	**0.047**	**0.052**	0.106	0.000	0.050	0.050	0.340	0.341	0.526
Overall Mental Health	0.000	0.210	0.210	**0.043**	**0.046**	**0.098**	0.000	-0.014	-0.014	0.546	0.554	0.658
Self-Esteem	0.000	0.313	0.313	**0.005**	**0.006**	**0.024**	0.000	0.073	0.073	0.282	0.295	0.563
Mastery	0.000	0.286	0.286	**0.012**	**0.019**	**0.058**	0.000	0.198	0.198	**0.053**	**0.059**	0.182
Welfare (Child Ages 1 - 12 Years)												
AFDC/TANF	-2744.043	-46.414	-0.017	0.557	0.549	0.549	-2743.386	439.910	0.159	**0.074**	**0.074**	**0.074**
Food Stamp	-2996.965	164.090	0.089	0.218	0.208	0.303	-3263.273	347.770	0.202	**0.039**	**0.038**	**0.072**
Medicaid	-3543.761	167.036	0.090	0.217	0.211	0.286	-3823.048	317.413	0.191	**0.048**	**0.045**	**0.073**

Note: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Control Mean) of each result set shows the unconditional mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Difference in Means) gives the unconditional difference in means between the treatment group and the control group. As mentioned in Section 2, the control group stands for the 2 of the NFP experiment and the treatment group stands for the original group 4. The third column (Effect Size) presents the unconditional effect size for the respective group. The fourth column (Asymp. p-value) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the unconditional difference in means between treatment and control groups. The fifth column (Unrestricted Permutation – Single p-value) presents the one-sided unrestricted permutation p-values for the single-hypothesis testing based on the t-statistic associated with the treatment indicator. Finally, the last column (Unrestricted Permutation – Stepdown) provides p-values that account for multiple-hypothesis testing based on the Stepdown procedure of Romano and Wolf (2005). Blocks of outcomes that are tested jointly are separated by lines. The selection of blocks of outcomes is done on the basis of their meaning. Outcomes that share similar meaning are grouped together. Female maternal outcomes allude to mothers whose first child is a girl. Likewise, male maternal outcomes alludes to mothers whose first child is a boy.

Table F.3: Maternal cumulative subsequent births (Unconditional Analysis)

Outcome Description	Female Sample						Male Sample					
	Basic Statistics				Unrestricted Permutation		Basic Statistics				Unrestricted Permutation	
	Control Mean	Difference in Means	Effect Size	Asymp. p-value	Single p-value	Stepdown p-value	Control Mean	Difference in Means	Effect Size	Asymp. p-value	Single p-value	Stepdown p-value
Cumulative Subsequent Births (Years 2 - 12)												
Subsequent Children Birth (Years 9 - 12)	-0.259	-0.044	-0.073	0.713	0.725	0.918	-0.397	0.014	0.020	0.437	0.446	0.659
Subsequent Children Birth (Years 6 - 9)"	-0.344	-0.177	-0.300	0.991	0.992	0.992	-0.459	-0.046	-0.064	0.699	0.700	0.700
Subsequent Children Birth (Years 2 - 6)"	-0.884	-0.034	-0.040	0.630	0.627	0.939	-1.027	0.138	0.159	**0.087**	**0.079**	0.206
Subsequent Children Birth (Years 0 - 2)"	-0.298	0.050	0.110	0.171	0.173	0.532	-0.315	0.105	0.226	**0.023**	**0.023**	**0.090**

Note: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Control Mean) of each result set shows the unconditional mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Difference in Means) gives the unconditional difference in means between the treatment group and the control group. As mentioned in Section 2, the control group stands for the 2 of the NFP experiment and the treatment group stands for the original group 4. The third column (Effect Size) presents the unconditional effect size for the respective group. The fourth column (Asymp. p-value) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the unconditional difference in means between treatment and control groups. The fifth column (Unrestricted Permutation – Single p-value) presents the one-sided unrestricted permutation p-values for the single-hypothesis testing based on the t-statistic associated with the treatment indicator. Finally, the last column (Unrestricted Permutation – Stepdown) provides p-values that account for multiple-hypothesis testing based on the Stepdown procedure of Romano and Wolf (2005). Blocks of outcomes that are tested jointly are separated by lines. The selection of blocks of outcomes is done on the basis of their meaning. Outcomes that share similar meaning are grouped together. Female maternal outcomes allude to mothers whose first child is a girl. Likewise, male maternal outcomes alludes to mothers whose first child is a boy.

Table F.4: Cognitive Abilities and Achievement Outcomes (Unconditional Effects)

Outcome Description	Female Sample						Male Sample					
	Basic Statistics				Unrestricted Permutation		Basic Statistics				Unrestricted Permutation	
	Control Mean	Difference in Means	Effect Size	Asymp. p-value	Single p-value	Stepdown p-value	Control Mean	Difference in Means	Effect Size	Asymp. p-value	Single p-value	Stepdown p-value
Kaufman Assessment for Children (Year 6)												
Gestalt Closure	8.981	0.172	0.055	0.330	0.332	0.690	9.775	-0.456	-0.154	0.888	0.885	0.885
Hand Movements	9.282	0.351	0.158	0.106	0.110	0.454	9.287	0.043	0.019	0.438	0.447	0.763
Matrix Analogies	8.632	0.082	0.045	0.356	0.355	0.643	8.478	0.170	0.105	0.204	0.204	0.573
Number Recall	9.423	0.169	0.058	0.320	0.323	0.719	8.952	0.659	0.264	**0.029**	**0.041**	0.222
Photo Series	6.967	0.343	0.156	0.108	0.112	0.423	6.774	-0.063	-0.028	0.589	0.586	0.787
Spatial Memory	8.434	0.117	0.047	0.357	0.361	0.537	8.526	0.276	0.110	0.195	0.199	0.601
Triangles	8.868	0.193	0.082	0.253	0.255	0.672	9.120	0.078	0.032	0.400	0.401	0.773
Word Order	9.693	-0.193	-0.069	0.713	0.709	0.709	9.191	0.542	0.203	**0.059**	**0.063**	0.290
Kaufman Assessment for Children (Year 6)												
Nonverbal	89.203	1.409	0.148	0.127	0.144	0.253	89.244	0.778	0.079	0.269	0.275	0.333
Sequential Processing	96.582	0.714	0.054	0.335	0.341	0.341	94.507	2.339	0.193	**0.070**	**0.080**	0.158
Simultaneous Processing	88.844	1.268	0.117	0.180	0.189	0.300	89.919	0.180	0.017	0.447	0.446	0.446
WISC-III, PPVT-III for Children (Year 6)												
Wechsler Intelligence Scale (WISC-III)	96.256	1.091	0.059	0.312	0.309	0.492	90.657	0.321	0.018	0.443	0.441	0.441
Peabody Picture Vocabulary Test (PPVT-III)	83.299	0.508	0.040	0.373	0.373	0.373	82.466	1.534	0.135	0.158	0.176	0.292
Child Cognition (Year 6) - Factor Scores												
Cognition + Achievement (KABC, PPVT, WISC)	0.000	0.109	0.109	0.197	0.209	0.209	0.000	0.187	0.187	**0.074**	**0.076**	**0.076**
Cognitive skills (Mental Processing KABC)	0.000	0.119	0.119	0.174	0.186	0.242	0.000	0.270	0.270	**0.019**	**0.021**	**0.029**
Reading Achievement for the Child (Year 12)												
Average Reading Grade (Grades 1 - 5)	2.703	-0.028	-0.036	0.606	0.607	0.840	2.327	0.106	0.133	0.159	0.153	0.329
TCAP % Language (School Years 1 - 5, Grd 3+)	50.854	-2.399	-0.099	0.760	0.756	0.879	38.063	5.143	0.224	**0.055**	**0.053**	0.161
TCAP % Reading (School Years 1 - 5, Grd 3+)	41.607	-1.699	-0.080	0.717	0.712	0.870	34.912	2.116	0.099	0.236	0.234	0.391
PIAT Total Reading (Derived Score)	90.246	-0.405	-0.040	0.619	0.615	0.812	89.292	1.158	0.084	0.250	0.242	0.300
PIAT Reading Comprehension (Derived Score)	88.307	-1.091	-0.114	0.811	0.805	0.805	87.585	2.108	0.172	**0.091**	**0.092**	0.230
PIAT Reading Recognition (Derived Score)	94.221	0.870	0.069	0.306	0.310	0.583	92.456	0.752	0.050	0.344	0.339	0.339
Math Achievement for the Child (Year 12)												
Average Math Grade (Grades 1 - 5)	2.634	-0.021	-0.025	0.577	0.583	0.734	2.368	0.149	0.184	**0.078**	**0.071**	0.119
TCAP % Math (School Years 1 - 5, Grd 3+)	46.935	-0.115	-0.005	0.514	0.513	0.724	40.346	3.749	0.161	0.126	0.128	0.128
PIAT Mathematics (Derived Score)	87.188	-0.790	-0.080	0.728	0.724	0.724	86.316	2.102	0.198	**0.062**	**0.065**	0.150

Note: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Control Mean) of each result set shows the unconditional mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Difference in Means) gives the unconditional difference in means between the treatment group and the control group. As mentioned in Section 2, the control group stands for the 2 of the NFP experiment and the treatment group stands for the original group 4. The third column (Effect Size) presents the unconditional effect size for the respective group. The fourth column (Asymp. p-value) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the unconditional difference in means between treatment and control groups. The fifth column (Unrestricted Permutation – Single p-value) presents the one-sided unrestricted permutation p-values for the single-hypothesis testing based on the t-statistic associated with the treatment indicator. Finally, the last column (Unrestricted Permutation – Stepdown) provides p-values that account for multiple-hypothesis testing based on the Stepdown procedure of Romano and Wolf (2005). Blocks of outcomes that are tested jointly are separated by lines. The selection of blocks of outcomes is done on the basis of their meaning. Outcomes that share similar meaning are grouped together. Female maternal outcomes allude to mothers whose first child is a girl. Likewise, male maternal outcomes alludes to mothers whose first child is a boy.

Table F.5: Socio-emotional Outcomes (Unconditional Effects)

Outcome Description	Female Sample: Basic Statistics: Control Mean	Female Sample: Basic Statistics: Difference in Means	Female Sample: Basic Statistics: Effect Size	Female Sample: Asymp. p-value	Female Sample: Unrestricted Permutation: Single p-value	Female Sample: Unrestricted Permutation: Stepdown p-value	Male Sample: Basic Statistics: Control Mean	Male Sample: Basic Statistics: Difference in Means	Male Sample: Basic Statistics: Effect Size	Male Sample: Asymp. p-value	Male Sample: Unrestricted Permutation: Single p-value	Male Sample: Unrestricted Permutation: Stepdown p-value
Child Behavior Checklist (Year 2) - Factor Scores												
Affective Problems	0.000	0.336	0.336	**0.001**	**0.000**	**0.002**	0.000	-0.163	-0.163	0.903	0.903	0.903
Anxiety Problems	0.000	0.191	0.191	**0.048**	**0.040**	**0.040**	0.000	0.029	0.029	0.407	0.421	0.758
Pervasion Developmental Problems	0.000	0.262	0.262	**0.010**	**0.007**	**0.023**	0.000	-0.084	-0.084	0.757	0.763	0.924
Attention Deficit Hyperactivity Disorder	0.000	0.239	0.239	**0.021**	**0.016**	**0.041**	0.000	-0.078	-0.078	0.734	0.736	0.935
Oppositional Defiant Problems	0.000	0.224	0.224	**0.029**	**0.026**	**0.047**	0.000	-0.113	-0.113	0.832	0.847	0.947
Child Behavior Checklist (Year 6) - Factor Scores												
Affective Problems	0.000	0.063	0.063	0.310	0.331	0.631	0.000	0.115	0.115	0.164	0.161	0.503
Anxiety Problems	0.000	0.085	0.085	0.230	0.220	0.515	0.000	-0.106	-0.106	0.797	0.797	0.797
Somatic Problems	0.000	-0.084	-0.084	0.745	0.739	0.739	0.000	-0.061	-0.061	0.684	0.692	0.876
Attention Deficit Hyperactivity Problems	0.000	0.269	0.269	**0.013**	**0.013**	**0.056**	0.000	0.053	0.053	0.332	0.323	0.690
Oppositional Defiant Problems	0.000	0.031	0.031	0.398	0.400	0.617	0.000	0.103	0.103	0.208	0.220	0.578
Conduct Problems	0.000	0.269	0.269	**0.010**	**0.009**	**0.044**	0.000	0.021	0.021	0.430	0.441	0.766
MacArthur (Year 6) - Factor Scores												
Dysregulated Aggression	0.000	0.040	0.040	0.366	0.354	0.798	0.000	0.179	0.179	**0.096**	0.126	0.465
Warmth and Empathy	0.000	0.360	0.360	**0.003**	**0.004**	**0.017**	0.000	-0.097	-0.097	0.779	0.786	0.786
Emotional Integration	0.000	-0.045	-0.045	0.640	0.639	0.639	0.000	0.022	0.022	0.433	0.433	0.851
Performance Anxiety	0.000	0.037	0.037	0.375	0.359	0.649	0.000	-0.051	-0.051	0.644	0.622	0.903
Aggression	0.000	0.182	0.182	**0.057**	**0.050**	0.164	0.000	0.151	0.151	0.133	0.157	0.510
Internalizing, Externalizing, Absences (Year 12)												
Presence of Internalizing Disorders	-0.239	0.046	0.107	0.197	0.202	0.480	-0.403	0.099	0.201	**0.053**	**0.052**	**0.097**
Presence of Externalizing Disorders	-0.182	0.023	0.060	0.317	0.328	0.545	-0.187	-0.069	-0.177	0.908	0.907	0.907
Average # of Absences (School Years 1 - 5)	-10.186	-0.996	-0.131	0.844	0.845	0.845	-11.803	1.941	0.247	**0.021**	**0.022**	**0.061**

Note: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Control Mean) of each result set shows the unconditional mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Difference in Means) gives the unconditional difference in means between the treatment group and the control group. As mentioned in Section 2, the control group stands for the 2 of the NFP experiment and the treatment group stands for the original group 4. The third column (Effect Size) presents the unconditional effect size for the respective group. The fourth column (Asymp. p-value) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the unconditional difference in means between treatment and control groups. The fifth column (Unrestricted Permutation – Single p-value) presents the one-sided unrestricted permutation p-values for the single-hypothesis testing based on the t-statistic associated with the treatment indicator. Finally, the last column (Unrestricted Permutation – Stepdown) provides p-values that account for multiple-hypothesis testing based on the Stepdown procedure of Romano and Wolf (2005). Blocks of outcomes that are tested jointly are separated by lines. The selection of blocks of outcomes is done on the basis of their meaning. Outcomes that share similar meaning are grouped together. Female maternal outcomes allude to mothers whose first child is a girl. Likewise, male maternal outcomes alludes to mothers whose first child is a boy.

Table F.6: Using Logit to Obtain the Inverse Probability Weights

Sample	Psy. Res.	Pct. Pov.	Smoker	Drinker	Educ.	Mom Support	Hus./BF Support	Race	HH Emp.	Age	Age Squared	STDs	# STDs	Mental Health	Mastery	Bavolek	Income (5 Cat.)	Gest. Wks.	Height	Pre. Preg. Wt.	Schooling	LR Chi2	Prob. > Chi2	AIC	BIC
Treatment Females																									
Year 12	x				x					x	x	x		x	x	x	x	x	x	x	x	25.79	0.06	45.26	90.70
Year 9	x					x				x	x				x	x	x		x	x	x	22.53	0.05	51.39	88.81
Year 6					x	x	x	x		x						x			x	x	x	17.62	0.02	30.361	54.417
Year 4.5		x			x	x		x		x						x			x	x	x	16.67	0.05	21.291	48.019
Year 2	x				x	x				x	x	x	x	x	x	x						24.18	0.01	-20.92	8.79
Control Females																									
Year 12	x	x				x		x		x		x			x					x		20.31	0.01	205.54	237.16
Year 9		x				x		x		x	x	x							x	x		19.29	0.01	122.33	153.81
Year 6							x	x		x	x	x				x			x			17.84	0.01	103.72	131.73
Year 4.5					x					x	x		x		x			x			x	14.68	0.02	114.54	139.16
Year 2										x	x			x	x		x			x	x	10.00	0.12	-19.62	5.01
Treatment Males																									
Year 12					x			x				x	x	x		x	x	x	x		x	22.97	0.04	106.90	144.96
Year 9								x		x	x			x								9.38	0.05	82.55	96.14
Year 6					x			x		x	x		x	x		x			x			17.66	0.02	62.91	87.37
Year 4.5		x			x			x		x	x	x		x	x				x	x	x	14.03	0.23	46.04	78.66
Year 2						x						x		x		x			x	x		15.73	0.02	-3.54	15.49
Control Males																									
Year 12	x				x	x		x	x	x							x		x	x		32.07	0.00	155.12	200.37
Year 9					x			x			x			x			x		x	x		37.02	0.00	95.72	134.01
Year 6					x			x			x				x		x		x	x	x	34.24	0.00	68.20	109.91
Year 4.5			x	x		x		x	x			x		x	x	x			x			22.03	0.02	121.86	160.14
Year 2										x	x			x	x							12.38	0.01	-24.20	-6.63

Note: The table describes the pre-program variables used to calculate the inverse probability weights. The first column provides the four division groups: treatment females, control females, treatment males, and control males. Additionally, there is a corresponding time period for each row. The next 23 columns represents the set of pre-program characteristics that were used for logit. A "x" represents that that variable was used for the specific sample and time period. The column labeled "LR Chi2" is the chi-squared calculated using the logit regression and the next column, "Prob. > Chi2," provides the corresponding p-values. The last two columns, AIC and BIC, provides the Akaike information criterion and Bayesian information criterion respectively.

