The Nurse Family Partnership Program: a Reanalysis of the Memphis Randomized Controlled Trial

James Heckman, Maggie Holland, Kevin Makino, David Olds, Rodrigo Pinto, and Maria Rosales[1]

University of Chicago

This Draft: May 14, 2014

[1]James Heckman is the Henry Schultz Distinguished Service Professor of Economics at the University of Chicago, Professor of Science and Society, University College Dublin, Alfred Cowles Distinguished Visiting Professor, Cowles Foundation, Yale University, and Senior Fellow, American Bar Foundation. David Olds is NFP founder, Professor of Pediatrics, Psychiatry and Preventive Medicine University of Colorado. Rodrigo Pinto and Maria Rosales are graduate students at the University of Chicago. Margaret Holland and Kevin Makino are a senior researcher at the consortium that examines the Nurse-Family Partnership Program. We thank Terrance Oey and Willem van Vliet for superb research assistance. The research was supported in part by the American Bar Foundation, the Pritzker Children's Initiative, the Buffett Early Childhood Fund, NICHD 5R37HD065072, 1R01HD54702, an anonymous funder, the support of a European Research Council grant hosted by University College Dublin, DEVHEALTH 269874, and a grant from the Institute for New Economic Thinking (INET) to the Human Capital and Economic Opportunity Global Working Group (HCEO)an initiative of the Becker Friedman Institute for Research in Economics (BFI). The views expressed in this paper are those of the authors and not necessarily those of the funders or persons named here. Supplementary materials for this paper may be found at http://jenni.uchicago.edu/NFP/.

Abstract

As of 2000, home visiting programs have served over half million children in the U.S. Among those, the Nurse-Family Partnership (NFP) is the most cited home visitation program in the U.S. (Howard and Brooks-Gunn, 2009). It offers home visits conducted by nurses during pregnancy and early childhood for first-time disadvantaged mothers. In this paper, we analyze the treatment effects of the NFP randomized controlled trial performed in Memphis, TN (1990). We improved upon previous evaluations by performing a permutation-based inference that accounts for the characteristics of the NFP randomization protocol and is valid for small samples. We account for the arbitrary selection of statistically significant effects by performing a multiple-hypothesis testing based on the Stepdown procedure of Romano and Wolf (2005a). We also examine the underlying mechanisms generating the treatment effects. We estimate a mediation model that decomposes the NFP treatment effects into components associated with the intervention-enhanced parenting and early childhood skills. We find that NFP improved home investments, parenting attitudes and mental health for mothers of female and male infants at age 2. At age 6, NFP boosted cognitive skills for both genders and socio-emotional skills for females. We find evidence that these treatment effects are explained by the improvement of maternal traits and early-life family investments. At age 12, the treatment effects for males persisted in the form of boosting achievement test scores. These effects are mostly explained by enhanced cognitive skills at age 6.

Keywords: Cognitive Traits; Non-cognitive Traits; Health Traits; Factor Analysis; Human Capital; Human Development; Early Childhood Intervention; Social Experiment; Disadvantaged Mothers; Disadvantaged Children; Nurse-Family Partnership.

JEL Codes: I1; H5; C5.

1 Introduction

As of 2000, home visiting programs have helped over half-million disadvantaged children in the U.S.(Olds et al., 2000).[1] Among these programs, Nurse-Family Partnership (NFP) is the most cited home visiting program in the U.S. (Howard and Brooks-Gunn, 2009). It is considered a cornerstone for those that advocate the benefits of prenatal and early childhood interventions on health and socioeconomic outcomes. Today, NFP surrogates operate in 31 states across the U.S. and have provided service for more than $20,000$ families.

NFP offers prenatal and early childhood care to disadvantaged first-time unmarried mothers. The treatment consists of nurse visits from pregnancy until two years after birth. Its goal is to improve the odds of socio-economic long-term success for disadvantaged children. It operates by promoting healthy behaviors and fostering parenting skills.

In this paper, we analyze the NFP randomized controlled trial performed in Memphis, TN (1990). The NFP Memphis randomized trial data includes information on maternal behavior, home environment, parenting, children's cognitive and socio-emotional skills measured at ages 2 and 6, and middle childhood outcomes such as achievement and behavior problems, measured at age 12.

We contribute to the existing literature by performing a permutation-based inference that accounts for the characteristics of the NFP randomization protocol and is valid for small sample sizes. We address the problem of selective reporting of statistically significant outcomes (i.e., "cherry picking") by implementing a multiple-hypothesis testing that casts on the Stepdown procedure of Romano and Wolf (2005a). Our inference method follows Heckman et al. (2010), who show that a detailed study of the characteristics of the randomization protocol can improve the statistical power of traditional inference methods. Our analysis goes beyond causal inference. We study the underlying mechanisms generating the NFP effects. Specifically, we decompose statistically significant treatment effects into interpretable components associated with program-induced changes in children's early skills and parental investments. Our analysis is based on the concept of the technology of skill formation proposed by Cunha and Heckman (2007b), in which skills are multiple in nature and evolve over time according to a dynamic process of skill formation. This approach enables us to interpret early skills as building blocks of more complex skills at later periods. This is usually termed mediation analysis in the statistical literature (MacKinnon et al., 2006).

At age 2, we find statistically significant effects on home environment, parenting attitudes and maternal mental health for both boys and girls. At age 6, the program also improved cognitive skills for both boys and girls, while it enhanced early socio-emotional skills only for females. We show evidence that these treatment effects are likely to be explained by the improvement of maternal traits and early-life family investments.

[1] This number is expected to grow: the Patient Protection and Affordable Care Act guarantees $ 1.5 billion of new funding to home visiting programs over the next five years (Olds et al., 2000).

At age 12, the treatment effects for males persisted. We find that treated males outperformed controls in a range of achievement scores. We also find that 40% to 60% of male treatment effects at age 12 can be explained by enhanced cognitive skills at age 6.

Our results widen the current understanding of how early childhood interventions operate. Another example of mediation analysis applied to early childhood interventions is Heckman et al. (2012), who studied the Perry Preschool Project (PPP). NFP and PPP differ in several aspects. As mentioned, NFP is a home based intervention that targets prenatal health and parenting skills of disadvantaged mothers from pregnancy until 2 years after birth. In contrast, the Perry Preschool Project combined both center and home based components and targeted early cognitive and socio-emotional skills of children from 3 to 5 years old (Heckman et al., 2010). Heckman et al. (2012) focus on the study of adulthood outcomes instead of early achievement scores. They find that socio-emotional skills from ages 6-9 explained most of the treatment effects on adult anti-social behavior.

The remainder of this paper is organized as follows. Section 2 provides information on the experimental design and the background of the NFP Memphis randomized control trial, while Section 3 describes the data we use. Section 4 explains our inference approach. In Section 5, we describe our methodology to decompose the treatment effects. Sections 6 and 7 present the inference and mediation results. Section 8 concludes.

2 Experimental Design and Background

The NFP Memphis trial was conducted in 1990 and offered nurse home visits to disadvantaged first-time-pregnant mothers. According to Olds (2002) and Olds et al. (1997), the main goals of NFP are: (1) improvement in maternal and fetal health during pregnancy; (2) development of parenting skills; (3) planning of social and economic aspects of maternal life through counseling services.

Visits started during pregnancy and continued until the newborn reached the age of 2. During these visits, nurses encouraged mothers to adopt a healthy diet in addition to eliminating the use of tobacco, alcohol, and illegal drugs. They also taught mothers how to recognize pregnancy complications and how to achieve adequate prenatal care. After delivery, nurses promoted infant health care and good parenting skills. Mothers were instructed about how to interact with their children in order to foster emotional and cognitive development. NFP further assisted mothers in the process of establishing life goals related to work, education and future pregnancies.

The remaining of this section describes key aspects of the NFP Memphis intervention. We summarize main features of the NFP intervention in Table 1.

Eligibility

The NFP Memphis trial recruited pregnant women from June 1, 1990 to August 31, 1991 through the Memphis-Shelby County Health Department. Eligible mothers complied with 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 complied with two or more of the following socio-economic criteria: (1) unmarried; (2) less than 12 years of education; (3) unemployed.

Sample

The total eligible sample included $1,290$ invited participants. $1,138$ mothers effectively enrolled. The majority of the participants were African-American (92%), unmarried (97%), low income (95%), and under 18 years old (64%). Moreover, mothers who participated in the program were more likely to be younger, African-American and less likely to have completed high school compared to mothers who refused to participate (Kitzman et al., 1997).