Table F.7: Child Health Outcomes (Correcting for Attrition)

	Females						Males					
	Basic Statistics				Block Perm. FL		Basic Statistics				Block Perm. FL	
Outcome Description	Cntr. Mean	Cd. Diff. Mn.	Cd. Eff. Size	Asy P-val	Single P-val	Stepdown	Cntr. Mean	Cd. Diff. Mn.	Cd. Eff. Size	Asy P-val	Single P-val	Stepdown
Birth Outcomes for Child												
Placenta Weight	683.488	-11.638	-0.073	0.707	0.467	0.717	662.401	27.965	0.157	0.112	**0.014**	**0.036**
Birth Weight	3050.565	-128.456	-0.235	0.966	0.903	0.903	2993.726	204.977	0.292	**0.006**	**0.000**	**0.001**
Head Circumference	33.257	0.038	0.023	0.425	0.203	0.459	33.506	0.327	0.146	0.107	**0.060**	**0.060**
Length	49.652	0.234	0.087	0.236	0.202	0.513	49.908	0.711	0.196	**0.042**	**0.018**	**0.033**
Gestational Age at Delivery	39.092	-0.545	-0.242	0.940	0.854	0.919	38.526	0.745	0.214	**0.028**	**0.001**	**0.005**
Child Health Outcomes (Year 12)												
Any Injuries Since Last Interview	0.175	-0.043	-0.122	0.171	0.216	0.386	0.232	-0.059	-0.149	0.132	0.120	0.474
# Hospitalizations for Injuries Since Last Interview	0.009	-0.011	-0.116	0.138	0.185	0.451	0.011	-0.013	-0.134	0.132	0.170	0.582
Total # Injuries Since Last Interview	0.200	-0.068	-0.156	0.102	**0.074**	0.224	0.278	-0.057	-0.110	0.212	0.268	0.685
Hospitalized Since Last Interview	0.059	-0.044	-0.226	**0.033**	**0.035**	0.140	0.040	0.054	0.299	0.975	0.890	0.890
Have Chronic Condition/Health Problem	0.203	-0.003	-0.009	0.473	0.639	0.639	0.360	0.077	0.163	0.885	0.849	0.965
Standardized Child BMI	1.090	-0.240	-0.277	**0.019**	**0.012**	**0.060**	0.778	0.224	0.257	0.968	0.833	0.988

Note: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Control Mean) of each result set shows the unconditional mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Difference in Means) gives the conditional difference in means between the treatment group and the control group. As mentioned in Section 2, the control group stands for the original treatment group 2 of the NFP experiment and the treatment group stands for the original group 4. The third column (Effect Size) presents the conditional effect size for the respective group. The fourth column (Asymp. p-value) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the conditional difference in means between treatment and control groups. The fifth column (Block Permutation – Single p-value) presents the one-sided restricted permutation p-values for the single-hypothesis testing based on the t-statistic associated with the treatment indicator in the Freedman and Lane (1983) regression as described in Section 3. By restricted permutation we mean that permutations are done within strata defined by the baseline variables used in the randomization protocol: maternal age and race, gestational age at enrollment, employment status of the head of the household, and geographic region. The covariates used in the Freedman and Lane (1983) regression are: maternal height, household income, grandmother support, maternal parenting attitudes and mother currently in school. Finally, the last column (Block Permutation – Stepdown) provides p-values that account for multiple-hypothesis testing based on the Stepdown procedure of Romano and Wolf (2005). Blocks of outcomes that are tested jointly are separated by lines. The selection of blocks of outcomes is done on the basis of their meaning. Outcomes that share similar meaning are grouped together. Female maternal outcomes allude to mothers whose first child is a girl. Likewise, male maternal outcomes alludes to mothers whose first child is a boy. The results in this table use an inverse probability weighting scheme to address attrition. The weights are based on the predicted probability to drop the sample. The prediction is based on a Logit model that is described at the beginning of this section.

Table F.8: Family Environment (Correcting for Attrition)

Outcome Description	Females						Males					
	Basic Statistics				Block Perm. FL		Basic Statistics				Block Perm. FL	
	Cntr. Mean	Cd. Diff. Mn.	Cd. Eff. Size	Asy P-val	Single P-val	Stepdown	Cntr. Mean	Cd. Diff. Mn.	Cd. Eff. Size	Asy P-val	Single P-val	Stepdown
Home Environment, Parenting (Year 1) - Factor Scores												
Home Observation Measurement of the Environment (HOME)	0.000	0.338	0.338	**0.004**	**0.004**	**0.004**	-0.005	0.154	0.154	0.114	**0.079**	**0.079**
Non-Abussive Parenting Attitudes (Bavolek)	0.007	0.294	0.293	**0.010**	**0.003**	**0.006**	-0.002	0.364	0.364	**0.002**	**0.001**	**0.002**
Home Environment, Parenting (Year 2)- Factor Scores												
Home Observation Measurement of the Environment (HOME)	0.001	0.298	0.297	**0.010**	**0.004**	**0.007**	-0.008	0.116	0.116	0.186	0.111	0.111
Non-Abussive Parenting Attitudes (Bavolek)	0.012	0.374	0.372	**0.003**	**0.005**	**0.005**	-0.005	0.481	0.481	**0.000**	**0.001**	**0.001**
Maternal Mental Health (Year 2)												
Anxiety	-0.001	-0.226	-0.226	**0.042**	**0.038**	**0.086**	0.012	-0.052	-0.052	0.340	0.348	0.633
Depression	0.000	-0.115	-0.115	0.180	0.102	0.169	0.010	-0.011	-0.011	0.465	0.524	0.692
Positive Well-Being	-0.002	0.096	0.096	0.222	0.413	0.413	-0.006	-0.213	-0.214	0.950	0.947	0.947
Emotional Stability	0.001	0.185	0.185	**0.076**	**0.056**	0.113	-0.012	0.042	0.042	0.367	0.427	0.689
Overall Mental Health	0.000	0.193	0.193	**0.066**	**0.066**	0.122	-0.011	-0.047	-0.047	0.644	0.666	0.772
Self-Esteem	0.011	0.283	0.283	**0.014**	**0.003**	**0.014**	-0.011	0.045	0.045	0.367	0.467	0.707
Mastery	0.009	0.251	0.250	**0.030**	**0.018**	**0.057**	-0.010	0.253	0.252	**0.026**	**0.040**	0.137
Total Cost of Govt. Programs (Child Ages 1 - 12 Years)												
AFDC/TANF	2585.286	-177.226	-0.070	0.280	0.627	0.627	2657.084	-426.434	-0.165	**0.073**	**0.087**	0.156
Food Stamp	2900.613	-374.602	-0.229	**0.026**	0.241	0.362	3191.672	-288.782	-0.187	**0.061**	0.118	0.155
Medicaid	3462.064	-367.166	-0.221	**0.035**	0.275	0.377	3747.045	-271.420	-0.183	**0.068**	0.153	0.153

Note: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Control Mean) of each result set shows the unconditional mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Difference in Means) gives the conditional difference in means between the treatment group and the control group. As mentioned in Section 2, the control group stands for the original treatment group 2 of the NFP experiment and the treatment group stands for the original group 4. The third column (Effect Size) presents the conditional effect size for the respective group. The fourth column (Asymp. p-value) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the conditional difference in means between treatment and control groups. The fifth column (Block Permutation – Single p-value) presents the one-sided restricted permutation p-values for the single-hypothesis testing based on the t-statistic associated with the treatment indicator in the Freedman and Lane (1983) regression as described in Section 3. By restricted permutation we mean that permutations are done within strata defined by the baseline variables used in the randomization protocol: maternal age and race, gestational age at enrollment, employment status of the head of the household, and geographic region. The covariates used in the Freedman and Lane (1983) regression are: maternal height, household income, grandmother support, maternal parenting attitudes and mother currently in school. Finally, the last column (Block Permutation – Stepdown) provides p-values that account for multiple-hypothesis testing based on the Stepdown procedure of Romano and Wolf (2005). Blocks of outcomes that are tested jointly are separated by lines. The selection of blocks of outcomes is done on the basis of their meaning. Outcomes that share similar meaning are grouped together. Female maternal outcomes allude to mothers whose first child is a girl. Likewise, male maternal outcomes alludes to mothers whose first child is a boy. The results in this table use an inverse probability weighting scheme to address attrition. The weights are based on the predicted probability to drop the sample. The prediction is based on a Logit model that is described at the beginning of this section.

Table F.9: Cognitive Abilities and Achievement Outcomes (Correcting for Attrition)

	Females						Males					
	Basic Statistics				Block Perm. FL		Basic Statistics				Block Perm. FL	
Outcome Description	Cntr. Mean	Cd. Diff. Mn.	Cd. Eff. Size	Asy P-val	Single P-val	Stepdown	Cntr. Mean	Cd. Diff. Mn.	Cd. Eff. Size	Asy P-val	Single P-val	Stepdown
Kaufman Assessment Battery for Children (Year 6)												
Gestalt Closure	9.026	0.244	0.081	0.266	0.193	0.487	9.787	-0.388	-0.134	0.837	0.636	0.636
Hand Movements	9.267	0.438	0.203	**0.065**	**0.025**	0.150	9.319	0.127	0.060	0.332	0.398	0.711
Matrix Analogies	8.636	0.136	0.080	0.273	0.300	0.579	8.480	0.285	0.180	**0.092**	0.124	0.434
Number Recall	9.390	0.437	0.152	0.120	**0.086**	0.327	8.886	1.004	0.421	**0.002**	**0.004**	**0.029**
Photo Series	7.040	0.424	0.212	**0.055**	**0.064**	0.284	6.791	0.030	0.014	0.458	0.496	0.700
Spatial Memory	8.441	0.204	0.084	0.264	0.341	0.516	8.568	0.218	0.090	0.258	0.194	0.531
Triangles	8.845	0.435	0.188	**0.070**	0.129	0.402	9.201	0.094	0.041	0.382	0.213	0.522
Word Order	9.737	-0.079	-0.030	0.591	0.386	0.386	9.148	0.763	0.298	**0.016**	**0.006**	**0.039**
Kaufman Assessment Battery for Children (Year 6)												
Nonverbal	89.267	2.118	0.239	**0.041**	**0.051**	0.104	89.466	1.104	0.118	0.196	0.187	0.235
Sequential Processing	96.587	1.685	0.131	0.160	**0.071**	0.124	94.353	3.676	0.312	**0.013**	**0.011**	**0.023**
Simultaneous Processing	88.981	1.980	0.196	**0.072**	**0.094**	**0.094**	90.128	0.498	0.050	0.359	0.231	0.231
WISC-III, PPVT-III for Children (Year 6)												
Wechsler Intelligence Scale for Children (WISC-III)	96.518	0.900	0.050	0.348	0.352	0.352	90.692	1.746	0.102	0.227	0.298	0.298
Peabody Picture Vocabulary Test (PPVT-III)	83.682	1.685	0.154	0.119	0.164	0.286	82.695	2.325	0.221	**0.062**	**0.013**	**0.024**
Child Cognition (Year 6) - Factor Scores												
Cognition + Achievment (KABC, PPVT, WISC)	0.005	0.118	0.118	0.188	**0.067**	**0.092**	-0.010	0.182	0.182	**0.092**	**0.063**	**0.063**
Cognitive skills (Mental Processing Composite-KABC)	0.000	0.137	0.137	0.150	**0.073**	**0.073**	-0.007	0.277	0.277	**0.023**	**0.015**	**0.021**
Reading Achievement for the Child (Year 12)												
Average Reading Grade (Grades 1 - 5)	2.694	0.076	0.101	0.228	0.107	0.271	2.348	0.058	0.078	0.296	**0.100**	0.164
TCAP % Language (School Years 1 - 5, Grd 3+)	51.600	0.372	0.016	0.456	0.180	0.307	37.918	5.076	0.233	**0.067**	**0.005**	**0.021**
TCAP % Reading (School Years 1 - 5, Grd 3+)	42.374	0.197	0.010	0.473	0.164	0.310	35.020	1.750	0.088	0.280	**0.043**	0.117
PIAT Total Reading (Derived Score)	90.420	0.662	0.069	0.307	0.344	0.404	89.350	1.381	0.103	0.221	**0.063**	0.129
PIAT Reading Comprehension (Derived Score)	88.458	-0.232	-0.026	0.576	0.546	0.546	87.641	2.369	0.203	**0.072**	**0.022**	**0.070**
PIAT Reading Recognition (Derived Score)	94.486	2.175	0.180	0.102	0.156	0.325	92.620	0.266	0.018	0.447	0.136	0.136
Math Achievement for the Child (Year 12)												
Average Math Grade (Grades 1 - 5)	2.622	0.093	0.113	0.196	0.146	0.270	2.391	0.099	0.130	0.183	**0.072**	**0.072**
TCAP % Math (School Years 1 - 5, Grd 3+)	47.610	1.896	0.080	0.291	0.188	0.279	40.176	3.086	0.139	0.185	**0.033**	**0.082**
PIAT Mathematics (Derived Score)	87.413	-0.080	-0.008	0.525	0.727	0.727	86.538	1.947	0.193	**0.086**	**0.048**	**0.085**

Note: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Control Mean) of each result set shows the unconditional mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Difference in Means) gives the conditional difference in means between the treatment group and the control group. As mentioned in Section 2, the control group stands for the original treatment group 2 of the NFP experiment and the treatment group stands for the original group 4. The third column (Effect Size) presents the conditional effect size for the respective group. The fourth column (Asymp. p-value) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the conditional difference in means between treatment and control groups. The fifth column (Block Permutation – Single p-value) presents the one-sided restricted permutation p-values for the single-hypothesis testing based on the t-statistic associated with the treatment indicator in the Freedman and Lane (1983) regression as described in Section 3. By restricted permutation we mean that permutations are done within strata defined by the baseline variables used in the randomization protocol: maternal age and race, gestational age at enrollment, employment status of the head of the household, and geographic region. The covariates used in the Freedman and Lane (1983) regression are: maternal height, household income, grandmother support, maternal parenting attitudes and mother currently in school. Finally, the last column (Block Permutation – Stepdown) provides p-values that account for multiple-hypothesis testing based on the Stepdown procedure of Romano and Wolf (2005). Blocks of outcomes that are tested jointly are separated by lines. The selection of blocks of outcomes is done on the basis of their meaning. Outcomes that share similar meaning are grouped together. Female maternal outcomes allude to mothers whose first child is a girl. Likewise, male maternal outcomes alludes to mothers whose first child is a boy. The results in this table use an inverse probability weighting scheme to address attrition. The weights are based on the predicted probability to drop the sample. The prediction is based on a Logit model that is described at the beginning of this section.

Table F.10: Socio-Emotional Abilities (Correcting for Attrition)

	Females						Males					
	Basic Statistics			Block Perm. FL			Basic Statistics			Block Perm. FL		
Outcome Description	Cntr. Mean	Cd. Diff. Mn.	Cd. Eff. Size	Asy P-val	Single P-val	Stepdown	Cntr. Mean	Cd. Diff. Mn.	Cd. Eff. Size	Asy P-val	Single P-val	Stepdown
Child Behavior Checklist (Year 2) - Factor Scores												
Affective Problems	-0.001	-0.337	-0.337	**0.002**	**0.003**	**0.015**	0.004	0.287	0.287	0.985	0.955	0.955
Anxiety Problems	-0.002	-0.181	-0.181	**0.066**	0.249	0.249	0.007	0.016	0.016	0.550	0.636	0.907
Pervasion Developmental Problems	-0.005	-0.261	-0.261	**0.013**	**0.060**	**0.100**	0.005	0.185	0.185	0.925	0.817	0.950
Attention Deficit Hyperactivity Disorder	-0.001	-0.243	-0.242	**0.025**	**0.019**	**0.060**	0.003	0.056	0.056	0.670	0.706	0.923
Oppositional Defiant Problems	-0.001	-0.217	-0.217	**0.040**	**0.053**	0.120	0.005	0.126	0.126	0.853	0.880	0.962
Child Behavior Checklist (Year 6) - Factor Scores												
Affective Problems	-0.010	-0.007	-0.007	0.479	0.612	0.796	-0.004	-0.103	-0.103	0.203	0.151	0.481
Anxiety Problems	-0.008	-0.061	-0.061	0.306	0.492	0.759	0.009	0.083	0.082	0.729	0.813	0.813
Somatic Problems	-0.003	0.130	0.130	0.832	0.884	0.884	0.007	0.063	0.063	0.678	0.442	0.757
Attention Deficit Hyperactivity Problems	-0.012	-0.230	-0.230	**0.035**	**0.096**	0.307	-0.006	-0.040	-0.040	0.379	0.310	0.713
Oppositional Defiant Problems	0.000	-0.027	-0.027	0.415	0.286	0.608	-0.013	-0.083	-0.083	0.270	0.317	0.672
Conduct Problems	-0.002	-0.267	-0.266	**0.013**	**0.003**	**0.015**	-0.009	-0.011	-0.011	0.467	0.485	0.665
MacArthur Story Stem Battery (MSSB) (Year 6) - Factor Scores												
Dysregulated Aggression	-0.006	-0.027	-0.027	0.413	0.135	0.269	-0.009	-0.130	-0.130	0.189	0.137	0.496
Warmth and Empathy	-0.011	0.388	0.388	**0.002**	**0.005**	**0.019**	-0.011	-0.099	-0.099	0.770	0.535	0.832
Emotional Integration	-0.005	-0.028	-0.028	0.585	0.765	0.765	-0.015	0.055	0.055	0.349	0.429	0.849
Performance Anxiety	0.010	-0.038	-0.038	0.373	**0.093**	0.259	-0.009	0.077	0.077	0.701	0.843	0.843
Aggression	-0.005	-0.164	-0.164	**0.084**	**0.003**	**0.012**	-0.010	-0.095	-0.095	0.260	0.177	0.570
Internalizing, Externalizing, Absences (Year 12)												
Internalizing Disorders	0.240	-0.028	-0.066	0.309	0.453	0.813	0.397	-0.087	-0.183	**0.090**	**0.082**	0.154
Externalizing Disorders	0.183	-0.013	-0.032	0.402	0.641	0.866	0.183	0.089	0.239	0.951	0.859	0.859
Average # of Absences (School Years 1 - 5)	10.144	0.263	0.035	0.605	0.666	0.666	11.548	-1.838	-0.246	**0.029**	**0.027**	**0.077**

Note: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Control Mean) of each result set shows the unconditional mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Difference in Means) gives the conditional difference in means between the treatment group and the control group. As mentioned in Section 2, the control group stands for the original treatment group 2 of the NFP experiment and the treatment group stands for the original group 4. The third column (Effect Size) presents the conditional effect size for the respective group. The fourth column (Asymp. p-value) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the conditional difference in means between treatment and control groups. The fifth column (Block Permutation – Single p-value) presents the one-sided restricted permutation p-values for the single-hypothesis testing based on the t-statistic associated with the treatment indicator in the Freedman and Lane (1983) regression as described in Section 3. By restricted permutation we mean that permutations are done within strata defined by the baseline variables used in the randomization protocol: maternal age and race, gestational age at enrollment, employment status of the head of the household, and geographic region. The covariates used in the Freedman and Lane (1983) regression are: maternal height, household income, grandmother support, maternal parenting attitudes and mother currently in school. Finally, the last column (Block Permutation – Stepdown) provides p-values that account for multiple-hypothesis testing based on the Stepdown procedure of Romano and Wolf (2005). Blocks of outcomes that are tested jointly are separated by lines. The selection of blocks of outcomes is done on the basis of their meaning. Outcomes that share similar meaning are grouped together. Female maternal outcomes allude to mothers whose first child is a girl. Likewise, male maternal outcomes alludes to mothers whose first child is a boy. The results in this table use an inverse probability weighting scheme to address attrition. The weights are based on the predicted probability to drop the sample. The prediction is based on a Logit model that is described at the beginning of this section.

G A Framework for Mediation Analysis

This section develops a theoretical framework to conduct our mediation analysis. Our model is motivated by the literature on the technology of skill formation (Cunha and Heckman, 2007). In it, subsequent skills build on earlier skills to generate human capital. Notationally, let $\boldsymbol{\theta}_{i,t}$ be the vector of skills during childhood for individual i at period t and $t \in \{0, 1, \ldots, T\}$, where T is the number of periods of childhood. Let $\boldsymbol{I}_{i,t}$ represent investments at the same period. We use $\boldsymbol{X}_i$ for family background characteristics and $\upsilon_{i,t}$ for an exogenous error term independent of $\boldsymbol{\theta}_{i,t}, \boldsymbol{I}_{i,t}$ and $\boldsymbol{X}_i$. The structural equations that govern the evolution of skills are given by:

$$\boldsymbol{\theta}_{i,t+1} = \boldsymbol{q}_{t+1}(\boldsymbol{\theta}_{i,t}, \boldsymbol{I}_{i,t+1}, \boldsymbol{X}_i, \upsilon_{i,t+1}); t \in \{0, 1, \ldots, T-1\}. \tag{17}$$

By the term "structural equations," we mean autonomous functions in the language of Frisch (1938), i.e. deterministic functions whose functional forms do not change as their arguments vary. We also allow for skills to affect investments, that is:

$$\boldsymbol{I_{i,t+1}} = \boldsymbol{h}_{t+1}(\boldsymbol{\theta}_{i,t}, \boldsymbol{X}_i, \varepsilon_{i,t+1}); t \in \{0, 1, \ldots, T-1\}, \tag{18}$$

where $\varepsilon_{i,t+1}$ is an exogenous error term independent of $\boldsymbol{\theta}_{i,t}$ and $\boldsymbol{X}_i$. Our model is completed by the following structural outcome equation at period T :

$$\boldsymbol{Y_i} = \boldsymbol{g}_T(\boldsymbol{\theta}_{i,T}, \boldsymbol{X}_i, \xi_{i,T}). \tag{19}$$

where $\xi_{i,T}$ is an exogenous error term independent of $\boldsymbol{\theta}_{i,T}$ and $\boldsymbol{X}_i$.

We can use a recursive substitution of investments and skills of Equations (17)–(18) into (19) to generate the following equation:

$$\boldsymbol{Y_i} = \boldsymbol{f}_{t'}(\boldsymbol{\theta}_{i,t'}, \boldsymbol{X}_i, \{\upsilon_{i,\tilde{t}}\}_{\tilde{t}=t'}^{T}, \{\varepsilon_{i,\tilde{t}}\}_{\tilde{t}=t'}^{T}, \xi_{i,T}), \tag{20}$$

where $\{\upsilon_{i,\tilde{t}}\}_{\tilde{t}=t'}^{T} = \{\upsilon_{i,t'}, \upsilon_{i,t'+1}, \ldots, \upsilon_{i,T}\}$ and $\{\varepsilon_{i,\tilde{t}}\}_{\tilde{t}=t'}^{T} = \{\varepsilon_{i,t'}, \varepsilon_{i,t'+1}, \ldots, \varepsilon_{i,T}\}$.

Suppose that an intervention occurs at period t' where $t' \in \{1, \ldots, T\}$. Let $D_i \in \{0, 1\}$ be the treatment indicator of this intervention which takes value 1 if participant i is treated and 0 otherwise. The intervention enters our technology of skill formation model as a form of skill investment. Thus we append the investment Equation (18) at period t' by:

$$\boldsymbol{I_{i,t'}} = \boldsymbol{h}_{t'}(\boldsymbol{\theta}_{i,t'-1}, D_i, \boldsymbol{X}_i, \varepsilon_{i,t'}); \text{ for some } t' \in \{0, 1, \ldots, T-1\}, \tag{21}$$

The counterfactual values investment $\boldsymbol{I}_{i,t'}$ are defined by the value $\boldsymbol{I}_{i,t'}$ takes when the intervention D_i is fixed at a level $d \in \{0, 1\}$. By fixing, I mean the causal operation defined in Haavelmo (1944) where D_i is set to $d \in \{0, 1\}$ as argument in the structural equation (21). That is:

$$\boldsymbol{I_{i,t',d}} = \boldsymbol{h}_{t'}(\boldsymbol{\theta}_{i,t'-1}, d, \boldsymbol{X}_i, \varepsilon_{i,t'}); d \in \{0, 1\} \text{ for some } t' \in \{0, 1, \ldots, T-1\}. \tag{22}$$

Let the counterfactual skills be defined in a symmetric fashion by:

$$\boldsymbol{\theta}_{i,t',d} = \boldsymbol{q}_{t'}(\boldsymbol{\theta}_{i,t'-1}, \boldsymbol{I}_{i,t',d}, \boldsymbol{X}_i, \upsilon_{i,t'}).$$

We also define the counterfactual skills and investments for periods $t > t'$ by:

$$\begin{aligned} \boldsymbol{I}_{i,t+1,d} &= \boldsymbol{h}_{t+1}(\boldsymbol{\theta}_{i,t,d}, \boldsymbol{X}_i, \varepsilon_{i,t+1}), \text{ and} \\ \boldsymbol{\theta}_{i,t+1,d} &= \boldsymbol{q}_{t+1}(\boldsymbol{\theta}_{i,t,d}, \boldsymbol{I}_{i,t+1,d}, \boldsymbol{X}_i, \upsilon_{i,t+1}); t > t'. \end{aligned}$$

We can also define the counterfactual outcomes by:

$$\boldsymbol{Y_{i,d}} = \boldsymbol{f}_{t'}(\boldsymbol{\theta}_{i,t',d}, \boldsymbol{X}_i, \{\epsilon_{i,\tilde{t}}\}_{\tilde{t}=t'}^{T}, \{\varepsilon_{i,\tilde{t}}\}_{\tilde{t}=t'}^{T}, \xi_{i,T}), \tag{23}$$

If the intervention assignment uses the method of randomization, then we have that:

$$(\boldsymbol{Y_{i,d}}, \boldsymbol{\theta}_{i,t',d}) \perp\!\!\!\perp D_i|\boldsymbol{X}_i;\ d \in \{0,1\}.$$

We can also write the realized values of skills and outcomes as:

$$\boldsymbol{Y_i} = \boldsymbol{Y}_{i,1}D_i + \boldsymbol{Y}_{i,0}(1-D_i),\ \text{and}$$
$$\boldsymbol{\theta}_{i,t} = \boldsymbol{\theta}_{i,t,1}D_i + \boldsymbol{\theta}_{i,t,0}(1-D_i); t \geq t'.$$

We use Equation (23) to generate a tractable equation to examine mediation effects. Note that Equation (23) holds not only for t' but for any $t \geq t'$.

$$\boldsymbol{Y_{i,d}} = \boldsymbol{f}_t(\boldsymbol{\theta}_{i,t,d}, \boldsymbol{X}_i, \{\upsilon_{i,\tilde{t}}\}_{\tilde{t}=t}^{T}, \{\varepsilon_{i,\tilde{t}}\}_{\tilde{t}=t}^{T}, \xi_{i,T}),\ \text{for any}\ t \in \{t', t'+1, \ldots, T\}. \tag{24}$$

Error terms $(\{\upsilon_{i,\tilde{t}}\}_{\tilde{t}=t}^{T}, \{\varepsilon_{i,\tilde{t}}\}_{\tilde{t}=t}^{T}, \xi_{i,T})$ are independent of $\boldsymbol{\theta}_{i,t,d}$ and $\boldsymbol{X}_i$. For sake of notational simplicity, we can substitute those error terms by ζ_t without loss of generality. Equation (24) then becomes:

$$\boldsymbol{Y}_{i,d} = f_t(\boldsymbol{\theta}_{i,t,d}, \boldsymbol{X}_i, \zeta_{i,t}). \tag{25}$$

We achieve a linear form of Equation (25) by approximating it through a Maclaurin expansion. This generates the following equation:

$$\boldsymbol{Y_{i,d,t}} = \boldsymbol{\kappa_t} + \boldsymbol{\alpha}_{t,d}\boldsymbol{\theta}_{i,t,d} + \boldsymbol{\beta}_{t,d}\boldsymbol{X}_i + \boldsymbol{\epsilon_{i,t,d}},\ \ d \in \{0,1\}. \tag{26}$$

where $\boldsymbol{\epsilon_{t,d}}$ accounts for the approximation error. Equations (25)–(26) are used in our mediation analysis in Section 5.

H Mediation Methodology

H.1 Three Step Procedure

This part of the appendix explains in detail the three step procedure that we use in order to decompose the NFP treatment effects. As noted in the paper, we perform two sets of analyses. First, we study whether the treatment effects on child skills at age 6 were mediated by program enhancement of birth weight, parenting attitudes and investments, and maternal socio-emotional skills at age 2. Second, we study whether the program impact on outcomes at age 12 was mediated by the NFP enhancement of skills at age 6. The results from these analysis shed light on the complementarity of investments and skills in explaining the NFP treatment effects.

Step One The idea is to develop a measurement system that links the observed items and the latent skills. In order to do that, we assume that our measurements are dedicated. This means that each observed measurement is linked to a unique skill. Specifically, let $\mathcal{M}^j$ be the index set of measures associated with trait j, where $j \in \mathcal{J} = \{P, C, SE\}$. P, C, SE denote, respectively, parenting skills, child cognitive skills, and child socio-emotional abilities.[3] Thus, our linear measurement system is as follows:[4]

$$M^j_{m^j,d} = \nu^j_{m^j} + \boldsymbol{\varphi^j_{m^j}} \boldsymbol{\theta^j_d} + \eta^j_{m^j,d}, \tag{27}$$

where $\nu^j_{m^j}$ is the intercept term and $\varphi^j_{m^j}$ represents the loading factor of trait j. We cannot reject the null hypothesis that the intercepts and loading factors depend on treatment status. $\eta_{m^j,d}$ is a mean zero idiosyncratic error term which, by assumption, is independent of $\boldsymbol{\theta}^j_d$ $\forall j \in \mathcal{J}$. We normalize the loading factor associated with the first measure of each factor

[3] This follows the same notation as Heckman et al. (2013)

[4] We control for pre-program variables X but we keep it implicit to shorten notation.

to 1 in order to set a scale, otherwise the scale is arbitrary.[5] Finally, we allow for factor correlation.

The parameters that identify the measurement system are the factor means, the factor covariances, the intercepts, the factor loadings, and the variances of the error terms: $E[\boldsymbol{\theta^j(d)}] = \mu^j_d$, $Var[\boldsymbol{\theta_d}] = \Sigma_{\boldsymbol{\theta_d}}$, $\boldsymbol{\nu^j_{m^j}}$, $\boldsymbol{\varphi^j_{m^j}}$, $Var[\eta^j_{m^j}]$). Heckman et al. (2013) show that the existence of at least three measures for each latent skill guarantees identification.[6] Broadly, means, variances, and covariances across the measures identify the parameters of the system.

We estimate the parameters of the measurement system that links skills with measures both at ages 2 and 6. Variables become potential mediators if we estimate an effect of the NFP on it, so that they are potential meaningful channels. For age 2, non-abusive parenting attitudes are approximated by the Adult-Adolescent Parenting Inventory (Bavolek), which comprises 32 items, and home investments are measured by the Bradley and Caldwell Home Observation for measurement of the Environment (HOME) inventory, which is composed of 45 items.

The maternal skills selected correspond to anxiety, assessed by the Rand Mental Health Inventory, self-esteem, measured by the Rosenberg scale, and mastery, approximated by the Pearlin scale. Similarly, for age 6, we select children's skills influenced by the NFP as plausible mediators . Child cognition is measured by 8 subtests from the K-ABC mental processing composite. For children's socio-emotional skills, we identify as potential mediators the treatment reduction in conduct, attention and aggression problems, as well as the enhancement of children's pro-social skills. Attention and conduct problems are approximated by items from the Child Behavior Checklist. Pro-social skills (warmth or empathy) and aggression problems are approximated by items from the MacArthur Story Stem Battery. Section C of the Appendix explains in more detail these tests, as well as the instruments they use.

We estimate the parameters of the measurement system by maximum likelihood. In

[5]Given that the first measure sets the scale, we choose it to be the most correlated with the skill. The results are robust to alterations of this.

[6] Carneiro et al. (2003) and Cunha et al. (2010) also discuss identification of factor models.

order to do this, we assume that the latent skills and the error terms, θ^j and $\eta^j_{m^j}$, are normal and i.i.d. We use full-information maximum likelihood to deal with the missing values in the measures for some individuals. FIML yields unbiased estimates that are more efficient than ad hoc methods like list-wise and pair-wise deletion, which work under the implicit assumption of random missing data. By missing at random we mean that the probability of data associated with a variable x can depend on other observed variables but not on the values of x itself.

For the case of the measurement system at age 2, we have 146 items. Although it is ideal to estimate the complete set of items (skills) jointly, it is not feasible. Thus, we estimate them in two blocks: one for parenting and home investments and other for maternal characteristics. This allows us to account for the correlation between the skills that are in the same block. For the case of the measurement system at age 6, the set of items is smaller and we do a joint estimation.

Step Two In the second step we use the parameter estimates from the first step to construct factor scores for each children. The objective of this is to construct approximations for the latent skills. The two most common linear scoring methods are the regression method and the Bartlett method, which resembles GLS (Thomson, 1934). We use the Bartlett (1937) method because it estimates unbiased approximations of the unobserved skills. Actually, this guarantees that the difference in means between the factor scores for children in the treatment and the control groups equals the difference in means in the true scores. The derivation of the Bartlett estimator begins with the measurement system summarized as:

$$\underbrace{\boldsymbol{M}_i}_{|\mathcal{M}|\times 1} = \underbrace{\boldsymbol{\varphi}}_{|\mathcal{M}|\times|\mathcal{J}|}\underbrace{\boldsymbol{\theta}_i}_{|\mathcal{J}|\times 1} + \underbrace{\boldsymbol{\eta}_i}_{|\mathcal{M}|\times 1}$$

where the dimension of each term is below the braces (recall that $\mathcal{J}$ and $\mathcal{M}$ are the indexing sets for skills and measures respectively). Assume that the $(\boldsymbol{\theta}_i, \boldsymbol{\eta}_i)$, $i \in \{1, \ldots, I\}$, are independent across i. For simplicity, we assume that they are i.i.d.[7] Let $Cov(\boldsymbol{M}_i, \boldsymbol{M}_i) = \boldsymbol{\Sigma}$, $Cov(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i) = \boldsymbol{\Phi}$ and $Cov(\boldsymbol{\eta}_i, \boldsymbol{\eta}_i) = \boldsymbol{\Omega}$. The linear relation between the factor scores and the measures is the following:

$$\boldsymbol{\theta}_{S,i} = \boldsymbol{L}'\boldsymbol{M}_i \tag{28}$$

In order to obtain unbiased estimates, Barlett imposes the restriction that $\boldsymbol{L}'\boldsymbol{\varphi} = \boldsymbol{I}_{|\mathcal{J}|}$. The Bartlett estimator for the vector of approximated skills ($\boldsymbol{\theta_i}$) is:

$$\boldsymbol{\theta}_{S,i} = (\hat{\boldsymbol{\varphi}}'\hat{\boldsymbol{\Omega}}^{\boldsymbol{-1}}\hat{\boldsymbol{\varphi}})^{-1}\hat{\boldsymbol{\varphi}}'\hat{\boldsymbol{\Omega}}^{\boldsymbol{-1}}\boldsymbol{M_i}, \tag{29}$$

where the matrix of loading factors, $\hat{\boldsymbol{\varphi}}$, and $\hat{\boldsymbol{\Omega}} = Cov(\boldsymbol{\eta}_i, \boldsymbol{\eta}_i)$ are both estimated in the first step. Bartlett's estimator is a Generalized Least Squares, GLS, procedure where measures are used as dependent variables and loading factors are treated as regressors. By the Gauss-Markov theorem, the Bartlett GLS estimator is optimal and hence leads to the best linear unbiased predictor (BLUE).

There are individuals that have missing data in some of the items that compose the measurement system. In order to take advantage of the information that they have (instead of list-wise delete them), we predict factor scores for them. We use the covariance between the measures and the factors from the sample with complete measurement system to predict scores for these people. Additionally, for the cases were individuals are missing a factor score because they did not have any item in that measurement system, we impute factor scores with the regression method.[8] This procedure recovers around 10% of the randomized sample.

Step 3 In this step, we use factor scores as approximations of the true skills to estimate the models that link the later outcomes with the intermediate skills. The factor scores are

[7] This is not strictly required but simplifies the notation.

[8] We impute factor scores for individuals that have at least two other factor scores.

measured with error, which produces downward-biased estimates of the parameters of the outcome equations. This bias corresponds to the traditional attenuation that results from classical measurement error. In factor scored regressions, Bolck et al. (2008) prove this. We adopt the bias correction strategy proposed by Croon (2002). In summary, this approach takes advantage of the fact that we have estimates of all the components of the bias. This strategy, also used by Heckman et al. (2013), can be summarized as follows:

Consider the model following model. To simplify notation, we use W to denote pre-program variables $\boldsymbol{X}$, treatment indicator and the intercept of equation 4:

$$\boldsymbol{Y_i} = \boldsymbol{\alpha}\boldsymbol{\theta}_i + \boldsymbol{\gamma}\boldsymbol{W}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, N. \tag{30}$$

The covariance matrix of $(\boldsymbol{\theta}_i, \boldsymbol{W}_i)$ is

$$\begin{pmatrix} Cov(\boldsymbol{\theta}, \boldsymbol{\theta}) & Cov(\boldsymbol{\theta}, \boldsymbol{W}) \\ Cov(\boldsymbol{W}, \boldsymbol{\theta}) & Cov(\boldsymbol{W}, \boldsymbol{W}) \end{pmatrix}.$$

We measure $\boldsymbol{\theta}_i$ with error. Thus,

$$\boldsymbol{\theta}_{S,i} = \boldsymbol{\theta}_i + \boldsymbol{V}_i, \quad i = 1, \ldots, N$$
$$(\boldsymbol{W}_i, \boldsymbol{\theta}_i) \perp\!\!\!\perp \boldsymbol{V}_i, \quad E(\boldsymbol{V}_i) = 0, \quad Cov(\boldsymbol{V}, \boldsymbol{V}) = \boldsymbol{\Sigma_{VV}}$$

Denote $Cov(\boldsymbol{\theta}_{S,i}, \boldsymbol{\theta}_{S,i}) = \boldsymbol{\Sigma}_{\boldsymbol{\theta_S}, \boldsymbol{\theta_S}}$. We assume that the $(\boldsymbol{\theta}_i, \boldsymbol{W}_i, \boldsymbol{\epsilon}_i)$ are i.i,d, but much weaker conditions suffice. Note that we do not assume that $\boldsymbol{\theta}_i \perp\!\!\!\perp \boldsymbol{W}_i$ as in traditional factor analysis. We do assume that $(\boldsymbol{\theta}_i, \boldsymbol{W}_i) \perp\!\!\!\perp \boldsymbol{\epsilon}_i$ and $E(\boldsymbol{\epsilon}_i) = 0$.