Treatment

The mothers that agreed to participate in the trial were randomized into four different groups that differed by treatment:

- Group 1: round-trip transportation from their homes to their prenatal session appointments (sample size: 166);
- Group 2: developmental screening and referral services when their babies were aged 6, 12, and 24 months, in addition to the benefits of Group 1 (sample size: 514);
- Group 3: nurse visits during pregnancy, one visit when in the hospital and one visit at home after childbirth, in addition to the benefits of Group 2 (sample size: 230);
- Group 4: nurse home visits during pregnancy and until the child's second birthday, in addition to the benefits of Group 2 (sample size: 228);

Groups 3 and 4 are often denominated experimental groups while groups 1 and 2 play the role of comparison groups. There is no available data on the participants of Groups 1 and 3 after the child's birth, as those participants were not followed. Thus, our paper can only examine data of the participants that were originally assigned to Groups 2 or 4. Henceforth, we refer to Group 2 as the control group and to Group 4 as the treatment group. Table 2 presents a statistical description of selected baseline characteristics of the treatment and control groups by gender of the child. The table shows that pre-program characteristics are fairly balanced across treatment groups. Some exceptions were that treated mothers had lower income, more populated households, lower employment status of the household head, and higher grandmother support

compared to mothers in the comparison group. Also, control mothers of male children were slightly taller than their treatment counterparts.

Randomization protocol

The NFP randomization protocol can be summarized as a randomization performed within 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; and (5) 4 geographic regions of residence.[2] See Appendix A for a detailed description of the randomization protocol.

Attrition

Post-randomization dropout was rare during the program period: four women in both treatment and control groups refused further participation in the intervention. There were 487 live births in the control group (out of 514 pregnant mothers) and 214 live births in the treatment group (out of 228 pregnant mothers). There were 27 miscarriages in the control group and 14 in the treatment group. Table 3 shows the percentage of non-missing data (retention rates) by gender and age of follow-up interviews. When the children were 12 years old, 86% of families who had no fetal or child death were interviewed (85% in the control group and 88% in the treatment group).

3 Data Description

The available data on the NFP Memphis trial consists of several surveys that collected information at different stages of child development. Surveys were collected at the onset of the intervention, at birth, and at the following ages: 6 months, 1 year, 2 years, 6 years and 12 years. Available data extends over a variety of topics including biometrics, mother and child physical and mental health, risky behaviors, achievement scores, welfare use and socio-demographic variables. NFP also offers many psychological instruments that measure child cognition, socio-emotional skills and mother's parenting abilities. Table 4 summarizes available data for each period of child development.

The NFP Memphis trial provides numerous background variables surveyed at the onset of the intervention. Among those variables we can cite: race, mother's age, marital status, family income, employment, education, fertility, delinquency, maternal mental health, family support and maternal risk behavior. For a statistical description of these variables, see Table 2. At birth, available data consists of biological measures of health at birth such as placenta and birth weight.

Data on home environment and parenting was collected when the children were 6 months, 1 year and

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

Table 1: Nurse Family Partnership Intervention: Memphis Trial

Intervention goals	Healthy prenatal behaviors Parenting Skills Life-planning strategies
NFP Groups	Group 1 (N=166): Free transportation to and from appointments (not followed after birth) Group 2 (N=515): Developmental Screening at ages 6, 12, 24 months plus group 1 benefits Group 3 (N=230): Home visits by nurse during pregnancy plus group 2 benefits (not followed after birth) Group 4 2 (N=228) : Regular home visits during child's infancy plus all the group 3 benefits
Treatment Groups	In our analysis, Group 2 is termed control group and Group 4 is termed treatment group.
Nurse Visits	Weekly to biweekly visits during pregnancy Weekly visits for the first 6 weeks Bi-monthly visits during infancy
Dosage	Average number of visits during pregnancy: 7 visits Average number of visits during infancy (0-24 months): 26 visits
Target population	Low Income First-time mothers During Pregnancy to two years of age
Eligibility criteria	**Biological Criteria:** 1. Less than 29 weeks of pregnancy 2. No previous live births 3. No specific chronic illness affecting fetus **Socio-economic Criteria:** at least two of the following indicators 1. Unmarried 2. Less than 12 years of education 3. Unemployed
Sample sociodemographic characteristics	Race: 92% African American Marital Status: 97% Unmarried Income: 85% Low Income Age: 64% Under 18

Notes: Basic information on the NFP Memphis Trial. See (Kitzman et al., 1997) for a detailed description.

2 years olds. The Home Observation for Measurement of the Environment (HOME) inventory measures home environment and family investments. It was surveyed at ages 1 and 2 and provides information on the resources and time allocated to cognitive and emotional stimuli. Regarding parenting skills, NFP administered the Adult-Adolescent Parenting Inventory (Bavolek), surveyed at 6 months, 1 and 2 years old. Bavolek measures mother's parenting and child rearing attitudes towards non-abusive and neglecting parenting.

Data on socio-emotional skills and hospital records is available at age 2. The Child-Behavior Checklist (CBCL) is an inventory of child behavioral and emotional problems as reported by the mother. It was developed by T.M. Achenbach and measures affective and anxiety disorders, somatic complaints, attention deficit/hyperactivity behaviors, defiant attitude, and general conduct problems (Association and on DSM-IV., 2000). Additionally, maternal psychological characteristics such as self-esteem, mastery and anxiety were measured using the Rosenberg Scale, the Pearlin Scale, and the Rand Mental Health Inventory respectively.

Table 2: Descriptive Statistics of NFP Baseline Characteristics

	Female Sample					Male Sample				
	Control Mean	Control Std.Dev.	Treated Mean	Treated Std.Dev.	p-val	Control Mean	Control Std.Dev.	Treated Mean	Treated Std.Dev.	p-val
Background Characteristics										
Maternal Race (White)	0.086	0.281	0.101	0.303	0.656	0.076	0.265	0.112	0.317	0.300
Marital Status (Married)	0.012	0.110	0.009	0.096	0.791	0.025	0.157	0.019	0.136	0.696
Maternal Age	18.318	3.290	18.257	3.531	0.877	17.950	3.130	17.963	2.962	0.970
Years of Education	10.384	1.833	10.147	2.040	0.300	10.252	1.919	10.112	1.997	0.543
Mother in School	0.541	0.499	0.596	0.493	0.333	0.630	0.484	0.533	0.501	**0.093**
Head of Household is Employed	0.607	0.490	0.495	0.502	**0.054**	0.517	0.501	0.519	0.502	0.972
% of Census Tract Below Poverty	32.532	20.225	36.289	21.820	0.128	36.047	22.099	34.249	18.324	0.431
Household Density	0.944	0.490	1.053	0.652	0.123	0.923	0.490	0.983	0.446	0.264
Total Household Income (Past 6 Months)										
Less than $3000	0.278	0.449	0.358	0.482	0.141	0.286	0.453	0.355	0.481	0.208
$3000 - $6999	0.237	0.426	0.220	0.416	0.732	0.248	0.433	0.224	0.419	0.632
$7000 - $10999	0.216	0.413	0.229	0.422	0.788	0.235	0.425	0.178	0.384	0.213
Greater than $11000	0.180	0.385	0.083	0.277	**0.008**	0.134	0.342	0.178	0.384	0.320
Income, No Response	0.090	0.286	0.110	0.314	0.565	0.097	0.296	0.065	0.248	0.311
Region of Residence										
Inner City	0.282	0.451	0.284	0.453	0.958	0.298	0.458	0.280	0.451	0.734
Bisson	0.171	0.378	0.229	0.422	0.220	0.210	0.408	0.206	0.406	0.925
Cawthon	0.229	0.421	0.174	0.381	0.233	0.185	0.389	0.224	0.419	0.410
Hollywood	0.318	0.467	0.312	0.465	0.905	0.307	0.462	0.290	0.456	0.750
Maternal Mental Health										
Maternal IQ (Shipley)	96.429	10.295	96.505	10.459	0.949	96.315	10.194	96.785	10.565	0.700
Maternal Bavolek Score	99.770	7.633	100.932	8.762	0.233	99.571	7.776	100.531	8.439	0.318
Maternal Mental Health	99.732	10.116	99.120	10.532	0.610	100.657	10.033	99.359	10.662	0.288
Self-Efficacy	100.705	9.836	100.553	9.283	0.889	99.437	10.064	98.892	11.045	0.664
Maternal Mastery	100.132	10.202	99.095	10.106	0.375	100.315	10.201	100.141	9.746	0.880
Maternal Psychological Resources	100.150	9.903	99.505	10.433	0.586	100.148	10.194	99.565	11.126	0.644
Maternal Health Characteristics										
Maternal Height (Cmt)	164.295	7.465	164.485	6.462	0.809	164.886	6.977	163.603	6.662	0.109
Pre-Pregnancy Weight (kgr)	62.826	14.130	61.217	12.027	0.273	61.307	15.571	63.392	14.674	0.233
Gestational Age	16.335	5.717	16.312	5.472	0.972	16.752	5.821	17.065	5.827	0.644
Maternal Social Support										
Grandmother Social Support	99.037	10.742	101.220	9.599	**0.059**	100.991	8.303	101.858	7.911	0.355
Husband/Boyfriend Social Support	99.806	10.200	99.962	9.772	0.891	100.421	9.976	101.068	10.248	0.585
Maternal Risky Behaviors										
Alcohol Consumption (Past 2 wks)	0.033	0.178	0.064	0.246	0.232	0.046	0.210	0.028	0.166	0.389
Smoking (Past 3 days)	0.090	0.287	0.128	0.336	0.303	0.101	0.302	0.093	0.292	0.830
Used Marijuana (Past 2 wks)	0.025	0.271	0.018	0.192	0.805	0.038	0.323	0.112	1.160	0.516
Used Cocaine (Past 2 wks)	0.000	0.000	0.000	0.000	–	0.013	0.194	0.000	0.000	0.318
Sexually Transmitted Diseases	0.335	0.473	0.349	0.479	0.801	0.324	0.469	0.402	0.493	0.167