If we use $\boldsymbol{\theta}_{S,i}$ in place of Y_i, it follows that:

$$\boldsymbol{Y_i} = \boldsymbol{\alpha}\boldsymbol{\theta}_{S,i} + \boldsymbol{\gamma}\boldsymbol{W}_i + \boldsymbol{\epsilon}_i - \alpha\boldsymbol{V}_i. \tag{31}$$

The estimation of Equation 31 using OLS produces estimates that are biased:

$$plim \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} Cov(\boldsymbol{\theta}_S, \boldsymbol{\theta}_S) & Cov(\boldsymbol{\theta}_S, \boldsymbol{W}) \\ Cov(\boldsymbol{W}, \boldsymbol{\theta}_S) & Cov(\boldsymbol{W}, \boldsymbol{W}) \end{pmatrix}^{-1} \begin{pmatrix} Cov(\boldsymbol{\theta}, \boldsymbol{\theta}) & Cov(\boldsymbol{\theta}, \boldsymbol{W}) \\ Cov(\boldsymbol{W}, \boldsymbol{\theta}) & Cov(\boldsymbol{W}, \boldsymbol{W}) \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \gamma \end{pmatrix}.$$

Let $\boldsymbol{\Sigma}_{\mathbf{B},\mathbf{C}}$ be $Cov(\mathbf{B}, \mathbf{C})$. Observe that $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{W}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}_S,\boldsymbol{W}}$ as a consequence of our assumptions. In this notation

$$plim \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\gamma} \end{pmatrix} = \underbrace{\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{V},\boldsymbol{V}} & \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{W}} \\ \boldsymbol{\Sigma}_{\boldsymbol{W},\boldsymbol{\theta}} & \boldsymbol{\Sigma}_{\boldsymbol{W},\boldsymbol{W}} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\theta}} & \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{W}} \\ \boldsymbol{\Sigma}_{\boldsymbol{W},\boldsymbol{\theta}} & \boldsymbol{\Sigma}_{\boldsymbol{W},\boldsymbol{W}} \end{pmatrix}}_{\boldsymbol{A}} \begin{pmatrix} \boldsymbol{\alpha} \\ \gamma \end{pmatrix} \tag{32}$$

which is the usual attenuation formula.

From the estimation of the measurement system, we can identify $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\theta}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{W}}$, $\boldsymbol{\Sigma}_{\boldsymbol{V},\boldsymbol{V}}$, and we have all the components of $\boldsymbol{A}$. Hence if we pre-multiply the least squares estimator by $\boldsymbol{A}^{-1}$, we obtain:

$$plim \boldsymbol{A}^{-1} \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha} \\ \gamma \end{pmatrix}.$$

This is called "Croon's method" in psychometrics (Croon, 2002). In our application, there are two groups corresponding to $D = 0$ and $D = 1$ (control and treatment, respectively). We allow $\boldsymbol{\theta}_i$ to vary by treatment status. Indeed, our method assumes that treatment only operates through shifting the distribution of $\boldsymbol{\theta}$. We do not normalize the means of $\boldsymbol{\theta}$ (or $\boldsymbol{W}$) to be zero.

In the third step of our estimation procedure we compute bootstrapped p-values for each decomposition channel of the treatment effects. We take 100,000 resamples with replacement. The bootstrapped p-value for the null hypothesis $H_0 : \alpha_j = 0$ is calculated as follows:

$$p\text{-value} = \frac{1}{B}\sum_{b=1}^{B} 1(t_b^{j,*} > t^j) \text{ with } t^j = \frac{\hat{\alpha}^j}{\hat{\sigma}(\hat{\alpha}^j)} \text{ and } t_b^{j,*} = \frac{(\hat{\alpha}_b^j - \hat{\alpha}^j)}{\hat{\sigma}(\hat{\alpha}_b^j)} \tag{33}$$

where $\hat{\alpha_b^j}$ is bootstrapped estimated in the b^{th} resample and $\hat{\alpha}_j$ is estimated from the original data. Given the estimates of the outcome equation and of the factor scores, we construct the bootstrapped p-value for the contribution of skill k under the null hypothesis H_0 : $\hat{\alpha}^j E(\theta_1^j - \theta_0^j) = 0$ as follows:

$$p\text{-value} = \frac{1}{B}\sum_{b=1}^{B} 1(T_b^{j,*} > T^j) \text{ with } T^j = \frac{\hat{\alpha}^j * E(\theta^j \widehat{(1) - \theta^j}(0))}{\hat{\sigma}(\hat{\alpha}^j * E(\theta^j \widehat{(1) - \theta^j}(0)))} \tag{34}$$

where $T_b^{j,*}$ is the statistic T^j computed with the parameters obtained in the b^{th} resample. Notice that the p-value combines the variation in two population parameters: 1) the coefficient of the outcome equation; 2) the experimentally induced difference in means in the skills. It could be the case that each of these parameters are, separately, statistically significant. However, the p-value may increase due to a loss in power when they are combined.

Tables H.1 - H.4 shows the parameters of the outcome equations as wells as the decompositions components.

Table H.1: Female Decomposition (Year 6)

	Treatment		Birth Weight		Home y2		Parenting y2		Anxiety y2		Self-Esteem y2		Mastery y2		Sample Size
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	
Outcome Coefficients															
Cognitive	0.04	0.391	0.08	**0.089**	0.23	**0.035**	0.06	0.139	0.21	**0.087**	0.16	0.288	-0.08	0.363	304
Attention Problems	-0.15	**0.083**	-0.11	**0.013**	-0.11	0.169	-0.07	**0.042**	-0.14	0.139	0.24	0.206	-0.19	0.181	304
Conduct Problems	-0.15	**0.036**	-0.09	**0.018**	-0.07	0.249	-0.03	0.192	-0.18	**0.032**	-0.20	0.190	0.11	0.255	304
Warmth/Empathy	0.18	**0.060**	0.05	0.192	0.29	**0.003**	0.09	**0.014**	-0.01	0.481	-0.41	0.101	0.26	0.125	304
Aggression	-0.13	0.103	-0.04	0.218	-0.15	0.107	-0.01	0.416	0.13	0.177	-0.13	0.299	-0.06	0.386	304
Treatment Effect															
Cognitive	0.04	0.391	-0.01	0.110	0.04	**0.032**	0.02	**0.079**	0.03	**0.081**	0.03	0.246	-0.02	0.321	304
Attention Problems	-0.15	**0.083**	0.01	**0.099**	-0.02	0.144	-0.02	**0.046**	-0.02	0.115	0.04	0.168	-0.05	0.134	304
Conduct Problems	-0.15	**0.036**	0.01	0.112	-0.01	0.223	-0.01	0.170	-0.03	**0.065**	-0.04	0.153	0.03	0.208	304
Warmth/Empathy	0.18	**0.060**	-0.01	0.169	0.05	**0.007**	0.03	**0.018**	-0.00	0.434	-0.07	**0.073**	0.08	**0.093**	304
Aggression	-0.13	0.103	0.00	0.173	-0.03	**0.090**	-0.00	0.401	0.02	0.127	-0.02	0.263	-0.02	0.351	304
Treatment Effect Fraction															
Cognitive	0.29	0.391	-0.09	0.110	0.35	**0.032**	0.14	**0.079**	0.25	**0.081**	0.24	0.246	-0.19	0.321	304
Attention Problems	0.73	**0.083**	-0.07	**0.099**	0.10	0.144	0.09	**0.046**	0.10	0.115	-0.20	0.168	0.26	0.134	304
Conduct Problems	0.79	**0.036**	-0.06	0.112	0.06	0.223	0.05	0.170	0.14	**0.065**	0.19	0.153	-0.16	0.208	304
Warmth/Empathy	0.71	**0.060**	-0.03	0.169	0.21	**0.007**	0.11	**0.018**	-0.01	0.434	-0.29	**0.073**	0.29	**0.093**	304
Aggression	0.74	0.103	-0.03	0.173	0.16	**0.090**	0.02	0.401	-0.12	0.127	0.13	0.263	0.10	0.351	304

Notes: The first column provides the outcome description and the top row provides information on the mediators. For Year 6, the mediators are treatment, birth weight, home environment, parenting, anxiety, self-esteem and mastery. The last column provides the sample size for the corresponding outcome in the first column. The rows are divided into 3 groups: Outcome Coefficients, Treatment Effect and Treatment Effect Fraction. The last of these groups is also shown visually in Figure 3. Each mediator has two subcolumns of information: the coefficient and the p-value. Bold p-values are significant at the 10% level. We used the following controls: maternal race, maternal age, maternal height, gestational age, household density, region, employment status of household head, grandmother support, randomization wave, income category, mother currently in school, and maternal parenting attitudes.

Table H.2: Male Decomposition (Year 6)

	Treatment		Birth Weight		Home y2		Parenting y2		Anxiety y2		Self-Esteem y2		Mastery y2		Sample Size
	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	
Outcome Coefficients															
Cognitive	0.08	0.240	0.08	**0.093**	0.35	**0.004**	0.11	**0.014**	0.06	0.327	-0.04	0.447	0.02	0.464	305
Aggression	-0.08	0.186	-0.02	0.331	0.07	0.241	-0.05	**0.091**	-0.23	**0.014**	0.37	**0.058**	-0.11	0.310	305
Treatment Effect															
Cognitive	0.08	0.240	0.02	**0.064**	0.04	**0.054**	0.02	**0.047**	0.00	0.281	-0.00	0.362	0.00	0.422	305
Aggression	-0.08	0.186	-0.01	0.296	0.01	0.165	-0.01	**0.091**	-0.01	0.252	0.02	0.173	-0.02	0.230	305
Treatment Effect Fraction															
Cognitive	0.50	0.240	0.14	**0.064**	0.22	**0.054**	0.11	**0.047**	0.02	0.281	-0.02	0.362	0.02	0.422	305
Aggression	0.84	0.186	0.05	0.296	-0.07	0.165	0.08	**0.091**	0.12	0.252	-0.23	0.173	0.20	0.230	305

Notes: The first column provides the outcome description and the top row provides information on the mediators. For Year 6, the mediators are treatment, birth weight, home environment, parenting, anxiety, self-esteem and mastery. The last column provides the sample size for the corresponding outcome in the first column. The rows are divided into 3 groups: Outcome Coefficients, Treatment Effect and Treatment Effect Fraction. The last of these groups is also shown visually in Figure 4. Each mediator has two subcolumns of information: the coefficient and the p-value. Bold p-values are significant at the 10% level. We used the following controls: maternal race, maternal age, maternal height, gestational age, household density, region, employment status of household head, grandmother support, randomization wave, income category, mother currently in school, and maternal parenting attitudes.

Table H.3: Female Decomposition (Year 12)

	Treatment		Cognition		Attention problems		Conduct Problems		Warmth/Empathy		Aggression		Sample Size
	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	
Outcome Coefficients													
Total # Days Child Used Marijuana	-0.15	**0.060**	-0.12	**0.086**	0.03	0.373	0.00	0.515	0.12	0.158	0.04	0.192	271
Child Used Alcohol, Marijuana, or Tobacco in Last 30 Days	-0.02	0.198	-0.02	**0.048**	0.00	0.467	0.02	0.338	-0.00	0.477	0.00	0.457	268
Standardized Child BMI (Year 12)	-0.30	**0.016**	-0.11	0.100	-0.22	0.115	0.37	**0.030**	0.08	0.206	-0.11	0.197	272
Treatment Effect													
Child Ever Used Marijuana	-0.15	**0.060**	-0.01	0.218	-0.00	0.318	-0.00	0.484	0.04	0.111	-0.01	0.155	271
Child Used Alcohol, Marijuana, or Tobacco in Last 30 Days	-0.02	0.198	-0.00	0.217	-0.00	0.431	-0.00	0.273	-0.00	0.469	-0.00	0.424	268
Standardized Child BMI (Year 12)	-0.30	**0.016**	-0.01	0.209	0.03	0.109	-0.05	**0.060**	0.02	0.162	0.02	0.145	272
Treatment Effect Fraction													
Child Ever Used Marijuana	1.12	**0.060**	0.05	0.218	0.03	0.318	0.00	0.484	-0.26	0.111	0.06	0.155	271
Child Used Alcohol, Marijuana, or Tobacco in Last 30 Days	0.83	0.198	0.05	0.217	0.02	0.431	0.07	0.273	0.01	0.469	0.01	0.424	268
Standardized Child BMI (Year 12)	1.06	**0.016**	0.02	0.209	-0.11	0.109	0.19	**0.060**	-0.08	0.162	-0.08	0.145	272

Notes: The first column provides the outcome description and the top row provides information on the mediators. For Year 12, the mediators are treatment, cognition, attention problems, Conduct Problems, Warmth/Empathy and Aggression. The last column provides the sample size for the corresponding outcome in the first column. The rows are divided into 3 groups: Outcome Coefficients, Treatment Effect and Treatment Effect Fraction. The last of these groups is also shown visually in Figure 7. Each mediator has two subcolumns of information: the coefficient and the p-value. Bolded p-values are significant at the 10% level. Bold p-values are significant at the 10% level. We used the following controls: maternal race, maternal age, maternal height, gestational age, household density, region, employment status of household head, grandmother support, randomization wave, income category, mother currently in school, and maternal parenting attitudes.

Table H.4: Male Decomposition (Year 12)

	Treatment		Cognition		Attention problems		Conduct Problems		Warmth/Empathy		Aggression		Sample Size
	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	
Outcome Coefficients													
Average TCAP percentile, y1-5: language composite	2.88	0.188	11.93	**0.000**	2.21	0.379	-0.94	0.442	2.40	0.178	-3.88	0.170	222
PIAT reading comprehension derived score	1.69	0.134	7.44	**0.000**	-1.41	0.326	0.45	0.480	0.31	0.391	0.57	0.371	272
Average math grade grades 1-5	0.03	0.368	0.50	**0.000**	-0.21	0.145	0.22	0.137	0.05	0.288	-0.13	0.118	243
Average math grade. Years 1-5 after KG	-0.00	0.538	0.50	**0.000**	-0.15	0.205	0.16	0.198	0.03	0.324	-0.14	**0.096**	246
average tcap percentile y1-5: math	0.81	0.348	15.29	**0.000**	5.32	0.231	-2.86	0.336	-0.98	0.380	-3.47	0.204	223
PIAT math derived score	1.64	0.106	7.86	**0.000**	-1.36	0.324	0.20	0.515	0.71	0.232	1.53	0.146	270
SC ever tried smoking: 1=yes	-0.05	**0.079**	0.02	0.254	0.04	0.279	0.02	0.341	-0.02	0.192	0.01	0.404	274
SC use alc, mar, tob last 30 days	-0.05	**0.04**	-0.00	0.40	-0.01	0.41	0.04	0.25	-0.02	**0.07**	0.03	0.20	272
Internalizing disorders - Youth report	-0.05	0.213	-0.07	**0.047**	0.02	0.421	0.03	0.406	0.03	0.295	0.10	**0.072**	274
Anxious/depressed - clinical or borderline disorder, youth report	-0.05	**0.082**	-0.05	**0.016**	-0.05	0.167	0.06	0.101	0.01	0.332	0.07	**0.083**	273
Average number of absences, school years 1-5	-1.05	0.146	-2.25	**0.001**	4.36	**0.010**	-3.87	**0.009**	0.22	0.372	-1.10	0.156	267
Treatment Effect													
Average TCAP percentile, y1-5: language composite	2.88	0.188	2.09	**0.064**	-0.02	0.432	0.03	0.361	-0.39	0.135	0.45	0.161	222
PIAT reading comprehension derived score	1.69	0.134	1.44	**0.041**	0.11	0.255	-0.03	0.364	-0.04	0.313	-0.07	0.270	272
Average math grade grades 1-5	0.03	0.368	0.08	**0.080**	-0.00	0.366	0.00	0.449	-0.01	0.200	0.02	0.107	243
Average math grade. Years 1-5 after KG	-0.00	0.538	0.08	**0.061**	0.00	0.451	-0.00	0.335	-0.00	0.230	0.02	0.121	246
average tcap percentile y1-5: math	0.81	0.348	2.68	**0.070**	-0.09	0.367	0.09	0.313	0.16	0.303	0.41	0.203	223
PIAT math derived score	1.64	0.106	1.55	**0.042**	0.11	0.242	-0.01	0.420	-0.08	0.169	-0.18	0.102	270
SC ever tried smoking: 1=yes	-0.05	**0.079**	0.00	0.198	-0.00	0.234	-0.00	0.334	0.00	0.194	-0.00	0.324	274
SC use alc, mar, tob last 30 days	-0.05	**0.043**	-0.00	0.342	0.00	0.328	-0.00	0.262	0.00	0.144	-0.00	0.170	272
Internalizing disorders - Youth report	-0.05	0.213	-0.01	**0.058**	-0.00	0.342	-0.00	0.303	-0.00	0.245	-0.01	0.116	274
Anxious/depressed - clinical or borderline disorder, youth report	-0.05	**0.082**	-0.01	**0.028**	0.00	0.213	-0.00	0.219	-0.00	0.251	-0.01	**0.085**	273
Average number of absences, school years 1-5	-1.05	0.146	-0.36	**0.063**	-0.27	0.258	0.07	0.424	-0.03	0.272	0.14	0.127	267
Treatment Effect Fraction													
Average TCAP percentile, y1-5: language composite	0.57	0.188	0.41	**0.064**	-0.00	0.432	0.01	0.361	-0.08	0.135	0.09	0.161	222
PIAT reading comprehension derived score	0.54	0.134	0.46	**0.041**	0.03	0.255	-0.01	0.364	-0.01	0.313	-0.02	0.270	272
Average math grade. Years 1-5 after KG	0.23	0.368	0.68	**0.080**	-0.03	0.366	0.01	0.449	-0.05	0.200	0.16	0.107	243
average tcap percentile y1-5: math	0.20	0.348	0.66	**0.070**	-0.02	0.367	0.02	0.313	0.04	0.303	0.10	0.203	223
PIAT math derived score	0.54	0.106	0.51	**0.042**	0.04	0.242	-0.00	0.420	-0.03	0.169	-0.06	0.102	270
SC use alc, mar, tob last 30 days	0.92	**0.043**	0.02	0.342	-0.02	0.328	0.04	0.262	-0.04	0.144	0.08	0.170	272
Internalizing disorders - Youth report	0.63	0.213	0.17	**0.058**	0.01	0.342	0.02	0.303	0.03	0.245	0.14	0.116	274
Anxious/depressed - clinical or borderline disorder, youth report	0.71	**0.082**	0.14	**0.028**	-0.05	0.213	0.05	0.219	0.02	0.251	0.12	**0.085**	273
Average number of absences, school years 1-5	0.70	0.146	0.24	**0.063**	0.18	0.258	-0.05	0.424	0.02	0.272	-0.09	0.127	267

Notes: The first column provides the outcome description and the top row provides information on the mediators. For Year 12, the mediators are treatment, cognition, attention problems, Conduct Problems, Warmth/Empathy and Aggression. The last column provides the sample size for the corresponding outcome in the first column. The rows are divided into 3 groups: Outcome Coefficients, Treatment Effect and Treatment Effect Fraction. The last of these groups is also shown visually in Figures 5 - 6. Each mediator has two subcolumns of information: the coefficient and the p-value. Bold p-values are significant at the 10% level. We used the following controls: maternal race, maternal age, maternal height, gestational age, household density, region, employment status of household head, grandmother support, randomization wave, income category, mother currently in school, and maternal parenting attitudes.

I Mediation Specification Tests

In this section we specify how do we empirically test the effect that the mediators have on the final outcomes. We use $\mathcal{J}$ for an indexing set of skills. We use $\mathcal{J}_p \subseteq \mathcal{J}$ for the subset of measured skills. Our model for the outcome equation is:

$$Y_d = \kappa_d + \sum_{j \in \mathcal{J}} \alpha_d^j \theta_d^j + \boldsymbol{\beta}_d \boldsymbol{X} + \tilde{\epsilon}_d, \;\; d \in \{0,1\},$$

where κ_d is an intercept, $(\alpha_d^j; j \in \mathcal{J})$ are loading factors and $\boldsymbol{\beta}_d$ are $|\boldsymbol{X}|$-dimensional vectors of parameters. The error term $\tilde{\epsilon}_d$ is a zero-mean i.i.d. random variable assumed to be independent of regressors $(\theta_d^j; j \in \mathcal{J})$ and $\boldsymbol{X}$.

The NFP analysts collected a rich array of measures of cognitive and personality skills. However, it is likely that there are skills that they did not measure. As noted before, we use $\mathcal{J}_p \subseteq \mathcal{J}$ be the index set of measured skills. Namely, skills for which we have enough psychological instruments for estimation. We rewrite the equation for scalar potential outcome Y_d as:

$$\begin{aligned}
Y_d &= \kappa_d + \sum_{j \in \mathcal{J}} \alpha_d^j \theta_d^j + \boldsymbol{\beta}_d \boldsymbol{X} + \tilde{\epsilon}_d \\
&= \kappa_d + \underbrace{\sum_{j \in \mathcal{J}_p} \alpha_d^j \theta_d^j}_{\substack{\text{skills that} \\ \text{we measure}}} + \underbrace{\sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha^j \theta_d^j}_{\substack{\text{skills that we} \\ \text{do not measure}}} + \boldsymbol{\beta}_d \boldsymbol{X} + \tilde{\epsilon}_d \\
&= \underbrace{\kappa_d + \sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j \,\mathrm{E}(\theta_d^j)}_{\text{new intercept}} + \underbrace{\sum_{j \in \mathcal{J}_p} \alpha_d^j \theta_d^j}_{\substack{\text{skills that} \\ \text{we measure}}} + \underbrace{\sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j (\theta_d^j - \mathrm{E}(\theta_d^j))}_{\substack{\text{skills that we} \\ \text{do not measure}}} + \boldsymbol{\beta}_d \boldsymbol{X} + \tilde{\epsilon}_d, \\
&= \underbrace{\tau_d}_{\text{new intercept}} + \underbrace{\sum_{j \in \mathcal{J}_p} \alpha_d^j \theta_d^j}_{\substack{\text{skills that} \\ \text{we measure}}} + \boldsymbol{\beta}_d \boldsymbol{X} + \underbrace{\sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j (\theta_d^j - \mathrm{E}(\theta_d^j)) + \tilde{\epsilon}_d}_{\text{new error term}}
\end{aligned} \tag{35}$$

where $d \in \{0,1\}$, $\tau_d = \kappa_d + \sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j \,\mathrm{E}(\theta_d^j)$.

Any differences in the error terms between treatment and control groups can be attributed to differences in unmeasured skills. Thus, we assume, without loss of generality, that $\tilde{\epsilon}_1 \overset{d}{=} \tilde{\epsilon}_0$, where $\overset{d}{=}$ means equality in distribution.

The goal of this section is to examine the statistical assumptions needed to estimate unbiased parameters $(\alpha_d^j : j \in \mathcal{J}_p, d \in \{0,1\})$. These parameters are used to perform the decomposition of outcome treatment effects into parts associated with skills enhancement $(\theta_1^j - \theta_0^j : j \in \mathcal{J}_p)$. Parameters α may suffer from confounding effects if measured and unmeasured skills are not independent. We can solve this confounding problem by assuming that unmeasured skills are independent of measured skills. Namely,

$$(\theta_d^j;\, j \in \mathcal{J} \setminus \mathcal{J}_p) \perp\!\!\!\perp (\theta_d^j;\, j \in \mathcal{J}_p) | \boldsymbol{X};\; d \in \{0,1\},$$

then the regression:

$$Y_d = \tau_d + \sum_{j \in \mathcal{J}_p} \alpha_d^j \theta_d^j + \boldsymbol{\beta}_d \boldsymbol{X} + \epsilon_d, \tag{36}$$

produces unbiased estimates of parameter $(\alpha_d^j;\, j \in \mathcal{J}_p);\; d \in \{0,1\}$. Indeed error terms ϵ_d in equation (36) are given by

$$\epsilon_d = \tilde{\epsilon}_d + \sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j (\theta_d^j - \mathrm{E}(\theta_d^j))$$

which are independent of $(\theta_d^j;\, j \in \mathcal{J}_p)$ conditional on $\boldsymbol{X}$ under the assumption that skills are independent.

Now suppose that instead of the skills independence assumption for both groups, we focus only on the control group, thus,

$$(\theta_0^j;\, j \in \mathcal{J} \setminus \mathcal{J}_p) \perp\!\!\!\perp (\theta_0^j;\, j \in \mathcal{J}_p) | \boldsymbol{X}.$$

Moreover, suppose we also assume that $\alpha_1^j = \alpha_0^j;\; j \in \mathcal{J}$. Equivalently, the outcome loading factors for both treatment and control groups are the same. In this new setup, the regression

$$Y_0 = \tau_0 + \sum_{j \in \mathcal{J}_p} \alpha^j \theta_0^j + \boldsymbol{\beta}_0 \boldsymbol{X} + \epsilon_0, \tag{37}$$

also produces unbiased estimates of $(\alpha^j;\ j \in \mathcal{J}_p)$. Now consider the regression

$$Y_1 = \tau_1 + \sum_{j \in \mathcal{J}_p} \alpha^j \theta_1^j + \boldsymbol{\beta}_1 \boldsymbol{X} + \epsilon_1.$$

According to our rationale, this regression only produces unbiased estimates of $(\alpha^j;\ j \in \mathcal{J}_p)$ if:

$$(\theta_1^j;\ j \in \mathcal{J} \setminus \mathcal{J}_p) \perp\!\!\!\perp (\theta_1^j;\ j \in \mathcal{J}_p) | \boldsymbol{X}, \tag{38}$$

or, alternatively,

$$(\theta_1^j - \theta_0^j;\ j \in \mathcal{J} \setminus \mathcal{J}_p) \perp\!\!\!\perp (\theta_1^j - \theta_0^j;\ j \in \mathcal{J}_p) | \boldsymbol{X}. \tag{39}$$

Thus, under this new set of assumptions, testing $H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_0$ is translated into testing the independence relations of equations (38)–(39).

While the skill independence assumption in equation (38) may appear strong, the rich settlement of information on NFP surveys makes this assumption more plausible. NFP data has a huge selection of psychological questionnaires that aims to measure both cognitive and non-cognitive skills though childhood. We examine all the available data and only a subset of these measures turns out to be statistically relevant for mediation analysis. We use these measures to estimate factors that are able to explain the majority of the treatment effects. Thus, it seems unlikely that some unobserved skills overlooked by psychologists could have a major impact on mediating treatment effects.

I.1 Skills and the Measurement System

The assumption that the loading factors in the measurement system (Equation 27) are the same for treatment and control is not necessary to identify the model. It is useful for clarity in the interpretation because the treatment operates by the shift in latent skills and not by the map between measures and skills.

Ultimately, we need the decomposition of the treatment effects, (6), to be invariant to the choice of the measurement system we used. Thus, for each skill's contribution to treatment effect on each outcome, we want to test the null hypothesis that:

$$H_0 : \boldsymbol{\alpha}_0(\mathrm{E}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)) = \boldsymbol{\alpha}_1(\mathrm{E}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)) \tag{40}$$

where $\boldsymbol{\alpha}_d = (\alpha_d^j \ : \ j \in \mathcal{J}_p)$ and $\boldsymbol{\theta}_d = (\theta_d^j \ : \ j \in \mathcal{J}_p)$ such that $d \in \{0,1\}$ denotes treatment status.

Let $\hat{\boldsymbol{\theta}}_i$ be the estimated factor score for individual i, assigned to treatment status $D_i \in \{0,1\}$, using the estimated loading factors from the subsample of individuals with the same treatment status, i.e. for each individual factor score:

$$\hat{\boldsymbol{\theta}}_i = (\boldsymbol{\varphi}_{D_i}{}'(\boldsymbol{\Omega}_{D_i})^{-1}\boldsymbol{\varphi}_{D_i})^{-1}\boldsymbol{\varphi}_{D_i}{}'(\boldsymbol{\Omega}_{D_i})^{-1}\boldsymbol{M}_i.$$

We would like to test whether the contribution to the treatment effects is independent if we use the parameters from a different measurement system (i.e if we estimate a different set of loading factors for the treatment and control group).

Hence, an appropriate single hypothesis test statistic for each skill $j \in \mathcal{J}_p$ becomes:

$$\hat{\alpha}_0^j(\hat{\theta}_1^j - \hat{\theta}_0^j) - \hat{\alpha}_1^j(\hat{\theta}_1^j - \hat{\theta}_0^j)$$

where we use a hat superscript to denote estimated parameters. $\hat{\alpha}$ are Croon corrected estimates of α. We can use a summary statistic to test the joint hypothesis stated in (40).

Independence between $\hat{\boldsymbol{\alpha}}_d$ and $\hat{\boldsymbol{\theta}}_d - \hat{\boldsymbol{\theta}}_d$ yields:

$$\mathrm{Var}(\hat{\boldsymbol{\alpha}}_d(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0)) = (\hat{\boldsymbol{\alpha}}_d)^2 \,\mathrm{Var}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0) + \mathrm{Var}(\hat{\boldsymbol{\alpha}}_d)(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0)^2 + \mathrm{Var}(\hat{\boldsymbol{\alpha}})\,\mathrm{Var}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0)$$

Independence between the quantities estimated for each of the d's yields:

$$\mathrm{Var}(\hat{\boldsymbol{\alpha}}_0(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0) - \hat{\boldsymbol{\alpha}}^1(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0)) = \mathrm{Var}(\hat{\boldsymbol{\alpha}}_0(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0)) + \mathrm{Var}(\hat{\boldsymbol{\alpha}}_1(\bar{\hat{\boldsymbol{\theta}}}_1 - \bar{\hat{\boldsymbol{\theta}}}_0))$$

This variance helps us to get the z-statistic:

$$z = \frac{\hat{\boldsymbol{\alpha}}_0(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0) - \hat{\boldsymbol{\alpha}}_1(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0)}{\sqrt{\mathrm{Var}(\hat{\boldsymbol{\alpha}}_0(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0) - \hat{\boldsymbol{\alpha}}^1(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_0))}}$$

A two-sided z-test gives a p-value associated with the skill and outcome null hypothesis of invariance to the choice of the measurement system.

These paired (outcome, skill) p-values are shown in Tables I.1 and I.2. We find that we can not reject the null hypothesis for any skill-outcome pair, which suggests that our decompositions of the NFP treatment effects are not driven by the choice of the measurement system.

I.1.1 Additional Specification Tests for the Outcome Equations

In order to clearly interpret the channels through which the NFP affects later outcomes, (1) assumes that the parameters that map skills and pre-program variables with the outcomes are not affected by the programs. Put another way, the mediated channels operate exclusively through the program effect on skills. This assumption is not necessary to identify the model.

For each outcome decomposed, we test the hypothesis that $\alpha_1^j = \alpha_0^j, \forall j \in \mathcal{J}$ and $\boldsymbol{\beta_1} = \boldsymbol{\beta_0}$ with a Wald test. Tables I.3 and I.4 show the results of this test. We cannot reject the null hypothesis of equality of the coefficients for the treatment and control groups. This evidence strengthens the validity of our interpretation of the decomposition of the NFP treatment effect.

J Oaxaca-Blinder Decomposition Results

Oaxaca-Blinder decompositions are often used to examine sources of treatment effects. This method decomposes the difference in means between two groups (treatment and control) into the part that is due to the group differences in the channels and into the part that is due to group differences in the parameters that capture the relationship between the channels and the outcomes. In our context, the Oaxaca-Blinder decomposition is summarized as follows:[9]

$$\underbrace{E(\boldsymbol{Y}|D=1) - \mathrm{E}(\boldsymbol{Y}|D=0)}_{\text{Treatment Effects}} = \underbrace{(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_0)\boldsymbol{\theta}_0}_{\text{Differences unexplained by the skills}} + \underbrace{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)\boldsymbol{\alpha}}_{\text{Differences explained by the skills}} . \tag{41}$$

The decomposition that we propose summarizes the unexplained part in the above equation through the difference in the intercepts between the treatment and the control groups. In order to assess whether our decomposition is a plausible specification, we estimate an Oaxaca-Blinder decomposition. The results in Tables J.1 - J.5 present evidence that the unexplained component accounting for differences in the mapping of the skills on outcomes is not statistically significant for any outcome. Therefore, the results from the decomposition of the NFP treatment effects presented in the paper seem to be correctly specified.

[9] We implicitly control for pre-program variables.

Table I.1: Specification Test - Invariance of the Contribution of Skills to the Choice of the Measurement System (Females)

Factor Testing Results - Females

	Maternal Skills Age 2				
Age 6 outcomes	**Home**	**Parenting**	**anxiety**	**esteem**	**mastery**
Cognition	0.263	0.907	0.859	0.698	0.672
Attention problems	0.363	0.709	0.702	0.667	0.748
Conduct problems	0.421	0.694	0.922	0.721	0.677
Warmth-empathy (pro-social skills)	0.267	0.907	0.644	0.833	0.973
Agression Problems	0.692	0.819	0.862	0.821	0.786

	Children's Skills Age 6				
Age 12 outcomes	**Cognition**	**Attention probs**	**Conduct probs**	**Empathy**	**Agression**
SC # days ever used marijuana	0.867	0.878	0.592	0.885	0.280
SC use alc, mar, tob last 30 days	0.876	0.695	0.907	0.893	0.812
Standarized Child BMI	0.889	0.822	0.574	0.953	0.324

Notes: The table shows p-values for the Wald test: $z = \frac{\hat{\boldsymbol{\alpha}}^0(\bar{\hat{\boldsymbol{\theta}}}_1^0 - \bar{\hat{\boldsymbol{\theta}}}_0^0) - \hat{\boldsymbol{\alpha}}^1(\bar{\hat{\boldsymbol{\theta}}}_1^1 - \bar{\hat{\boldsymbol{\theta}}}_0^1)}{\sqrt{\mathrm{Var}(\hat{\boldsymbol{\alpha}}^0(\bar{\hat{\boldsymbol{\theta}}}_1^0 - \bar{\hat{\boldsymbol{\theta}}}_0^0) - \hat{\boldsymbol{\alpha}}^1(\bar{\hat{\boldsymbol{\theta}}}_1^1 - \bar{\hat{\boldsymbol{\theta}}}_0^1))}}$

Table I.2: Specification Test - Invariance of the Contribution of Skills to the Choice of the Measurement System (Males)

	Maternal Skills Age 2				
Age 6 outcomes	**Home**	**Parenting**	**anxiety**	**esteem**	**mastery**
Cognition	0.349	0.394	0.971	0.927	0.946
Agression Problems	0.928	0.959	0.950	0.843	0.537

	Children's Skills Age 6				
Age 12 outcomes	**Cognition**	**Attention probs**	**conduct probs**	**Empathy**	**Agression**
Average TCAP percetile. Years 1-5 after KG: Language	0.529	0.975	0.993	0.636	0.882
PIAT reading comprehension derived score	0.420	0.794	0.941	0.867	0.812
Average math grades. Years 1-5 after KG	0.425	0.953	0.830	0.871	0.792
Average TCAP percetile. Years 1-5 after KG: Math	0.571	0.940	0.951	0.940	0.875
PIAT mathematics derived score	0.433	0.503	0.817	0.976	0.652
SC use of alc, mar, tob. Lat 30 days	0.845	0.970	0.751	0.791	0.538
Internalizing disorders - youth report	0.582	0.934	0.911	0.905	0.509
Clinical or borderline anxious/depressed disorder	0.537	0.936	0.771	0.916	0.687
Average number of absences, school years 1-5 after KG	0.379	0.908	0.706	0.833	0.735

Notes: The table shows p-values for the Wald test: $z = \frac{\hat{\boldsymbol{\alpha}}^0(\bar{\hat{\boldsymbol{\theta}}}_1^0 - \bar{\hat{\boldsymbol{\theta}}}_0^0) - \hat{\boldsymbol{\alpha}}^1(\bar{\hat{\boldsymbol{\theta}}}_1^1 - \bar{\hat{\boldsymbol{\theta}}}_0^1)}{\sqrt{\mathrm{Var}(\hat{\boldsymbol{\alpha}}^0(\bar{\hat{\boldsymbol{\theta}}}_1^0 - \bar{\hat{\boldsymbol{\theta}}}_0^0) - \hat{\boldsymbol{\alpha}}^1(\bar{\hat{\boldsymbol{\theta}}}_1^1 - \bar{\hat{\boldsymbol{\theta}}}_0^1))}}$

Table I.3: Specification Test - Outcome Equation (Females)

Outcome	Test Stat	P-Val
6 Years		
Cognition	0.982	0.490
Attention prob.	1.753	0.018
Conduct Prob.	0.846	0.675
Pro-social	1.264	0.189
Aggression	0.558	0.955
12 Years		
SC use alc, mar, tob last 30 days	1.266	0.189
SC # days use of alc, mar, tob last 30 days	1.271	0.186
Standardized Child BMI (Year 12)	1.172	0.270

Notes: The table shows p-values for Wald tests for the equality of slopes between treatment and control group in the outcome equation.