Notes: This table presents the statistical description of selected pre-program variables at baseline. The first column of the table gives the variable description. The 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. Maternal social support corresponds to standardized indexes. Additional baseline tables using samples at years 6 and 12 can be found in Section D of the Appendix.

A battery of psychological instruments measuring child cognition and socio-emotional skills were administered at age 6. Three instruments were used to measure cognition: (1) the Kaufman Assessment Battery for Children (K-ABC), (2) the Peabody Picture Vocabulary Test (PPVT) and (3) the coding sub-test from the Wechsler Intelligence Scale for Children (WISC-III). K-ABC evaluates sequential and simultaneous processing to solve problems, language or literacy achievement, and math knowledge. PPVT tests receptive vocabulary and WISC-III examine short-term memory and visual perception. Socio-emotional skills were

Table 3: Retention Rates by Gender

	Female Sample			Female Sample		
	All Females	Control Group	Treatment Group	All Males	Control Group	Treatment Group
Time	Groups 2 and 4	Group 2	Group 4	Groups 2 and 4	Group 2	Group 4
Month 6	95.48	96.73	92.66	96.24	97.07	94.39
Month 12	98.02	97.55	99.08	97.69	97.91	97.2
Year 2	96.05	95.92	96.33	98.26	98.74	97.2
Year 4.5	90.96	89.8	93.58	94.19	94.09	94.39
Year 6	91.24	91.43	90.83	95.03	95.32	94.39
Year 9	89.55	90.2	88.07	93.57	94.47	91.59
Year 12	86.16	86.94	84.4	90.32	88.94	93.4

Notes: The table presents sample attrition over time. The first column of the table gives the time description. There are 2 blocks, one for females and one for males. Each block has 3 columns for the sub-samples consisting of Groups 2 and 4, Group 2 only and Group 4. Each column provides the percentages of non-missing data, that is, the percentage of complete mother interviews among children that did not die at birth.

surveyed through two instruments: the Child-Behavior Checklist (CBCL) and the McArthur Story Stem Battery (MSSB). The MSSB collects children's responses to eight story beginnings (stems), which are videotaped and coded to represent content themes. Its aim is to test observable affective expressions and coherence in completing stories (Olds et al., 2004). The codes are averaged across stories and mapped into the following four dimensions: 1) deregulated aggression, which combines representations of aggression, personal injury, dishonesty, danger, destruction, inappropriate child power, and negative parenting; 2) warmth and empathy, which encompasses parental warmth and representations of affection, affiliation, reparation, and guilt; 3) emotional integration, which captures the ability to maintain a coherent story; and 4) performance anxiety, which includes the unwillingness to verbalize, unresponsiveness, and anxiety behaviors. Additional descriptions of psychological instruments can be found in Appendix B.

Outcomes at age 12 consist of achievement test scores obtained through school records, and child mental health. Achievement scores include the Tennessee Comprehensive Assessment Program (TCAP), GPA records from grades 1 to 5 and the Peabody Individual Achievement Test (PIAT) collected at age 12. In addition, there is information on child's internalizing and externalizing behaviors from the CBCL. We also consider data on welfare use and government expenditures obtained from the Tennessee administrative records.

4 Inference

In this paper, we analyze the treatment effects of the NFP Memphis trial using inference techniques that account for the features of the randomization protocol and correct for multiple-hypothesis testing. Our

Table 4: NFP Classification of Outcome/Meditors

Instrument	Content		Time of Survey (Child's Age)					
	Subject	Classification	0	0.5	1	2	6	12
Birth Biometrics	Child/Mother	Health	✓					
Bavolek	Parents	Environment/Parenting		✓	✓	✓		
HOME	Family	Environment/Parenting			✓	✓		
Mathernal Mental Health	Mother	Mental Health				✓		
K-ABC	Child	Cognition					✓	
PPVT	Child	Cognition					✓	
WISC-III	Child	Cognition					✓	
CBCL	Child	Socio-emotional				✓	✓	
MSSB	Child	Socio-emotional					✓	
Hospital Visits	Child	Health	✓	✓	✓	✓		
Health Injuries	Child	Health						✓
PIAT	Child	Achievement Scores						✓
Schooling Grades	Child	Achievement Scores						✓
CBCL/Youth Self-Report	Child	Mental Health						✓
Welfare Use	Family	Governmental Aid						✓

Notes: This table presents a summary of the data available for the analysis of treatment effects in NFP mediators and outcomes. The first column presents the name of the outcome or instrument according to Section 3. The remaining columns point out the time of survey according to the age of the child in participating families. Additional information on the description of the available instruments of the NFP Data can be seen at Appendix B.

methodology relies on permutation-based tests, which do not rely on asymptotic properties of a test statistics and are valid for small sample sizes. Permutation tests are often termed *distribution-free* because they do not require parametric assumptions about the data generating process. In particular, they are valid when the data distribution is non-normal or skewed.

Our inference method relies on an exchangeability property generated by the randomization protocol. Specifically, under the null hypothesis of no treatment effect, randomization guarantees that treatment assignments among program participants are exchangeable within blocks defined by the baseline characteristics used in the stratification of the sample at intake (Heckman et al., 2010) By blocks we mean the set of participants that share the same value of the baseline variables used in the randomization protocol. We generate the exact distribution of a test statistic conditional on data by permuting the treatment status of participants that belong to the same block. It is a form of nonparametric conditioning. We also condition on background variables other than the ones used in the randomization protocol by casting on the Freedman and Lane (1983) procedure, which is a permutation test that evokes linearity. Finally, the presence of multiple outcomes leads to the danger of arbitrarily selecting statistically significant outcomes that may occur just by chance. This is often termed "cherry picking". We correct for this potential source of inference bias

by performing multiple-hypothesis testing that implements the Stepdown algorithm of Romano and Wolf (2005b).

The remainder of this section provides a broad discussion of our method. Section 4.1 explains the mechanics of our permutation-based inference. Section 4.2 discusses our solution for conditioning on pre-program variables other than the ones used in the randomization protocol. Section 4.3 explains our multiple-hypothesis testing approach. We also present a formal and detailed description of our inference method in Appendix C.