Table I.4: Specification Test - Outcome Equation (Males)

Outcome	Test Stat	P-Val
6 Years		
Cognition	0.609	0.926
Aggression	0.881	0.628
12 Years		
average tcap percentile, y1-5: language composite	1.162	0.283
PIAT reading comprehension derived score	1.286	0.175
Average math grade. Years 1-5 after KG	1.493	**0.073**
average tcap percentile y1-5: math	1.242	0.213
PIAT math derived score	1.102	0.343
SC ever tried smoking: 1=yes	0.838	0.686
Internalizing disorders - Youth report	0.993	0.477
Anxious/depressed - clinical or borderline disorder	0.682	0.867
Average number of absences, school years 1-5	0.798	0.738

Notes: The table shows p-values for Wald tests for the equality of slopes between treatment and control group in the outcome equation.

Table J.1: Oaxaca-Blinder Decomposition, outcomes at age 6 (Females)

	Cognition				Attention Problems				Conduct Problems				Warmth/Empathy				Aggression			
	Effect	SE	*P*-Val	Fraction	Effect	SE	*P*-Val	Fraction	Effect	SE	*P*-Val	Fraction	Effect	SE	*P*-Val	Fraction	Effect	SE	*P*-Val	Fraction
Overall																				
Total Diff. in Means	0.114	0.112	0.311	-	-0.189	0.092	**0.039**	-	-0.197	0.072	**0.006**	-	0.235	0.102	**0.021**	-	-0.187	0.094	**0.046**	-
Explained	0.083	0.041	**0.044**	0.706	-0.080	0.037	**0.033**	0.271	-0.051	0.030	**0.086**	0.213	0.069	0.035	**0.050**	0.291	-0.024	0.030	0.421	0.260
Unexplained	0.031	0.112	0.784	0.294	-0.110	0.093	0.237	0.729	-0.146	0.072	**0.045**	0.787	0.166	0.099	**0.093**	0.709	-0.163	0.090	**0.071**	0.740
Explained Portion																				
Home Index	0.032	0.020	0.113	0.355	-0.016	0.015	0.279	0.096	-0.010	0.013	0.420	0.063	0.034	0.019	**0.071**	0.214	-0.015	0.015	0.344	0.160
Parenting Index	0.028	0.026	0.284	0.137	-0.048	0.024	**0.046**	0.093	-0.018	0.018	0.314	0.046	0.040	0.023	**0.074**	0.106	0.002	0.019	0.935	0.015
Maternal Anxiety Index	0.024	0.020	0.217	0.254	-0.018	0.016	0.271	0.100	-0.019	0.014	0.181	0.137	0.006	0.014	0.671	-0.008	0.009	0.014	0.536	-0.116
Maternal Self-Esteem Index	0.007	0.021	0.751	0.238	0.009	0.020	0.642	-0.204	-0.032	0.022	0.145	0.187	-0.043	0.028	0.122	-0.290	-0.022	0.021	0.283	0.131
Maternal Mastery Index	0.002	0.021	0.913	-0.192	-0.019	0.023	0.409	0.256	0.017	0.019	0.375	-0.157	0.039	0.028	0.158	0.295	-0.003	0.023	0.893	0.096
Birthweight	-0.011	0.012	0.393	-0.086	0.012	0.013	0.353	-0.070	0.011	0.012	0.361	-0.063	-0.006	0.009	0.481	-0.027	0.005	0.007	0.483	-0.027
Unexplained Portion																				
Home Index	0.008	0.024	0.746	0.068	-0.004	0.018	0.842	0.019	0.000	0.014	0.994	0.001	0.028	0.023	0.222	0.118	-0.009	0.019	0.624	0.051
Parenting Index	0.020	0.028	0.461	0.179	0.033	0.022	0.130	-0.176	-0.001	0.016	0.965	0.004	0.007	0.021	0.741	0.030	0.010	0.019	0.615	-0.051
Maternal Anxiety Index	-0.017	0.021	0.434	-0.145	0.034	0.024	0.161	-0.181	0.032	0.020	0.111	-0.161	-0.007	0.017	0.696	-0.029	0.005	0.016	0.757	-0.026
Maternal Self-Esteem Index	-0.009	0.029	0.757	-0.080	0.041	0.029	0.164	-0.216	0.022	0.022	0.306	-0.113	-0.018	0.035	0.599	-0.077	0.006	0.024	0.802	-0.033
Maternal Mastery Index	0.017	0.034	0.628	0.147	-0.046	0.036	0.200	0.244	-0.038	0.026	0.148	0.193	0.018	0.037	0.625	0.077	0.001	0.029	0.974	-0.005
Birthweight	-0.002	0.007	0.719	-0.022	0.000	0.003	0.915	-0.002	0.002	0.006	0.695	-0.011	-0.002	0.007	0.758	-0.009	0.002	0.005	0.749	-0.009
Residual	0.014	0.117	0.904	0.124	-0.169	0.094	**0.074**	0.891	-0.163	0.074	**0.028**	0.828	0.140	0.099	0.156	0.595	-0.176	0.091	**0.052**	0.943

Notes: The indices are means of the non-missing items. The fractions are proportions of the total conditional difference in means.

Table J.2: Oaxaca-Blinder Decomposition, outcomes at age 6 (Males)

	Cognition				Attention Problems				Conduct Problems				Warmth/Empathy				Aggression			
	Effect	SE	P-Val	Fraction	Effect	SE	P-Val	Fraction	Effect	SE	P-Val	Fraction	Effect	SE	P-Val	Fraction	Effect	SE	P-Val	Fraction
Overall																				
Total Diff. in Means	0.173	0.105	**0.100**	-	-0.025	0.095	0.797	-	-0.012	0.088	0.891	-	-0.080	0.104	0.442	-	-0.099	0.086	0.250	-
Explained	0.092	0.037	**0.012**	0.496	-0.064	0.033	**0.051**	-5.856	-0.056	0.035	0.111	-1.130	0.022	0.025	0.387	-0.254	-0.025	0.025	0.322	0.158
Unexplained	0.081	0.101	0.422	0.504	0.039	0.092	0.670	6.856	0.044	0.090	0.628	2.130	-0.102	0.105	0.332	1.254	-0.074	0.085	0.383	0.842
Explained Portion																				
Home Index	0.021	0.022	0.344	0.220	-0.003	0.006	0.627	-0.303	-0.009	0.010	0.395	-0.395	0.006	0.009	0.513	-0.161	0.002	0.005	0.728	-0.074
Parenting Index	0.044	0.022	**0.042**	0.112	-0.021	0.019	0.258	-0.479	-0.011	0.018	0.540	-0.041	0.008	0.017	0.640	-0.007	-0.019	0.013	0.150	0.085
Maternal Anxiety Index	0.001	0.004	0.847	0.019	-0.007	0.012	0.561	-0.544	-0.008	0.014	0.546	-0.297	0.003	0.007	0.629	-0.046	-0.011	0.018	0.538	0.125
Maternal Self-Esteem Index	-0.005	0.009	0.607	-0.016	0.005	0.010	0.592	1.536	-0.011	0.015	0.490	-0.514	-0.003	0.009	0.690	0.034	0.012	0.014	0.399	-0.230
Maternal Mastery Index	0.009	0.018	0.598	0.021	-0.015	0.018	0.385	-3.978	-0.001	0.014	0.964	0.553	0.004	0.018	0.831	-0.020	-0.003	0.013	0.833	0.197
Birthweight	0.022	0.017	0.192	0.140	-0.023	0.018	0.214	-2.089	-0.016	0.015	0.283	-0.435	0.004	0.014	0.747	-0.054	-0.007	0.010	0.534	0.055
Unexplained Portion																				
Home Index	0.002	0.011	0.834	0.013	-0.003	0.009	0.763	0.111	0.002	0.009	0.846	-0.152	0.008	0.012	0.529	-0.096	-0.002	0.009	0.836	0.019
Parenting Index	-0.007	0.024	0.760	-0.043	0.012	0.024	0.627	-0.471	0.016	0.022	0.467	-1.314	0.010	0.027	0.698	-0.129	0.034	0.021	0.110	-0.340
Maternal Anxiety Index	0.006	0.012	0.622	0.035	0.000	0.006	0.945	0.016	-0.005	0.011	0.625	0.451	0.003	0.009	0.708	-0.042	-0.004	0.008	0.627	0.040
Maternal Self-Esteem Index	0.004	0.010	0.665	0.024	-0.002	0.008	0.771	0.091	0.004	0.010	0.688	-0.321	0.005	0.010	0.642	-0.059	0.000	0.006	0.970	-0.002
Maternal Mastery Index	-0.033	0.028	0.241	-0.192	-0.007	0.026	0.780	0.294	0.020	0.023	0.396	-1.627	-0.023	0.029	0.421	0.288	0.029	0.026	0.268	-0.289
Birthweight	0.020	0.024	0.418	0.113	-0.060	0.028	**0.034**	2.422	-0.065	0.028	**0.022**	5.335	-0.015	0.024	0.525	0.191	-0.003	0.018	0.877	0.028
Residual	0.090	0.107	0.400	0.519	0.100	0.093	0.284	-4.061	0.072	0.098	0.462	-5.970	-0.090	0.116	0.436	1.121	-0.128	0.076	**0.091**	1.293

Notes: The indices are means of the non-missing items. The fractions are proportions of the total conditional difference in means.

Table J.3: Oaxaca-Blinder Decomposition, outcomes at age 12 (Females)

	SC # days ever used marijuana				SC use alc, mar, tob last 30 days				Standardized Child BMI (12Y)			
	Effect	SE	*P*-Val	Fraction	Effect	SE	*P*-Val	Fraction	Effect	SE	*P*-Val	Fraction
Total Diff. in Means	-0.141	0.085	**0.098**	-	-0.021	0.020	0.281	-	-0.197	0.110	**0.072**	-
Explained	0.026	0.038	0.488	-0.184	-0.005	0.007	0.460	0.242	0.020	0.038	0.601	-0.102
Unexplained	-0.167	0.111	0.133	1.184	-0.016	0.021	0.429	0.758	-0.217	0.114	**0.056**	1.102
Explained												
Cognitive	-0.006	0.018	0.723	0.046	-0.001	0.003	0.708	0.058	-0.004	0.013	0.777	0.019
Attention Problems	0.000	0.013	0.991	0.001	-0.001	0.005	0.827	0.051	0.016	0.017	0.354	-0.081
Conduct Problems	-0.003	0.010	0.775	0.019	-0.003	0.004	0.468	0.149	-0.027	0.022	0.215	0.138
Warmth/Empathy	0.038	0.038	0.318	-0.268	0.000	0.004	0.934	-0.015	0.023	0.023	0.331	-0.115
Aggression	-0.003	0.005	0.644	0.018	0.000	0.002	0.988	-0.001	0.013	0.021	0.554	-0.064
Unexplained												
Cognitive	0.012	0.018	0.523	-0.084	0.003	0.004	0.496	-0.135	0.004	0.015	0.773	-0.022
Attention Problems	0.002	0.015	0.889	-0.015	0.008	0.007	0.288	-0.364	-0.036	0.034	0.293	0.181
Conduct Problems	0.013	0.018	0.482	-0.091	-0.008	0.012	0.522	0.364	0.095	0.047	**0.043**	-0.481
Warmth/Empathy	-0.029	0.032	0.374	0.202	-0.005	0.004	0.254	0.228	-0.026	0.023	0.260	0.132
Aggression	-0.002	0.009	0.803	0.016	-0.001	0.003	0.697	0.062	0.013	0.037	0.724	-0.066
Residual	-0.163	0.096	**0.089**	1.155	-0.013	0.027	0.629	0.603	-0.268	0.122	**0.028**	1.358

Notes: The indices are means of the non-missing items. The fractions are proportions of the total conditional difference in means.

Table J.4: Oaxaca-Blinder outcomes at age 12, Decomposition Part 1 (Males)

	language composite				derived score				Average math grade grades 1-5				after KG				average tcap percentile y1-5: math				PIAT math derived score			
	Effect	SE	*P*-Val	Fraction	Effect	SE	*P*-Val	Fraction	Effect	SE	*P*-Val	Fraction	Effect	SE	*P*-Val	Fraction	Effect	SE	*P*-Val	Fraction	Effect	SE	*P*-Val	Fraction
Total Diff. in Means	4.403	3.289	0.181	-	2.022	1.584	0.202	-	0.117	0.107	0.275	-	0.083	0.103	0.422	-	2.818	3.429	0.411	-	2.447	1.324	**0.065**	-
Explained	1.564	1.520	0.304	0.355	0.941	0.808	0.244	0.466	0.077	0.063	0.219	0.660	0.067	0.058	0.242	0.814	2.226	1.820	0.221	0.790	0.946	0.817	0.247	0.387
Unexplained	2.839	3.105	0.361	0.645	1.080	1.454	0.457	0.534	0.040	0.090	0.658	0.340	0.015	0.087	0.860	0.186	0.592	3.111	0.849	0.210	1.501	1.103	0.174	0.613
Explained																								
Cognitive	1.831	1.337	0.171	0.416	0.910	0.705	0.197	0.450	0.057	0.054	0.285	0.492	0.057	0.051	0.266	0.682	2.146	1.664	0.197	0.762	0.939	0.750	0.211	0.384
Attention Problems	0.075	0.275	0.787	0.017	0.060	0.131	0.647	0.030	0.009	0.016	0.570	0.080	0.007	0.012	0.548	0.087	-0.057	0.312	0.855	-0.020	0.067	0.123	0.589	0.027
Conduct Problems	-0.003	0.273	0.992	-0.001	0.037	0.122	0.765	0.018	0.000	0.010	0.964	-0.004	-0.001	0.008	0.933	-0.008	0.020	0.281	0.945	0.007	0.035	0.110	0.751	0.014
Warmth/Empathy	-0.535	0.468	0.253	-0.122	-0.079	0.156	0.611	-0.039	-0.009	0.013	0.475	-0.079	-0.012	0.014	0.362	-0.149	-0.063	0.385	0.870	-0.022	-0.094	0.122	0.441	-0.038
Aggression	0.197	0.419	0.639	0.045	0.014	0.161	0.930	0.007	0.020	0.018	0.267	0.171	0.017	0.017	0.328	0.203	0.181	0.439	0.681	0.064	-0.001	0.142	0.995	0.000
Unexplained																								
Cognitive	0.540	0.980	0.582	0.123	-0.015	0.288	0.958	-0.007	0.005	0.019	0.787	0.045	0.009	0.019	0.652	0.105	0.563	0.780	0.470	0.200	-0.016	0.228	0.943	-0.007
Attention Problems	0.699	0.769	0.363	0.159	0.449	0.342	0.189	0.222	0.009	0.020	0.659	0.077	0.006	0.020	0.752	0.075	0.518	0.663	0.435	0.184	-0.162	0.292	0.579	-0.066
Conduct Problems	0.003	0.427	0.995	0.001	-0.092	0.242	0.705	-0.045	-0.003	0.016	0.858	-0.024	-0.003	0.015	0.864	-0.031	0.002	0.402	0.995	0.001	0.004	0.183	0.983	0.002
Warmth/Empathy	0.373	0.691	0.590	0.085	0.134	0.271	0.622	0.066	0.011	0.020	0.588	0.092	0.012	0.021	0.559	0.150	-0.076	0.650	0.907	-0.027	0.070	0.222	0.752	0.029
Aggression	-0.078	0.495	0.875	-0.018	-0.071	0.270	0.793	-0.035	-0.006	0.017	0.740	-0.049	-0.001	0.013	0.933	-0.014	-0.266	0.614	0.664	-0.094	-0.168	0.231	0.468	-0.069
Residual	1.302	3.447	0.706	0.296	0.675	1.617	0.676	0.334	0.023	0.090	0.796	0.200	-0.008	0.089	0.927	-0.099	-0.150	3.289	0.964	-0.053	1.773	1.191	0.137	0.724

Notes: The indices are means of the non-missing items. The fractions are proportions of the total conditional difference in means.

Table J.5: Oaxaca-Blinder outcomes at age 12, Decomposition Part 2 (Males)

	SC ever tried smoking: 1=yes				SC use alc, mar, tob last 30 days				Internalizing disorders - Youth				Anxious/depressed - clinical or				Average number of absences			
	Effect	SE	*P*-Val	Fraction	Effect	SE	*P*-Val	Fraction	Effect	SE	*P*-Val	Fraction	Effect	SE	*P*-Val	Fraction	Effect	SE	*P*-Val	Fraction
Total Diff. in Means	-0.063	0.034	**0.065**	-	-0.034	0.025	0.176	-	-0.068	0.063	0.281	-	-0.053	0.030	**0.081**	-	-1.017	0.931	0.275	-
Explained	-0.005	0.010	0.644	0.076	-0.003	0.010	0.741	0.099	-0.030	0.022	0.163	0.445	-0.014	0.012	0.248	0.262	-0.524	0.346	0.130	0.515
Unexplained	-0.058	0.036	0.103	0.924	-0.031	0.025	0.231	0.901	-0.038	0.062	0.546	0.555	-0.039	0.030	0.192	0.738	-0.493	0.918	0.591	0.485
Explained																				
Cognitive	0.001	0.004	0.727	-0.021	0.000	0.003	0.886	0.013	-0.006	0.009	0.477	0.090	-0.006	0.006	0.330	0.108	-0.237	0.228	0.298	0.233
Attention Problems	-0.004	0.005	0.505	0.057	0.001	0.003	0.789	-0.023	-0.004	0.008	0.588	0.062	0.000	0.002	0.890	0.006	-0.212	0.237	0.370	0.209
Conduct Problems	-0.001	0.005	0.777	0.022	-0.001	0.004	0.740	0.042	-0.002	0.007	0.811	0.024	-0.001	0.004	0.742	0.026	0.056	0.180	0.758	-0.055
Warmth/Empathy	0.002	0.004	0.569	-0.034	0.004	0.004	0.296	-0.117	-0.002	0.007	0.762	0.030	0.001	0.004	0.811	-0.018	-0.125	0.131	0.339	0.123
Aggression	-0.003	0.006	0.556	0.054	-0.006	0.008	0.459	0.184	-0.016	0.015	0.267	0.239	-0.007	0.007	0.313	0.140	-0.004	0.119	0.970	0.004
Unexplained																				
Cognitive	-0.001	0.005	0.906	0.010	-0.006	0.007	0.382	0.187	-0.003	0.012	0.799	0.046	-0.001	0.007	0.879	0.020	0.103	0.135	0.444	-0.102
Attention Problems	-0.004	0.010	0.650	0.069	0.000	0.005	0.980	-0.003	-0.007	0.016	0.657	0.107	-0.001	0.005	0.907	0.010	-0.101	0.199	0.612	0.099
Conduct Problems	0.002	0.006	0.783	-0.025	0.002	0.005	0.734	-0.047	0.001	0.010	0.960	-0.007	0.002	0.006	0.765	-0.034	0.003	0.096	0.977	-0.003
Warmth/Empathy	-0.007	0.007	0.335	0.105	0.000	0.004	0.913	0.012	0.014	0.014	0.338	-0.203	0.003	0.008	0.678	-0.064	-0.051	0.205	0.803	0.050
Aggression	0.001	0.006	0.872	-0.014	0.006	0.009	0.520	-0.171	0.004	0.014	0.742	-0.066	0.001	0.008	0.932	-0.012	0.054	0.169	0.749	-0.053
Residual	-0.049	0.042	0.239	0.780	-0.031	0.026	0.235	0.923	-0.046	0.069	0.503	0.679	-0.044	0.035	0.211	0.818	-0.501	1.072	0.640	0.493

Notes: The indices are means of the non-missing items. The fractions are proportions of the total conditional difference in means.

K Summary of Previous Analyses of NFP

In this section, we summarize the findings from previous studies that examine the treatment effects of the NFP by each of the three trials. Tables K.1-K.8 present the studies for Elmira; Tables K.9-K.13 for Memphis and Tables K.14-K.15 for Denver.

Table K.1: Summary of Olds et al. (1986), Elmira Trial

A. Paper Title

Improving the Delivery of Prenatal Care and Outcomes of Pregnancy: A Randomized Trial of Nurse Home Visitation

B. Period of Investigation

Time of registration in the program, at the 32nd week of pregnancy and medical records at labor delivery

C. Sample Size

500 women invited, 400 enrolled. Comparison: 165 (group 1 and 2). Treatment: 189 (Group 3 and 4).
From the initial 400 women enrolled, 46 non-white women were removed because of the small sample sizes (when conditioned on other pre-program variables of interest).

D. Main Goal

Evaluation of the effectiveness of the comprehensive prenatal program as means of improving antepartum social support, health habits and obstetrician health status on on length of gestation and birth weight

E. Outcomes

Use of services and support systems, obstetrician complication after enrollment, obstetrician conditions and health habits, number of cigarettes, birth weight and length of gestation

F. Methods

Differences in means. OLS for continuous outcomes and logistic linear model for dichotomous outcomes

G. Main Results

Nurse home visited group improved in the use of community services, informal social support, and health habits. No overall effect on either birth weight or length of gestation. But, positive effects were present for the children of young adolescents (< 17) and smokers

Table K.2: Summary of Olds et al. (1986), Elmira Trial

A. Paper Title

Preventing Child Abuse and Neglect: A Randomized Trial of Nurse Home Visitation

B. Period of Investigation

Time of registration in the program, at 6, 12, 24 months of the child's life.

C. Sample Size

Comparison: 165 (group 1 and 2) Treatment: 189 (Group 3 and 4). From the initial 400 women enrolled 46 non-white women were removed in this analysis because of the small number to cross-classify race with other variables important for the statistical analysis. In the 2 years of child's life attrition between 12% and 21%

D. Main Goal

Effect of prenatal program on childhood health and developmental problems in the 2 years of child's life, including abuse and neglect

E. Outcomes

Child abuse and neglect.
Reports of infant temperament, behavior problems, and maternal reaction to behavioral problems.
Restriction and punishment and provision of play material.
Infant mental development (Bayley and Cattell)
Emergency room visits.

F. Methods

Differences in means (Adjusted for baseline characteristics). Simultaneous statistical inference. For continuous outcomes, OLS and for dichotomous outcomes logistic linear model (Assuming a binominal distribution). And low incidence outcomes, in the form of counts (number of emergency room visits) in the log-linear model (assuming a poisson distribution)

G. Main Results

Positive results concentrated among women at greater risk (younger mothers, poor, unmarried) of caregiving dysfunction. These group had fewer records of child abuse and neglect during the first two years of child lives; they punished their children less; they provided with more playing material. Children had less emergency room visits. For the infants of all the nurse-visited women: they visited the emergency room less, they were seen by physicians less frequently for accidents and poisoning in the second years of life

Table K.3: Summary of Olds et al. (1988), Elmira Trial

A. Paper Title

Improving the Life-Course Development of Socially Disadvantaged Mothers: A Randomized Trial of Nurse Home Visitation

B. Period of Investigation

Time of registration in the program, at 6, 10, 22, 46 months of children life. SSA records of number of days that women and their children received public assistance from the index child's birth to fourth birthday

C. Sample Size

Comparison: 165 (Groups 1 and 2).
Treatment: 189 (Group 3 and 4).
During the first 4 years of child's life attrition was between 15% and 21%

D. Main Goal

Effect on improving maternal life-course development

E. Outcomes

Mother's educational achievement (enrollment, graduation, years of schooling)
Employment
Child Care
Public assistance.
Subsequent-pregnancy

F. Methods

Differences in means (Adjusted for baseline characteristics, classification factors and interaction). For continuous outcomes, OLS and for dichotomous outcomes logistic linear model (Assuming a binominal distribution). And low incidence outcomes, in the form of counts (number of emergency room visits) in the log-linear model (assuming a poisson distribution)

G. Main Results

Up through the four year old index child's life, the nurse visited women who had not graduated from high school returned to school more rapidly than the comparison group. Treated poor, unmarried women showed an 82% increase in the number of months employed, had 43% fewer subsequent pregnancies, and postponed the birth of the second child on average by 12 months. During the first two years after delivery, nurse-visited, poor unmarried older women received 40% less of public assistance than comparison group

Table K.4: Summary of Olds et al. (1994), Elmira Trial

A. Paper Title

Intellectual Impairment in Children of Women who Smoke Cigarettes During Pregnancy

B. Period of Investigation

Time of registration in the program, 34th week of gestation, measures at 6, 10, 22, 36, 48 months of the child's life

C. Sample Size

Comparison: 165 (Groups 1 and 2).
Treatment: 189 (Groups 3 and 4).
During the first 4 years of child's life attrition was between 15% and 21%.
Analysis limited to whites.
From the initial 400 women enrolled 46 non-white women were removed in this analysis because of the small number to cross-classify race with other variables important for the statistical analysis. The estimation of the effect of smoking focused on the comparison sample because the nurse visited group altered the relationship prenatal smoking and Children IQ

D. Main Goal

Study the effect of maternal cigarette smoking during pregnancy on children's intellectual functioning during the first 4 years of life, adjusting for the primary confounding influences

E. Outcomes

Intellectual functioning scores: Bayley mental development index (12 months), Cattell (24 months), Stanford-Binet (36 months and 48 months)

F. Methods

General linear model methods, including mixed models to analyse repeated measures with missing data. Newton Raphson and EM algorithms. Adjustment for baseline characteristics, classification factors (marital status, SES), covariates and their interactions. To analyze the effect of smoking, the comparison is made between women in the comparison group who smoke 10 or more cigarettes per day during pregnancy and comparison women who smoke 0

G. Main Results

Children in the comparison group whose mothers smoke 10 or more cigarettes per day during pregnancy had Stanford-Binet scores at 3 and 4 years that were 4.35 points lower (after controlling for several variables) than their counterparts who did not smoke prenatally

Table K.5: Summary of Olds et al. (1994), Elmira Trial

A. Paper Title

Does Prenatal and Infancy Nurse Home Visitation Have Enduring Effects on Qualities of Parental Caregiving and Child Health at 25 to 50 Months of Life?

B. Period of Investigation

Time of registration in the program, and at 34, 36, 46, and 48 months of the child's life.

C. Sample Size

Comparison: 165 (group 1 and 2)
Treatment: 189 Group 3 and 4.
During the first 4 years of child's life attrition was between 15% and 21%.
Analysis limited to whites.
From the initial 400 women enrolled 46 non-white women were removed in this analysis because of the small number to cross-classify race with other variables important for the statistical analysis

D. Main Goal

Examine the effect of a randomized trial of a nurse home visitation program on the health, development, rates of child maltreatment, and living conditions of children from 3 to 4 years of age

E. Outcomes

Cases of abuse and neglect
Intellectual functioning: Stanford-Binet
Home hazards
Health care encounters
Home inventory

F. Methods

Differences in means (Adjusted for baseline characteristics, classification factors, covariates and interactions). For continuous outcomes, OLS and for dichotomous outcomes logistic linear model (Assuming a binominal distribution). And low incidence outcomes, in the form of counts (number of emergency room visits) in the log-linear model (assuming a poisson distribution)

G. Main Results

No treatment differences in the rates of child abuse and neglect children's intellectual function from 25 to 48 moths of age. However, nurse-visited children lived in homes with fewer hazards, they had 40% fewer injuries and 45% fewer behavioral and parental coping problems. They made 35% fewer visits to the emergency room. treatment mothers were more involved with and punished their children to a greater extend than comparison mothers. The functional meaning of punishments is different between the treatment group and the comparison group

Table K.6: Summary of Olds et al. (1997), Elmira Trial

A. Paper Title

Long-term Effects of Home Visitation on Maternal Life Course and Child Abuse and Neglect: Fifteen-Year Follow-up of a Randomized Trial

B. Period of Investigation

Time of registration in the program, and at 15 years of the child's life

C. Sample Size

Of the 400 pregnant women who enrolled,
324 participated in the fifteen year follow up.
Comparison group (Groups 1 and 2)
Treatment group (Group 4)

D. Main Goal

Evaluate the long-term effects of the program on women's life course and child abuse and neglect

E. Outcomes

*Rates of subsequent births (self-report) *Use of welfare (AFDC, food stamps, medicaid, self report) *Maternal substance abuse, arrests, convictions, and child abuse and neglect reports from birth up to 15 years of child life (New York State records)

F. Methods

Intent to treat approach. Differences in means (Adjusted for baseline characteristics, classification factors, covariates and interactions). For continuous outcomes, OLS was used and for low frequency count data (number of reports pf child maltreatment) the log-linear model was used (assuming a poisson distribution). The analysis reported here was not limited to one race

G. Main Results

Women visited by nurses during pregnancy and infancy were involved in fewer child abuse and neglect episodes than comparison group women. Among unmarried and low SES women at initial of enrollment, treated women had 1.3 vs. 1.6 subsequent births (in contrast to comparison group), 65 vs 37 months between the birth of the first child and the second, 60 vs 90 months receiving AFDC, 0.41 vs 0.73 behavioral impairments due to alcohol and drugs, 0.18 vs 0.58 arrest by self report, and 0.16 vs 0.90 arrest according to the state. All differences are significant at the 95% confidence level

Table K.7: Summary of Olds et al. (1998), Elmira Trial

A. Paper Title

Long-term Effects of Nurse Home Visitation on Children's Criminal and Antisocial Behavior: 15-Year Follow-up of a Randomized Controlled Trial

B. Period of Investigation

Time of registration in the program, and at 15 years of the child's life

C. Sample Size

400 pregnant women enrolled. A total of 315 adolescent offspring participated in the 15 years follow up study. Comparison group (1 and 2) and treatment group (group 3 and 4 separately)

D. Main Goal

Evaluate the long-term effects of the program on children's criminal and antisocial behavior

E. Outcomes

Children's self-reports of running away, arrests, convictions, initiation of sexual intercourse, number of sex partners, pregnancy, and use of illegal substances.
School records of suspensions.
Teachers'reports of children's disruptive behavior in school.
Parents'reports of the children arrests and behavioral problems

F. Methods

Intent to treat approach. Differences in means (Adjusted for baseline characteristics, classification factors, covariates and some interactions). For continuous outcomes, OLS was used and for low frequency count data (eg, number of arrests) the log-linear model was used (assuming a poisson distribution). Low incidence count outcomes with values higher than 20 were analyzed in a log-linear model, correcting for over-dispersion. For outcomes reported by more than one respondent (eg, child, teacher), they used repeated measured analysis (adding fixed factors for respondent and random factor for individual). For children self-reports of antisocial and delinquent acts, they used factor analysis and created two factors for multiple hypothesis testing: major delinquency and minor antisocial acts. The analysis reported here was not limited to one race

G. Main Results

Adolescents born to women who received the program during pregnancy and infancy and who were unmarried and from low SES at registration, in contrast to the comparison group, reported lower incidence of running away (0.24 vs 0.60), fewer arrests (0.20 vs 0.45), fewer convictions (0.09 vs 0.47), fewer lifetime sexual partners (0.92 vs 2.48), fewer cigarettes per day (1.50 vs 2.50), and fewer days of alcohol consumption (1.09 vs 2.49). Parents in the treatment group (4) reported that their children had fewer problems related to alcohol and drugs use (0.15 vs 0.35). Differences statistically significant. No effect on teachers' reports, short-term or long term suspensions, adolescent initiation of sexual life , and the two factors: major delinquency and minor antisocial acts

Table K.8: Summary of Eckenrode et al. (2010), Elmira Trial

A. Paper Title

Long-term Effects of Prenatal and Infancy Nurse Home Visitation on the Life Course of Youths: 19 year follow up

B. Period of Investigation

Time of registration in the program, and at 19 years of the child's life

C. Sample Size

400 pregnant women enrolled. A total of 310 adolescent offspring participated in the 19 years follow up study. Comparison group (1 and 2) and treatment group (group 3 and 4 separately)

D. Main Goal

Evaluate the impact of the prenatal and infancy nurse visits on youths' life course development

E. Outcomes

Youth self reports of educational achievement, reproductive behaviors, welfare use, criminal involvement, and drug use

F. Methods

Differences in means (Adjusted for baseline characteristics, classification factors, covariates). For continuous outcomes, OLS was used. For dichotomous outcomes, generalized linear model with log link and binomial error distributions was used. For count data, log link and negative binomial was assumed. To estimate the hazard of first arrest, the Cox proportional hazards method was used. Growth curves for arrest episodes over time were estimated in a generalized mixed model with cubic age, with log link and negative binomial error. The analysis reported here was not limited to one race

G. Main Results

In contrast to the comparison group, girls born to women in the pregnancy and infancy nurse-visited group were less likely to be arrested (10% vs 30%), and convicted (4% vs 20%) and had fewer lifetime arrests (0.10 vs 0.54) and convictions (0.04 vs 0.37).
Nurse-visited girls born to unmarried and low SES mothers had fewer children and were less likely to use Medicaid use than their comparison group counterparts

Table K.9: Summary of Kitzman et al. (1997), Memphis Trial

A. Paper Title

Effect of Prenatal and Infancy Home Visitation by Nurses on Pregnancy Outcomes, Childhood Injuries, and Repeated Childbearing: A Randomized Controlled Trial

B. Period of Investigation

Time of registration in the program, at 28th and 36th week of pregnancy,and at 6, 12, 24 months of the child's life. Medical and social service records were abstracted

C. Sample Size

1290 women invited, 1139 enrolled. Comparison group 1 (166), Comparison group 2 (515), treatment group 3 (230), treatment group 4 (228)

D. Main Goal

To examine the impact of pregnancy and infancy home visits by nurses on pregnancy-induced hypertension, pre-term delivery, and low birth weight; on children's injuries, immunizations, mental development, and behavioral problems; and on maternal life course

E. Outcomes

Medical records: Pregnancy-induced hypertension (PIH), preterm delivery, low birth weight, children's injuries, ingestions and immunizations.
Mothers' reports of children's behavioral problems;
Children mental development (Bayley scales and Achenbach Child Behavior Checklist)
Mothers'reports of subsequent pregnancy, educational achievement, and labor force participation
Use of welfare: AFDC, from state records

F. Methods

Differences in means (Adjusted for baseline characteristics, classification factors, covariates and interactions). For continuous outcomes, OLS and for dichotomous outcomes (eg, PIH) logistic linear model (Assuming a binominal distribution).
And low incidence outcomes, in the form of counts (number of health care encounters) in the log-linear model (assuming a Poisson distribution). Pregnancy models contrast comparison group (group 1 and 2) vs treatment group 3, and vs treatment group 4.
Postnatal models contrast comparison group (1 and 2) with treatment group 4 (the one that received both prenatal and infancy nurse visits)

G. Main Results

Women visited by nurses during pregnancy had had less PIH (13% vs 20%) compared to the comparison group. During the first two years after delivery, women in the treatment group had fewer health care encounters for children in which injuries were detected (0.43 vs 0.55), fewer days of children's hospitalization (0.03 vs 0.16), and fewer second pregnancies (36% vs 47%).
No program effects on pre-term delivery, or low birth weight; children's immunization rates, mental development or behavioral problems; or mothers' education and employment

Table K.10: Summary of Kitzman et al. (2000), Memphis Trial

A. Paper Title

Enduring Effects of Nurse Home Visitation on Maternal Life Course: A 3-Year Follow-up of a Randomized Trial

B. Period of Investigation

Data from assessments at time of registration in the program and 54th months of the child's life

C. Sample Size

Of those cases randomized with no fetal or child death, follow up interviews were completed on 91% of the cases (443 in comparison group 2 and 203 in treatment group 4)

D. Main Goal

Effectiveness of the NFP prenatal and infancy home visitation program on the maternal life course 3 years after the program ended

E. Outcomes

Mothers: Rate of subsequent pregnancy, mean of interval between first and second child, educational achievement, number of months in the labor force, and number of months enrolled in AFDC, food stamps (FS), Medicaid, WIC.
Administrative data from the Tennessee Dept. of Social Service were obtained for AFDC and Food Stamp.