4.1 Exchangeability and the Randomization Protocol

We follow the standard model of program evaluation that describes the outcome of participant i that belongs to sample set $\mathcal{I}$ by

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

where $D_i \in \{0, 1\}$ is an indicator of treatment assignment: $D_i = 1$ if participant i is randomized into the treatment group and $D_i = 0$ otherwise. The counterfactual outcomes[3] when the treatment assignment of individual i is *fixed* to control and treatment statuses are respectively given by $(Y_{i,0}, Y_{i,1})$.[4]

Randomized experiments solve the fundamental problem of *selection bias* due to treatment choice by the random assignment of treatment status. Heckman et al. (2010) explain that this fact is translated to the following assumption:

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

where $D = (D_i; i \in \mathcal{I})$ stands for the vector of treatment statuses, $X = (X_i; i \in \mathcal{I})$ for the vector of variables used in randomization protocol and $Y = (Y_i; i \in \mathcal{I})$ and $Y_d = (Y_{i,d}; i \in \mathcal{I})$ such that $d \in \{0, 1\}$ for the vectors of observed and counterfactual outcomes. In the case of NFP, variables X are maternal age and race, gestational age at enrollment, employment status of the head of the household, and geographic region.

Under Assumption **A-1**, we can assess the causal effect of the NFP intervention, i.e. $\mathbf{E}(Y_{i,1} - Y_{i,0}|X)$, by the conditional expectation $\mathbf{E}(Y|D = 1, X) - \mathbf{E}(Y|D = 0, X)$. Our goal is to test the null hypothesis of no treatment effect. This is equivalent to state that the conditional counterfactual outcome vectors have the same distribution:

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

Heckman et al. (2010) and Pinto (2012) show that it is possible to restate Hypothesis **H-1** as:

[3] Counterfactual outcomes are a common notion in the program evaluation literature and denote the actual outcomes and the outcomes that whould have happen in the absense of the intervention

[4] See Heckman and Pinto (2014a) for a discussion on causality and the concept of fixing.

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

We test Hypothesis **H-1′** using an exchangeability property of the vector of treatment statuses D. Specifically, participants that have the same values of X are *indistinguishable* under the randomization protocol. As a consequence, the distribution of D conditioned on X remains the same when elements in D are permuted within blocks formed by the values X takes. Moreover, if the no-treatment hypothesis **H-1′** holds, then the vector of treatment assignments must be independent of the outcome vector. Thereby, the joint distribution of (Y, D) must also be invariant with respect to permutations of D within blocks defined by baseline characteristics X. Notationally, we write:

$$(Y, D) \stackrel{d}{=} (Y, gD) \; \forall \; g \in \mathscr{G}_X. \tag{2}$$

where g is an action that permutes the elements of vector D and $\mathscr{G}_X$ is the set of all permutations that only permute within blocks formed by baseline characteristics X. Lehmann and Romano (2005, Chapter 9) terms Equation (2) as the Randomization Hypothesis.

We can use the Randomization Hypothesis (2) to conduct inference on the no-treatment hypothesis. For instance, let $T(Y, D)$ be a test statistic whose larger values provide evidence against **H-1′**. Let a standard statistical test rejects the null hypothesis of no treatment effect when $T(Y, D)$ is bigger than a critical value c. If our goal is to control for a Type-I error at significance level α; then the following equation must hold:

$$\begin{aligned} &\Pr(\text{Reject hypothesis } \textbf{H-1} \mid \text{hypothesis } \textbf{H-1} \text{ is true}) \\ &\qquad = \Pr(T(Y, D) \geq c \mid \text{Hypothesis } \textbf{H-1} \text{ is true}) \leq \alpha. \end{aligned} \tag{3}$$

In order to implement the test (3) above, we must compute the critical value c such that the probability of rejecting a null hypothesis of no treatment effect is less than α whenever Hypothesis **H-1** is true. But under Randomization Hypothesis (2), $T(Y, D)$ is uniformly distributed across the values of $T(Y, gD); g \in \mathscr{G}_X$.[5] Thus, we can compute the critical value c by taking the α quantile of the set $\{T(Y, gD) : \; g \in \mathscr{G}_X\}$. By permuting treatment statuses within blocks generated by X, our inference method is non-parametrically conditional on the values X takes.

In practice, permutation tests compare a test statistic computed on the original (not-permuted) data with a distribution of test statistics computed on re-samplings of that data through permutations. The measure of evidence against the Randomization Hypothesis, the p-value, is computed as the fraction of re-sampled data which generates a test statistic greater than the one obtained from the original data.

[5] See Lehmann and Romano (2005), Theorem 15.2.2.

4.2 Conditioning and Linearity

A typical problem in randomized trials is sampling variation, where pre-program variables differ across treatment groups by chance. One can increase the power of a statistical inference by conditioning on those pre-program variables.

We use Z to represent the pre-program variables not used in the randomization protocol that we ought to control due to imbalances between treatment and control groups.[6] Variables Z precede the treatment intervention and therefore $Z \perp\!\!\!\perp D \mid X$ holds due to randomization. But $Y \perp\!\!\!\perp D \mid X$ holds under the hypothesis of no-treatment effect. These two relations imply that $Y \perp\!\!\!\perp D \mid (X, Z)$. Like in Section 4.1, we can use this relation, $Y \perp\!\!\!\perp D \mid (X, Z)$, to generate a permutation test that consider the block formed by values of covariates X and Z. This way we can generate an inference method that non-parametrically conditions on variables X and Z.

Non-parametric conditioning through block permutation comes at a cost. As the set of conditioning variables increases, we reduce the number of eligible permutations. The reduction of the possible orbits of permutations can prohibit the implementation of a permutation-based test. We solve this problem by evoking linearity. That is, we avoid lessening the number of permutations by conditioning variables through a linear regression instead of a non-parametric block permutation. Anderson and Legendre (1999) investigate a range of permutation methods for linear models. They found that the Freedman and Lane (1983) method 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 a valid exchangeability property (i.e. we use permutations in $\mathscr{G}_X$). We linearly condition on additional pre-program variables Z not used in the randomization protocol. According to the Freedman and Lane (1983) method, our approach can be summarized by the following steps: (1) compute the residuals $Y - Z\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}} = (Z'Z)^{-1}Z'Y$; (2) permute these residuals according to permutations $g \in \mathscr{G}_X$; (3) add these permuted residuals to $Z\hat{\boldsymbol{\beta}}$, call it $\tilde{Y}$; (4) regress $\tilde{Y}$ on Z and the treatment statuses D, that is, $\tilde{Y}_i = \kappa + \delta D_i + \beta Z_i + \epsilon_i$; (5) store the t-statistic associated with D in the regression of item 4 as test statistic. A distribution of the test statistics is obtained by reiterating items (2)–(5). The p-value is computed as the fraction of re-sampled statistics that are greater than the non-permuted one. See Appendix C for a detailed discussion of the Freedman and Lane (1983) method.

In the case of NFP, the baseline variables used in the randomization protocol are: maternal age and race, gestational age at enrollment, employment status of the head of the household, and geographic region. These are the variables used for block permutation. Also, we adjust for imbalances in: maternal height, household

[6] These variables are: maternal height, household income, grandmother support, maternal parenting attitudes and mother currently in school.

income, grandmother support, maternal parenting attitudes and mother's currently in school. These are the covariates used as inputs in the Freedman and Lane (1983) method.

4.3 Multiple-Hypothesis Testing

A large number of outcomes allows for the selective reporting of statistically significant treatment effects that may occur by chance (i.e. "cherry picking"). Suppose that we use a single-hypothesis test statistic that rejects the true null hypothesis at significance level α. Therefore, the probability of rejecting at least one hypothesis out of K hypotheses is $1-(1-\alpha)^K$. Noticeably, as the number of outcomes increases, the rejection probability of a true null hypothesis departs from α and converges to 1. Put differently, Type-I error becomes more likely.

We correct for the possibility of arbitrary selection of statistically significant outcomes by correcting for Family-wise Error Rate (FWER), that is, the probability of rejecting at least one true hypothesis out of a set of multiple hypotheses. To do so, we use the stepdown method described in Romano and Wolf (2005a). The method strongly controls for FWER, i.e., it provides a valid multiple-hypothesis inference among a set of null hypotheses regardless of how many hypotheses are false.

The stepdown method targets the most significant test statistic out of the set of statistics associated with the hypotheses. This statistic is the one that most likely contributes to the significance of the joint null hypothesis. It then decides if the statistic is significant or not, accounting for the multiplicity of hypotheses. In the second step, stepdown disregards the most individually-significant hypothesis targeted in the first step and focuses on the remaining hypotheses to be tested. The method then iterates. In each step, we compute a stepdown p-value that adjusts the p-value of the most significant single hypothesis being targeted for the multiplicity of hypotheses. For a formal description of the stepdown method, see Appendix C.

5 Decomposing NFP Treatment Effects

In this section, we examine the mechanisms through which the NFP intervention generates treatment effects. Our goal is to decompose the treatment effects into interpretable components associated with intervention-induced changes of early skills. This method is usually termed mediation analysis in the statistical literature.

Our approach is based on a system of structural equations that models the process of child development. This framework is based on the concept of a technology of skill formation of Cunha and Heckman (2007b). In it, skills have multiple dimensions and encompass cognitive, socio-emotional and health related skills. They evolve through time as a function of previous skills, current investments and family characteristics. In this framework, early skills are building blocks for more advanced skills at later periods of development. Also,

family investments can respond to previous skills, and later skills foster adulthood outcomes. As examples of skills, we can cite cognition, socio-emotional abilities and health status.

Appendix F develops the theoretical background of the technology of skill formation model. This framework allows us to write the counterfactual outcome $Y_{i,d}$ of a participant i when the treatment D_i is fixed at $d \in \{0,1\}$ as a function of family background variables X_i, a vector of post-treatment counterfactual skills $\boldsymbol{\theta}_{i,d}$ and an exogenous error term ζ_i that is independent of X_i and $\boldsymbol{\theta}_{i,d}$:

$$Y_{i,d} = f(\boldsymbol{\theta}_{i,d}, X_i, \zeta_i). \tag{4}$$

We follow the approach of Heckman et al. (2013) which linearizes Equation (4) by approximating it through a Maclaurin expansion:

$$Y_{i,d} = \kappa + \boldsymbol{\alpha}_d \boldsymbol{\theta}_{i,d} + \boldsymbol{\beta}_d X_i + \epsilon_{i,d}, \;\; d \in \{0,1\}. \tag{5}$$

where $\epsilon_{i,d}$ accounts for a zero-mean error approximation.

The impact of skills and pre-program variables on the outcome is mapped through parameters $\boldsymbol{\alpha}_d$ and $\boldsymbol{\beta}_d$. If treatment effects are generated by changes in skills and not by the map between these skills and the outcomes of interest, then we have that

$$\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_0 \text{ and } \boldsymbol{\beta}_1 = \boldsymbol{\beta}_0. \tag{6}$$

Restrictions (6) simplify the interpretation of the channels through which the NFP intervention affect the outcome. [7] We use the method described in Heckman et al. (2013) to test Restrictions (6) in Tables H.14 and H.15 of Appendix H. As we do not reject those restrictions, we simplify Equation (5) by the following model:[8]

$$Y_{i,d} = \kappa + \boldsymbol{\alpha} \boldsymbol{\theta}_{i,d} + \boldsymbol{\beta} X_i + \epsilon_{i,d}, \;\; d \in \{0,1\}. \tag{7}$$

Variables $\boldsymbol{\theta}_{i,d}$ represent the set of all counterfactual skills that can affect outcome Y_i. These comprise the skills that can be measured using the available data from NFP surveys, as well as skills that can't be assessed due to the absense of related measures in the data. We follow Heckman et al. (2013) that uses $\mathcal{J}$ to index all skills and $\mathcal{J}_p \subseteq \mathcal{J}$ to index measured skills. Under this notation, the vector of counterfactual skills is given

[7] These are called structural invariance assumptions by Heckman et al. (2013).

[8] We can assume that $\epsilon_{i,1} \overset{d}{=} \epsilon_{i,0}$ as any discrepancies in the error term between treatment and control groups can be attributed to differences in skills.

by $\boldsymbol{\theta}_{i,d} = (\theta^j_{i,d} : j \in \mathcal{J}); d \in \{0,1\}$ and $\boldsymbol{\alpha} = (\alpha^j : j \in \mathcal{J})$. We can then rewrite Equation (7) as:

$$\begin{aligned} Y_{i,d} &= \kappa + \underbrace{\sum_{j \in \mathcal{J}_p} \alpha^j \theta^j_{i,d}}_{\text{measured skills}} + \underbrace{\sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha^j \theta^j_{i,d}}_{\text{unmeasured skills}} + \boldsymbol{\beta} X_i + \epsilon_{i,d} \\ &= \tau_d + \sum_{j \in \mathcal{J}_p} \alpha^j \theta^j_{i,d} + \boldsymbol{\beta} X_i + \tilde{\epsilon}_{i,d} \end{aligned} \tag{8}$$

where $\tau_d = \kappa + \sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha^j \, \mathrm{E}(\theta^j_{i,d}); d \in \{0,1\}$, and $\tilde{\epsilon}_{i,d} = \epsilon_{i,d} + \sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha^j (\theta^j_{i,d} - \mathrm{E}(\theta^j_{i,d}))$. According to Equation 1, the observed outcome for participant i is given by $Y_i = Y_{i,1} D_i + Y_{i,0}(1 - D_i)$. In the same fashion, we can express skills $\theta^j_i; j \in \mathcal{J}$ as:

$$\theta^j_i = \theta^j_{i,1} D_i + \theta^j_{i,0}(1 - D_i); \; j \in \mathcal{J}. \tag{9}$$

Substituting Equations (8)–(9) into (1) we obtain:

$$Y_i = \tau_0 + \tau D_i + \sum_{j \in \mathcal{J}_p} \alpha^j \theta^j_i + \boldsymbol{\beta} X_i + \tilde{\epsilon}_i \tag{10}$$

where $\tilde{\epsilon}_i = \tilde{\epsilon}_{i,1} D_i + \tilde{\epsilon}_{i,0}(1 - D_i)$ and $\tau = \tau_1 - \tau_0$. Parameter τ captures the NFP effect on outcome Y that is due to the average change of unmeasured skills. This is often called unexplained or residual effect. Parameters $\{\alpha^j\}_{j \in \mathcal{J}_p}$, on the other hand, map how program-induced changes in measured skills affect outcome Y. This is usually termed mediated or explained effects.

The unbiased estimation of the linear coefficients $\{\alpha^j\}_{j \in \mathcal{J}_p}$ in (7) requires independence of measured skills $\theta^j_i; j \in \mathcal{J}_p$ and the error term $\tilde{\epsilon}_i$. Heckman et al. (2013) explain that this requirement is achieved through the independence of measured and unmeasured skills, that is:

$$\{\theta_{i,d}\}_{j \in \mathcal{J}_p} \perp\!\!\!\perp \{\theta_{i,d}\}_{j \in \mathcal{J} \setminus \mathcal{J}_p} \mid (X_i, D_i); d \in \{0,1\} \tag{11}$$

The violation of (11) leads to the usual omitted-variable bias.[9] Appendix H presents robustness tests on the independence relation (11). We perform these tests using unused measures available in the data, in the same fashion as Heckman et al. (2013). We decompose the outcome conditional-mean treatment effects of the NFP program into changes associated to measured and unmeasured skills by:

$$\underbrace{\mathrm{E}(Y|D=1,X) - \mathrm{E}(Y|D=0,X)}_{\text{conditional treatment effect}} = \underbrace{(\tau_1 - \tau_0)}_{\text{role of unmeasured skills}} - \sum_{j \in \mathcal{J}_p} \underbrace{\alpha^j \, \mathrm{E}(\theta^j_1 - \theta^j_0)}_{\text{role of measured skill } j}. \tag{12}$$

[9] The linearity assumption enables us to require a weaker condition, mean independence.

Equation (12) implies that larger conditional treatment effects are a consequence of intermediate skills that are both significantly influenced by the program ($E(\theta_1^j - \theta_0^j) \neq 0$) and highly correlated with outcomes, that is, α^j is different from zero.