F. Methods

Intent to treat approach. Differences in means (classification factors, covariates and interactions). For continuous outcomes, OLS was used, for dichotomous outcomes (eg, cohabitation) the logistic linear model (assuming binominal distribution) and for low frequency count data (subsequent pregnancies) the log-linear model was used (assuming a poisson distribution) Models focused on contrasting comparison group 2 with treatment group 4

G. Main Results

Contrasted with women in the control group, women who received the NFP treatment had fewer pregnancies (1.15 vs 1.34), fewer closely spaced subsequent pregnancies (0.22 vs 0.32), longer intervals between the birth of the first and second child (30.25 vs 26.60), and fewer months using AFDC (32.55 vs 36.29) and FS (41.57 vs 45.04). Compared with the effect of the program while the visits were being conducted, the effect after it ended was essentially equal for AFDC, greater for FS, and greater for rates of closely spaced pregnancies.

Table K.11: Summary of Olds et al. (2004), Memphis Trial

A. Paper Title

Effects of Nurse Home-Visiting on Maternal Life Course and Child Development: Age 6 Follow-Up Results of a Randomized Trial

B. Period of Investigation

Data from assessments at time of registration in the program and 6 years of the child's life

C. Sample Size

Of those cases randomized with no fetal or child death, 6 yr follow up interviews were completed on 91% of the mothers (444 in comparison group 2 and 197 in treatment group 4) and 88% of the children (425 in comparison group 2 and 190 in treatment group 4)

D. Main Goal

Effectiveness of the NFP on mothers' fertility and economic self-sufficiency and the academic and behavioral adjustment of their children as they finish kindergarden near their sixth birthday

E. Outcomes

Mother: Number and timing of subsequent pregnancies; months of employment; use of welfare; food stamps; Medicaid; rates of marriage, cohabitation, and duration of relationships; Child Educational Achievement; Behavioral problems resulting from illegal substances;
Children's behavioral problems (Achenbach Child Behavior Check list), responses to story stems, Intellectual functioning (Kaufman Assessment Battery and Peabody Picture Vocabulary), Receptive language and academic achievement; Teachers completed the High-tower Teacher-Child Rating Scales

F. Methods

Differences in means (Adjusted for baseline characteristics, classification factors, covariates and some interactions). For continuous outcomes that didn't violate the normality assumption, OLS was used and for dichotomous correlated outcomes they used generalized estimating equations with logit link function. The timing of the first subsequent birth was explored using proportional-hazards analysis. For teaches' reports of children's classroom behavior, children's representation of aggressive behavior and parental warmth/empathy factors were obtained using principal components analysis

G. Main Results

Women visited by nurses had fewer subsequent pregnancies (1.16 vs 1.38) and births (1.08 vs 1.28), longer intervals between births (34.28 vs 30.23), longer relationships with current partners (54.36 vs 45 months), and since the previous follow up, fewer months of using AFDC (7.21 vs 8.96) and FS (9.67 vs 11.50) than control group mothers. Nurse visited children were more likely to have been enrolled in formal out of home care between 2 and 4.5 years (82% vs 74.9%). Children visited by nurses demonstrated higher intellectual functioning (scores 92.34 vs 90.24) and receptive vocabulary scores (84.32 vs 82.13) and fewer behavioral problems in the borderline or clinical range (1.8% vs 5.4%). For the cases of mother with low levels of psychological resources, children had higher achievement test scores, and expressed less aggression and incoherence in response to story stems. No statistically significant effect on women's education, duration of employment, rates of marriage, being in a partnered relationship, behavioral problems related to alcohol or drug abuse

Table K.12: Summary of Olds et al. (2007), Memphis Trial

A. Paper Title

Effects of Nurse Home-Visiting on Maternal and child functioning: Age-9 Follow-Up of a Randomized Trial

B. Period of Investigation

Data from assessments at time of registration in the program and 9 years of the child's life. However, whenever possible they use data from earlier phases

C. Sample Size

From the initially 743 primary black women randomize to comparison group 2 and treatment group 4 (Core of posnatal evaluations), follow up assessments at child age 9 were completed by 91% of the mothers, school records were obtained for 88% of the children and achievement test scores for 83% of the children.

D. Main Goal

To examine the impact of pregnancy and infancy home visits by nurses mothers' fertility and children development

E. Outcomes

Mothers: interval between births, number of children born per year, mothers' stability of relationships, use of welfare, FS and Medicaid, mother's use of substances, mothers' arrest and incarcerations.
Child: academic achievement (GPAs, Tennessee Comprehensive Assessment Test), school conduct, and mental disorders.
Secondary outcomes: women's employment, experience of domestic violence and children's mortality

F. Methods

Differences in means (Adjusted for baseline characteristics, classification factors, covariates and some interactions). For outcomes on which there are multiple assessments for each mother or child, mixed models were used. This is the first time in this trial where they examined the full longitudinal effects of some of the maternal outcomes. Quantitative dependent variables were analyzed using OLS; and, dichotomous outcomes using logit model. Low-frequency outcomes were analyzed in generalized linear models with negative binominal error and log link assumptions. Factor analysis was used to summarize the information from children Social Competence Scale, the Social Health Profile and the Teachers Observation of Child Adjustment revisited. 3 indices were produced: antisocial behavior, academically focused behavior, and peer affiliation.

G. Main Results

Nurse-visited women had longer intervals between births, few cumulative subsequent births per year, and longer relationships with current partners. From birth through child age 9, treated mothers used AFDC and FS for fewer months. Nurse-visited children whose mothers have low psychological resources, had better GPA and achievement test scores in math and reading in grades 1 through 3.

Table K.13: Summary of Kitzman et al. (2010), Memphis Trial

A. Paper Title

Enduring Effects of Prenatal and Infancy Home Visiting by Nurses on Children: Follow-up of a Randomized Trial Among children at Age 12

B. Period of Investigation

Data from assessments at time of registration in the program and 12 years of the child's life

C. Sample Size

From the initially 743 primary black women randomized to comparison group 2 and treatment group 4 (Core of posnatal evaluations), follow up assessments at child age 12 were abstracted for 613 children

D. Main Goal

To evaluate the impact of a nurse visiting program on 12-year-old first born children's use of substances, behavioral adjustment, and academic achievement

E. Outcomes

Use of cigarettes, alcohol, and marijuana;
Internalizing, externalizing and total behavior problems from parents', teachers' and children's reports.
Academic achievement.
Reading and math achievement using the Peabody Individual Achievement Tests (PIATS).
Reading and Math GPA from grade 1 to 6. Reading and Math from the Tennessee Comprehensive Assessment Program (grade 1 to 6).
Arrests reported by mother and child

F. Methods

Differences in means (Adjusted for baseline characteristics, classification factors, covariates and some interactions). For outcomes on which there are multiple assessments for each child (eg, GPAs), mixed models were used. Quantitative dependent variables were analyzed using OLS; and, dichotomous outcomes using logit model. Low-frequency outcomes were analyzed in generalized linear models with negative binominal error

G. Main Results

Nurse-visited children reported fewer days of using cigarettes, alcohol, and marijuana and were less likely to report the presence of internalizing disorders that met the borderline or clinical threshold compared to the control group kids. Treated children born to mothers with low psychological resources compared to the control group, had higher scores on PIATs, and on group-administered standardized tests of math and reading. No statistically significant program effects were found on children's externalizing or total behavior problems

Table K.14: Summary of Olds et al. (2002), Denver Trial

A. Paper Title

Home Visiting by Paraprofessionals and by Nurses: a randomized controlled trial

B. Period of Investigation

Time of registration in the program, and 36th week of pregnancy, and at 6, 12, 15, 21, 24 months of the child's life

C. Sample Size

1178 women invited, 735 enrolled and randomized. Control group (255), Paraprofessional group (245), and Nurse group (235)

D. Main Goal

To evaluate the effectiveness of home visiting by paraprofessionals and by nurses as separate means of improving maternal and child health when both types of visitors are trained with the same program model

E. Outcomes

Mothers: Urine cotinine over the course of pregnancy; women's use of auxiliary services during pregnancy, subsequent pregnancies and births, educational achievement, labor market participation, and use of welfare.
Mother-infant reponsive interactions; family home environments.
Infants emotional vulnerability in response to a fear stimuli and low emotional vitality in response to joy and anger stimuli; children's language and mental development index (MDI), temperament and behavioral problems

F. Methods

Differences in means (Adjusted for baseline characteristics, classification factors, covariates and some interactions). Comparisons between nurse vs control and paraprofessional vs control. Quantitative dependent variables were analyzed using OLS; and, dichotomous outcomes using logit model. For outcomes on which there are more than one observation in time (eg, maternal-child interaction and home environment), repeated measures were used, adding a fixed factor for time and random factor for individuals. The timing of subsequent pregnancy was analyzed with proportional hazards analysis. Factor analysis of measures for maternal and infancy interaction identified a single internally consisted factor: responsive interaction

G. Main Results

Mother-child pairs in the paraprofessional group in which the mother had low psychological resources interacted with one another more responsively than the control group. There are no statistically significant paraprofessional effects. Nurse-visited smokers, compared with the control group women, had larger reductions in nicotine; by the index child second birthday, women visited by nurses had fewer subsequent pregnancies; they delayed subsequent pregnancies for longer time, and they worked more during the second year of index child life. Mother-child pairs in the nurse group interacted with one another more responsively than the control group pairs. Nurse-visited children exhibit less emotional vulnerability. Nurse-visited children born to women with low psychological resources were less likely to exhibit low emotional vitality and language delays, and had higher MDI scores. No statistically significant effects on mothers' use of prenatal services, educational achievement, use of welfare or their children behavior problems

Table K.15: Summary of Olds et al. (2004), Denver Trial

A. Paper Title

Effects of Home Visits by Paraprofessionals and by Nurses: Age 4 Follow-Up Results of a Randomized Trial

B. Period of Investigation

Data from assessments at time of registration in the program and 48th months of the child's life

C. Sample Size

From the initial 735 mothers randomized, 635 completed 4-y interviews, and 605 completed 4-y child assessments

D. Main Goal

To evaluate the effects of prenatal and infancy home visiting by paraprofessionals and nurses from child age 2 through age 4

E. Outcomes

Mothers: Subsequent pregnancies, participation in education and work, use of welfare, marriage, cohabitation, domestic violence, mental health, substance abuse, and sense of mastery.
Mother-child interaction and home environment.
Children: tests of language and executive functioning, mother's report of child externalizing behavior

F. Methods

Differences in means (Adjusted for baseline characteristics, classification factors, covariates and some interactions). Comparisons between nurse vs control and paraprofessional vs control. Quantitative dependent variables were analyzed using OLS; and, dichotomous outcomes using logit model. The timing of subsequent pregnancy was analyzed with proportional hazards analysis. Principal component analysis was used to create factors from mothers and children externalizing behavior reports (one factor), cognitive tasks on children's ability to sustain attention and inhibitory control (single factor); examination of children's ability to regulate their behavior and emotion (two factors).

G. Main Results

In general, there are greater effects for paraprofessional-visited mothers than nurse-visited mothers, while greater effects for children in nurse-visited families than in paraprofessional ones.
Paraprofessional: Women were less likely to be married (compared to control), work more and reported better mental health and mastery, had fewer subsequent miscarriages and low birth weight babies. Mother and Children in this group showed greater responsiveness and sensitivity; and in cases of low levels of psychological resources they had home environments more supportive of children learning.
Nurses: Women reported greater intervals between births, less domestic violence, and enrolled the children less frequently in preschool. Children in this group and whose mothers had low levels of psychological resources had home environments that were supportive for learning, more advance language, superior executive functioning, and better behavioral adaptation.

References

Anderson, M. J. and P. Legendre (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation 62*(3), 271–303.

Bartlett, M. S. (1937, July). The statistical conception of mental factors. *British Journal of Psychology 28*(1), 97–104.

Beaton, A. E. (1978). Salvaging experiments: Interpreting least squares in non-random samples. In D. Hogben and D. W. Fife (Eds.), *Computer Science and Statistics: Tenth Annual Symposium on the Interface*, Washington, DC, pp. 137–145. U. S. Department of Commerce, National Bureau of Standards.

Bolck, A., M. Croon, and J. Hagenaars (2008). Estimating latent structure models with categorical variables: One-step versus three-step esimators. *Political Analysis 12*, 3–27.

Caldwell, B. M. and R. H. Bradley (1984). *HOME observation for measurement of the environment*. Little Rock, AR: University of Arkansas at Little Rock.

Carneiro, P., K. Hansen, and J. J. Heckman (2003, May). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review 44*(2), 361–422.

Croon, M. A. (2002). Using predicted latent scores in general latent structure models. In G. A. Marcoulides and I. Moustaki (Eds.), *Latent Variable and Latent Structure Models*, pp. 195–223. NJ: Lawrence Erlbaum Associates, Inc.

Cunha, F. and J. J. Heckman (2007, May). The technology of skill formation. *American Economic Review 97*(2), 31–47.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010, May). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

Eckenrode, J., M. Campa, D. Luckey, C. Henderson Jr., R. Cole, H. Kitzman, E. Anson, K. Sidora-Arcoleo, J. Powers, and D. Olds. (2010). Long-term Effects of Prenatal and Infancy Nurse Home Visitation on the Life Course of Youths 19-Year Follow-up of a Randomized Trial. *Arch Pediatr Adolesc Med. 1*, 9–15.

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

Frisch, R. (1938). Autonomy of economic relations: Statistical versus theoretical relations in economic macrodynamics. Paper given at League of Nations. Reprinted in D.F. Hendry and M.S. Morgan (1995), *The Foundations of Econometric Analysis*, Cambridge University Press.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica 12*(Supplement), iii–vi and 1–115.

Heckman, J., R. Pinto, and P. Savelyev (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review 103*(6), 768–806.

Kennedy, P. E. (1995, January). Randomization tests in econometrics. *Journal of Business and Economic Statistics 13*(1), 85–94.

Kitzman, H., D. Olds, R. Cole, C. Hanks, E. Anson, K. Arcoleo, D. Luckey, M. Knudtson, C. Henderson Jr., and J. Holmberg (2010). Enduring Effects of Prenatal and Infancy Home Visiting by Nurses on Children. *Arch Pediatr Adolesc Med. 5*, 412–418.

Kitzman, H., D. Olds, C. Henderson, C. Hanks, R. Cole, R. Tatelbaum, K. McConnochie, K. Sidora, D. Luckey, D. Shaver, et al. (1997). Effect of prenatal and infancy home visitation by nurses on pregnancy outcomes, childhood injuries, and repeated childbearing. *Jama 278*(8), 644–652.

Kitzman, H., D. Olds, K. Sidora, C. Henderson Jr., C. Hanks, R. Cole, D. Luckey, J. Bondy, K. Cole, and J. Glazner (2000). Enduring Effects of Nurse Home Visitation on Maternal Life Course: A Three-Year Follow-up of a Randomized Trial. *Jama 283*, 1983–1989.

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

Olds, D., C. J. Henderson, and T. R. (1994). Intellectual impairment in children of women who smoke cigarettes during pregnancy. *Pediatrics 93*(2), 221–227.

Olds, D., J. C. Henderson, R. Tatelbaum, and R. Chamberlin (1988). Improving the life-course development of socially disadvantaged mothers: a randomized trial of nurse home visitation. *American Journal of Public Health 78*, 1436–1445.

Olds, D., C. Henderson Jr, R. Chamberlin, and R. Tatelbaum (1986). Preventing Child Abuse and Neglect: A Randomized Trial of Nurse Home Visitation. *Pediatrics 78*, 65–78.

Olds, D., C. Henderson Jr., R. Cole, J. Eckenrode, H. Kitzman, D. Luckey, L. Pettitt, K. Sidora, P. Morris, and J. Powers (1998). Long-term Effects of Nurse Home Visitation on Children's Criminal and Antisocial Behavior: 15-Year Follow-up of a Randomized Controlled Trial. *JAMA*, 1238–1244.

Olds, D., C. Henderson Jr, and H. Kitzman (1994). Does Prenatal and Infancy Nurse Home Visitation Have Enduring Effects on Qualities of Parental Caregiving and Child Health at 25 to 50 Months of Life? *Pediatrics*, 93–89.

Olds, D., C. Henderson Jr, H. Kitzman, J. Powers, R. Cole, K. Sidora, P. Morris, L. Pettitt, and D. Luckey (1997). Long-term Effects of Home Visitation on Maternal Life Course and Child Abuse and Neglect: Fifteen-year Follow-up of a Randomized Trial. *JAMA 8*, 637–643.

Olds, D., C. Henderson Jr, R. Tatelbaum, and R. Chamberlin (1986). Improving the delivery of prenatal care and outcomes of pregnancy: A randomized trial of nurse home visitation. *Pediatrics 77*(1), 16.

Olds, D., H. Kitzman, R. Cole, J. Robinson, K. Sidora, D. Luckey, C. Henderson Jr, C. Hanks, J. Bondy, and J. Holmberg (2004). Effects of nurse home-visiting on maternal life course and child development: age 6 follow-up results of a randomized trial. *Pediatrics 114*(6), 1550.

Olds, D., H. Kitzman, C. Hanks, R. Cole, E. Anson, K. Sidora-Arcoleo, D. Luckey, J. C. Henderson, J. Holmberg, a. Tutt, A. Stevenson, and J. Bondy (2007). Effects of Nurse Home Visiting on Maternal and Child Functioning: Age Nine Follow-up of a Randomized Trial. *Pediatrics 120*, 832–845.

Olds, D. and J. Korfmacher (1998). Maternal psychological characteristics as influences on home visitation contact. *Journal of Community Psychology 26*(1), 23–36.

Olds, D., J. Robinson, R. O'brien, D. Luckey, L. Pettitt, C. Henderson Jr, R. Ng, K. Sheff, J. Korfmacher, S. Hiatt, et al. (2002). Home visiting by paraprofessionals and by nurses: A randomized, controlled trial. *Pediatrics 110*(3), 486.

Olds, D., J. Robinson, L. Pettitt, D. Luckey, J. Holmberg, R. Ng, K. Isacks, K. Sheff, and C. Henderson Jr (2004). Effects of home visits by paraprofessionals and by nurses: age 4 follow-up results of a randomized trial. *Pediatrics 114*(6), 1560.

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

Soares, J. and C. Wu (1983). Some restricted randomization rules in sequential designs. *Communications in Statistics-Theory and Methods 12*(17), 2017–2034.

Thomson, G. H. (1934, May). Hotelling's Method modified to give Spearman's g. *Journal of Educational Psychology 25*(5), 366–374.

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