Our mediation analysis is motivated by a technology of skill formation that allows for measured and unmeasured skills. This approach differs from the traditional mediation literature, which does not examine unmeasured skills. Nevertheless we can map our assumptions into the ones used in the standard mediation literature. For instance, a standard counterfactual outcome equation on this literature is:[10]

$$Y_{i,d} = f(d, \boldsymbol{\theta}_{i,d}^p, X_i, \zeta_i); d \in \{0,1\}, \tag{13}$$

where $\boldsymbol{\theta}_{i,d}^p = \{\theta_{i,d}\}_{j \in \mathcal{J}_p}$ represents the counterfactual vector of measured skills. Equation (13) is a simplification of Equation (4). It allows for the decomposition of outcome treatment effects into the mediated effect of enhancement of skills, often called "indirect effect", and the "direct" or "residual" effects that operated through the treatment indicator d. Specifically, let $Y_{i,d}(\boldsymbol{\theta})$ denotes the counterfactual outcome Y for participant i, where treatment status is fixed at d and skills fixed at $\boldsymbol{\theta}$. In this notation, the direct effects are given by $Y_{i,1}(\boldsymbol{\theta}_{i,d}^p) - Y_{i,0}(\boldsymbol{\theta}_{i,d}^p); d \in (0,1)$ and the indirect effects given by $Y_{i,d}(\boldsymbol{\theta}_{i,1}^p) - Y_{i,d}(\boldsymbol{\theta}_{i,0}^p); d \in \{0,1\}$. The direct effects capture the impact of the treatment itself while the intermediate skills are kept constant. The indirect effect accounts for the sole impact of the intermediate skills. We can map these quantities into our decomposition (12) as:

$$\begin{aligned}
\mathrm{E}(Y|D=1) - \mathrm{E}(Y|D=0) &= \mathrm{E}(Y_{i,1}(\boldsymbol{\theta}_{i,1}^p) - Y_{i,0}(\boldsymbol{\theta}_{i,0}^p)) \\
&= \underbrace{\mathrm{E}\left(Y_{i,1}(\boldsymbol{\theta}_{i,d}^p) - Y_{i,0}(\boldsymbol{\theta}_{i,d}^p)\right)}_{\text{role of unmeasured skills}} + \underbrace{\mathrm{E}\left(Y_{i,d}(\boldsymbol{\theta}_{i,1}^p) - Y_{i,d}(\boldsymbol{\theta}_{i,0}^p)\right)}_{\text{role of measured skills}}; d \in \{0,1\}
\end{aligned}$$

Imai et al. (2010) summarizes that the standard assumptions in mediation analysis are given by the so-called Sequential Ignorability Assumption:

$$\left(Y_{i,d}(\boldsymbol{\theta}^p), \boldsymbol{\theta}_{i,d}^p\right) \perp\!\!\!\perp D_i | X_i, \text{ and} \tag{14}$$

$$Y_{i,d}(\boldsymbol{\theta}^p) \perp\!\!\!\perp \boldsymbol{\theta}_{i,d}^p | \left(D_i = d, X_i\right). \tag{15}$$

Equation (14) states that counterfactual outcomes and counterfactual skills are independent of treatment statuses. In NFP, this relation comes as a consequence of the randomization of treatment assignments. Equation (15) assumes that measured skills are independent of any other variable that affects outcomes. In

[10] See (Imai et al., 2010; Jo, 2008) for examples of this literature.

NFP, this can only occur if the independence relation (11) holds. For a detailed discussion on mediation, see Heckman and Pinto (2014b). The next section describes our estimation methodology.

5.1 Empirical Strategy

We estimate a factor model in which measured skills $\boldsymbol{\theta}_i^p$ are evaluated as latent variables called factors and measured by a range of psychological instruments available in the NFP data. We then use this factor model to forecast skills that are then used to explain final outcomes. Factor models are particularly useful to summarize a large set of response variables, i.e. psychological instruments, into a small number of latent variables, i.e. skills. Unobserved skills are estimated through a weighted average of relevant observed measures. Factor models outperform other arbitrary indexes by reducing the measurement error associated with the factor estimation (Gorsuch, 1983).

Our empirical estimation is based on the three-step procedure of Heckman et al. (2013). The first step consists of the estimation of a linear measurement system in which latent skills are a function of measures, that is, observed item-level data on psychological instruments. In the second step, we use the parameters estimated in the previous step to forecast skills. These forecasts are termed factor scores and are obtained by the Bartlett (1937) method. In the third step, we explain later outcomes in terms of previous skills. Specifically, we use the computed factor scores as covariates in the linear regression represented by Equation (10). See Appendix G for a detailed explanation of the three-step procedure used in our estimation.

A benefit of a three-step procedure is to clearly distinguish the effect of the intervention on skills and the effect of the experimentally induced changes in these skills on later outcomes. Our standard errors and p-values are computed using the bootstrap method.

The NFP psychological instruments target well-specified traits of personality (see Appendix B for a description). We incorporate this fact into our methodology by adopting a measurement system that is based on dedicated measures. By this we mean that each observed measure is linked to a unique factor.[11] We also allow the latent factors to correlate. This approach is termed confirmatory factor analysis in the factor model terminology (Gorsuch, 1983). Likewise Heckman et al. (2013), we correct for the measurement error that arise from using an estimated factor, i.e. factor scores, instead of the true factor.

We conduct two sets of analyses. We first examine if child skill improvements at age 6 were mediated by variables surveyed at 2 years old. Next, we examine if child outcomes at age 12 were mediated by a set of skills at age 6. At age 2, we investigate five classes of mediators: (1) non-abusive parenting attitudes measured by the Bavolek Inventory; (2) home investments measured by the HOME inventory; (3) maternal anxiety measured by the maternal anxiety subscale from the Rand Mental Health Inventory; (4) maternal

[11] Heckman et al. (2013) show that the existence of at least three measures for each latent skill guarantees identification.

self-esteem measured by the Rosenberg Scale; and (5) maternal mastery measured by the Pearlin Scale. At age 6, we examine three broad psychological instruments as mediators: (1) child's cognitive skills measured by the K-ABC mental processing composite; (2) child's socio-emotional skills measured by the Child Behavior Checklist scales of attention and conduct problems; (3) child's warmth, empathy and aggression measured by the MacArthur Story Stem Battery.

6 Inference Results

We investigate the statistical significance of the NFP treatment effects using the methodology described in Section 4. We find statistically significant treatment effects that survive the multiple-hypothesis testing in a range of outcomes.

Table 5 presents a synopsis of the overall significance of the NFP treatment effects. The table shows the percentage of outcomes whose treatment effects are statistically significant at different significance levels by gender. These percentages are larger than what we would expect by chance.

Table 5: Percentage of Test Statistics Exceeding Various Significance Levels

Significance Level	Female Sample	Male Sample
Percentage of p-values smaller than 1%	7.6%	2.5%
Percentage of p-values smaller than 5%	22.3%	17.2%
Percentage of p-values smaller than 10%	33.8%	35.7%

Notes: This table presents the percentage of statistically significant treatment effects based on a selection of 157 outcomes in the NFP study. Among those, 5 outcomes are surveyed at birth; 1 at age 6 months old; 2 at age 1; 51 outcomes are surveyed at age 2; 37 outcomes are surveyed at age 6; 3 outcomes are surveyed at age 9; and 58 outcomes are surveyed at age 12.

Our main analysis is presented in Tables 6-9. We investigate outcomes on child health, family environments, child cognition, socio-emotional development and achievement scores. We divide the outcomes into blocks of variables according to their meaning. Each block contains variables of similar content surveyed at the same period. Each table consists of six columns for each gender. The first four columns display the basic statistics for each block of outcomes: (1) control group mean, (2) conditional difference in mean, (3) conditional effect size, and (4) asymptotic p-value. The fifth column presents the one-sided single hypothesis permutation p-value that accounts for the characteristics of the NFP randomization protocol as described in Section 4. The sixth column shows the adjusted stepdown p-values that corrects for multiple-hypothesis testing.

We find statistically significant effects in both maternal and child outcomes. For mothers, the NFP

intervention significantly improved mental health, home environment and parenting skills. On average, treated boys were healthier at birth and experienced an increase in cognitive abilities by age 6. By age 12, treated boys outperformed controls in math and reading achievement. Meanwhile, treated girls experienced an improvement in cognitive and socio-emotional skills at age 6. At age 12, treated girls had significantly lower body-mass index (BMI). We also find that controlling for background characteristics increases the precision of treatment effect estimates, often rendering significance in multiple-hypothesis inference. Tables E.4-E.7 of Appendix E investigates if our findings hold when correcting for attrition using the inverse probability weighting method. The inference results presented in Tables E.4-E.7 are very close to findings of Tables 6-9. This corroborates our premise that attrition does not play a substantial role in the NFP intervention. The remainder of this section discusses each outcome category in more detail.

Child Health Outcomes

Table 6 displays the NFP results on child's health. The first block shows the treatment effects on birth outcomes. Treated boys were relatively healthier at birth: there are strong positive effects on placenta weight, birth weight, head circumference, length and gestational age at delivery. The results are both significant at 5% significance level and robust to the multiple-hypothesis testing correction. Treated boys were, on average, 193 grams heavier and were delivered 0.7 weeks later than their control counterparts. In contrast, girls had no statistically significant effects. The second block focus on child health outcomes at age 12. We find statistically significant treatment effects on Body mass index (BMI) for girls. Boys have fewer statistically significant results, which do not survive the multiple-hypothesis correction.

Family environment

Table 7 displays the NFP results on family environment and parenting. The first two blocks display treatment effects on parenting beliefs and the home environment for ages 1 and 2. Treatment effects are statistically significant for both genders and robust to the stepdown procedure. By age 2, the NFP improved maternal parenting attitudes such as non-abusive and non-neglecting behaviors. We find effect sizes of approximately 0.3 standard deviations (SDs) for mothers of both females and males. The intervention improved home environment by 0.15 SDs for males and by 0.3 SDs for females. This positive impact in the home environment is mainly driven by treatment effects on HOME subscales associated with maternal involvement with the child, variety in daily stimulations and provision of appropriate playing material. Additionally, we find significant treatment effects on maternal mental health. Mothers of female children reported less anxiety, and better emotional stability, self-esteem and mastery skills. The effects range between 0.2 and 0.3 SDs. Mothers of male children also experienced an improvement in their mastery skills. However,

this effect does not survive the multiple-hypothesis correction. The bottom block examines the treatment effects on the cost of welfare programs participation. The NFP reduced the total cost of participation in AFDC/TANF, Food Stamps and Medicaid for mothers with male children at age 12.

Child Cognitive Outcomes

Table 8 displays the NFP results on cognition and achievement for children at ages 6 and 12. The first block analyzes the KABC (Kaufmann Assessment Battery for Children) measured at age 6. NFP increased IQ for both genders. However, only the results for boys are robust to the stepdown procedure. The second block displays different subsets of instruments from the KABC and the results are robust for boys and girls. Also, treated boys scored 0.25 SDs higher in the PPVT (Peabody Picture Vocabulary Test) than their control counterparts. We estimate two factor scores that summarize these measures: one that captures mainly cognitive skills, and another that combines both cognition and achievement. The treatment effect sizes are larger for boys than girls. For instance, NFP increased cognitive skills by 0.27 SDs for boys versus 0.12 SDs for girls.

Achievement Scores

The bottom part of table 8 presents the treatment effects on achievement outcomes at age 12. The results, again, are stronger for treated boys than girls, and are robust to the stepdown procedure. Treated boys scored .24 SDs higher in the reading section of the Tennessee Comprehensive Assessment Program (TCAP). Also, NFP improved math achievement scores exclusively for boys. Treatment effect sizes range between .15 and .22 SDs. In general, Table 8 documents a wide variety of effects on the cognitive skills of males. These results extend previous evaluations, which found this effect only for children born to mothers with low psychological resources (Kitzman et al., 2010).

Child Socio-Emotional Outcomes

Table 9 displays the NFP results on socio-emotional development. The upper block investigates the treatment effects on psychosocial functioning based on the Child Behavior Checklist (CBCL) at ages 2 and 6. The results are statistically significant only for girls, who had less conduct/attention problems (effect size: .27 SDs). Also, children's socio-emotional skills were assessed using the MacArthur Story Stem Battery (MSSB) surveyed at age 6. We find that treated girls improved pro-social skills (warmth and empathy) by .36 SDs and decreased aggressive behavior by .18 SDs. The bottom block of Table 9 considers socio-emotional outcomes at age 12. Treated boys experienced less internalizing disorders and fewer school absences (effect sizes of approximately .2 SDs). In summary, we find that the NFP intervention generated stronger effects on socio-emotional skills for girls and stronger effects on academic achievement for boys.

Table 6: Child Health Outcomes

	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	682.995	-1.826	-0.011	0.534	0.362	0.599	663.819	18.360	0.103	0.210	**0.032**	**0.083**
Birth Weight	3055.224	-126.604	-0.232	0.963	0.893	0.893	2997.486	193.305	0.276	**0.010**	**0.000**	**0.002**
Head Circumference	33.262	-0.007	-0.004	0.513	0.254	0.595	33.511	0.310	0.139	0.122	**0.053**	**0.053**
Length	49.665	0.184	0.069	0.291	0.285	0.588	49.918	0.556	0.153	**0.092**	**0.046**	**0.080**
Gestational Age at Delivery	39.119	-0.491	-0.217	0.925	0.805	0.887	38.544	0.698	0.201	**0.038**	**0.003**	**0.010**
Child Health Outcomes (Year 12)												
Any Injuries since Last Interview	0.164	-0.052	-0.149	0.119	**0.085**	0.230	0.224	-0.048	-0.122	0.176	0.124	0.425
# Hospitalizations for Injuries since Last Interview	0.009	-0.011	-0.112	0.156	0.213	0.388	0.010	-0.012	-0.117	0.142	**0.078**	0.341
Total # Injuries since Last Interview	0.197	-0.079	-0.181	**0.067**	**0.027**	0.113	0.268	-0.044	-0.085	0.264	0.257	0.672
Hospitalized since Last Interview	0.042	-0.036	-0.183	**0.050**	**0.050**	0.182	0.039	0.053	0.295	0.974	0.890	0.981
Have Chronic Condition/Health Problem	0.197	-0.002	-0.004	0.488	0.692	0.692	0.361	0.068	0.144	0.858	0.811	0.986
Standardized Child BMI	1.121	-0.255	-0.294	**0.014**	**0.005**	**0.025**	0.797	0.209	0.235	0.959	0.907	0.907

Notes: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Cntr. Mean) of each result set shows the mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Cd. Diff. Mn.) gives the conditional difference in means between the treatment group and the control group. The third column (Cd. Eff. Size) calculates the conditional effect size for the respective group. The fourth column (Ass. P-val.) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the difference in means between treatment and control groups. 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 fifth column (Block Perm. FL/Single P-val.) 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 4. 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 Perm. FL/Stepdown) provides p-values that accounts for multiple-hypothesis testing based on the stepdown algorithm of Romano and Wolf (2005a). 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 7: Family environment

	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
Home Environment, Parenting (Year 1) - Factor Scores												
Home Observation Measurement of the Environment (HOME)	0.000	0.354	0.354	**0.003**	**0.004**	**0.007**	0.000	0.208	0.208	**0.051**	**0.041**	**0.041**
Non-Abussive Parenting Attitudes (Bavolek)	0.000	0.289	0.288	**0.012**	**0.005**	**0.005**	0.000	0.273	0.273	**0.015**	**0.003**	**0.006**
Home Environment, Parenting (Year 2)- Factor Scores												
Home Observation Measurement of the Environment (HOME)	0.000	0.302	0.301	**0.010**	**0.003**	**0.006**	0.000	0.169	0.169	**0.092**	**0.075**	**0.075**
Non-Abussive Parenting Attitudes (Bavolek)	0.000	0.371	0.370	**0.003**	**0.006**	**0.006**	0.000	0.316	0.316	**0.006**	**0.003**	**0.006**
Home Environment, sub-scales (Year 1)												
Emotional/Verbal Responsivity of Mother	11.591	0.333	0.189	**0.070**	**0.033**	**0.093**	10.236	0.284	0.171	**0.091**	0.207	0.580
Avoidance of Restriction and Punishment	7.461	-0.037	-0.026	0.583	0.807	0.807	1.112	-0.039	-0.027	0.580	0.767	0.767
Organization of Environment	3.967	0.007	0.007	0.478	0.320	0.530	3.198	0.095	0.095	0.228	0.340	0.577
Provision of Appropiate Play material	11.266	0.391	0.223	**0.037**	**0.000**	**0.017**	1.149	0.118	0.074	0.286	0.233	0.520
Maternal Involvement with Child	7.735	0.335	0.237	**0.031**	**0.007**	**0.047**	5.991	0.148	0.104	0.207	**0.067**	0.340
Oppor. For Variety in Daily stimulation	3.259	0.249	0.218	**0.041**	**0.020**	**0.090**	2.643	0.305	0.267	**0.019**	0.107	0.403
Home Environment, sub-scales (Year 2)												
Emotional/Verbal Responsivity of Mother	13.046	-0.078	-0.052	0.650	0.607	0.607	8.591	-0.269	-0.179	0.914	0.853	0.853
Avoidance of Restriction and Punishment	2.061	0.299	0.171	**0.097**	0.327	0.510	4.286	0.053	0.027	0.416	0.400	0.767
Maternal Involvement with Child	6.150	0.108	0.088	0.248	0.193	0.503	7.368	0.177	0.141	0.131	0.340	0.743
Organization of Environment	6.742	0.176	0.187	**0.071**	0.200	0.433	5.175	-0.009	-0.011	0.534	0.640	0.850
Provision of Appropiate Play material	8.161	0.785	0.428	**0.000**	**0.000**	**0.013**	6.754	0.483	0.287	**0.012**	**0.087**	0.290
Oppor. For Variety in Daily stimulation	4.786	0.501	0.400	**0.001**	**0.000**	**0.013**	4.322	0.486	0.407	**0.001**	**0.007**	**0.047**
Maternal Mental Health (Year 2) - Factor Scores												
Anxiety	0.000	-0.247	-0.247	**0.030**	**0.029**	**0.068**	0.000	-0.038	-0.038	0.382	0.457	0.660
Depression	0.000	-0.129	-0.129	0.151	**0.094**	0.157	0.000	-0.062	-0.062	0.311	0.427	0.717
Positive Well-Being	0.000	0.101	0.101	0.209	0.344	0.344	0.000	-0.136	-0.136	0.853	0.887	0.887
Emotional Stability	0.000	0.207	0.207	**0.055**	**0.048**	**0.097**	0.000	0.050	0.049	0.347	0.470	0.634
Overall Mental Health	0.000	0.210	0.210	**0.051**	**0.057**	0.106	0.000	-0.014	-0.014	0.544	0.637	0.743
Self-Esteem	0.000	0.313	0.313	**0.007**	**0.002**	**0.008**	0.000	0.073	0.073	0.291	0.446	0.716
Mastery	0.000	0.286	0.286	**0.016**	**0.011**	**0.037**	0.000	0.198	0.198	**0.062**	**0.083**	0.244
Total Cost of Govt. Programs (Child Ages 1 - 12 Years)												
AFDC/TANF	2744.043	-166.474	-0.066	0.300	0.643	0.643	2743.386	-485.836	-0.188	**0.055**	**0.054**	0.101
Food Stamp	2996.965	-313.891	-0.192	**0.056**	0.349	0.482	3263.273	-357.933	-0.232	**0.030**	**0.067**	**0.093**
Medicaid	3543.761	-262.590	-0.158	**0.097**	0.442	0.558	3823.048	-271.631	-0.183	**0.070**	0.122	0.122

Notes: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Cntr. Mean) of each result set shows the mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Cd. Diff. Mn.) gives the conditional difference in means between the treatment group and the control group. The third column (Cd. Eff. Size) calculates the conditional effect size for the respective group. The fourth column (Ass. P-val.) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the difference in means between treatment and control groups. 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 fifth column (Block Perm. FL/Single P-val.) 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 4. 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 Perm. FL/Stepdown) provides p-values that accounts for multiple-hypothesis testing based on the stepdown algorithm of Romano and Wolf (2005a). 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 8: Cognitive Abilities and Achievement Outcomes

	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	8.981	0.235	0.078	0.276	0.163	0.430	9.775	-0.385	-0.133	0.836	0.647	0.647
Hand Movements	9.282	0.402	0.186	**0.082**	**0.025**	0.147	9.287	0.211	0.099	0.233	0.421	0.737
Matrix Analogies	8.632	0.148	0.087	0.257	0.271	0.536	8.478	0.316	0.199	**0.071**	0.122	0.428
Number Recall	9.423	0.397	0.138	0.144	0.109	0.383	8.952	0.954	0.400	**0.003**	**0.004**	**0.030**
Photo Series	6.967	0.434	0.216	**0.053**	**0.045**	0.216	6.774	0.050	0.024	0.431	0.578	0.779
Spatial Memory	8.434	0.204	0.084	0.266	0.295	0.457	8.526	0.278	0.115	0.204	0.170	0.439
Triangles	8.868	0.378	0.163	0.104	0.155	0.457	9.120	0.126	0.055	0.344	0.160	0.464
Word Order	9.693	0.002	0.001	0.497	0.300	0.300	9.191	0.751	0.293	**0.017**	**0.008**	**0.050**
Kaufman Assessment Battery for Children (Year 6)												
Nonverbal	89.203	2.035	0.230	**0.049**	**0.045**	**0.090**	89.244	1.421	0.152	0.136	0.175	0.221
Sequential Processing	96.582	1.696	0.132	0.161	**0.067**	0.117	94.507	3.768	0.320	**0.011**	**0.010**	**0.023**
Simultaneous Processing	88.844	1.928	0.191	**0.080**	**0.075**	**0.075**	89.919	0.709	0.071	0.303	0.216	0.216
WISC-III, PPVT-III for Children (Year 6)												
Wechsler Intelligence Scale for Children (WISC-III)	96.256	1.061	0.059	0.324	0.336	0.336	90.657	1.778	0.104	0.223	0.279	0.279
Peabody Picture Vocabulary Test (PPVT-III)	83.299	1.773	0.163	0.108	0.165	0.289	82.466	2.646	0.252	**0.039**	**0.008**	**0.016**
Child Cognition (Year 6) - Factor Scores												
Cognition + achievment (KABC, PPVT, WISC)	0.000	0.110	0.110	0.208	**0.084**	**0.084**	0.000	0.185	0.185	**0.088**	**0.083**	**0.083**
Cognitive skills (Mental Processing Composite-KABC)	0.000	0.119	0.119	0.186	**0.082**	0.113	0.000	0.273	0.273	**0.024**	**0.022**	**0.030**
Reading Achievement for the Child (Year 12)												
Average Reading Grade (Grades 1 - 5)	2.703	0.032	0.042	0.380	0.220	0.368	2.327	0.071	0.096	0.251	0.104	0.200
TCAP % Language (School Years 1 - 5, Grd 3+)	50.854	0.137	0.006	0.484	0.189	0.370	38.063	5.243	0.240	**0.058**	**0.009**	**0.035**
TCAP % Reading (School Years 1 - 5, Grd 3+)	41.607	0.146	0.007	0.480	0.163	0.351	34.912	1.511	0.076	0.305	**0.099**	0.237
PIAT Total Reading (Derived Score)	90.246	0.715	0.075	0.295	0.358	0.417	89.292	1.356	0.101	0.222	0.111	0.148
PIAT Reading Comprehension (Derived Score)	88.307	-0.196	-0.022	0.563	0.576	0.576	87.585	2.272	0.195	**0.077**	**0.037**	0.112
PIAT Reading Recognition (Derived Score)	94.221	2.276	0.189	**0.095**	0.147	0.347	92.456	0.379	0.026	0.422	0.171	0.171
Math Achievement for the Child (Year 12)												
Average Math Grade (Grades 1 - 5)	2.634	0.036	0.044	0.371	0.287	0.405	2.368	0.123	0.162	0.127	**0.054**	**0.054**
TCAP % Math (School Years 1 - 5, Grd 3+)	46.935	1.566	0.066	0.326	0.217	0.369	40.346	3.392	0.153	0.159	**0.030**	**0.076**
PIAT Mathematics (Derived Score)	87.188	-0.125	-0.013	0.538	0.758	0.758	86.316	2.247	0.223	**0.054**	**0.043**	**0.077**

Notes: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Cntr. Mean) of each result set shows the mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Cd. Diff. Mn.) gives the conditional difference in means between the treatment group and the control group. The third column (Cd. Eff. Size) calculates the conditional effect size for the respective group. The fourth column (Ass. P-val.) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the difference in means between treatment and control groups. 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 fifth column (Block Perm. FL/Single P-val.) 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 4. 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 Perm. FL/Stepdown) provides p-values that accounts for multiple-hypothesis testing based on the stepdown algorithm of Romano and Wolf (2005a). 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 9: Socio-emotional Outcomes

	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	0.000	-0.336	-0.336	**0.002**	**0.004**	**0.016**	0.000	0.163	0.163	0.893	0.848	0.946
Anxiety Problems	0.000	-0.191	-0.191	**0.057**	0.217	0.217	0.000	-0.029	-0.029	0.411	0.556	0.858
Pervasion Developmental Problems	0.000	-0.262	-0.262	**0.013**	**0.054**	0.125	0.000	0.084	0.084	0.747	0.642	0.885
Attention Deficit Hyperactivity Disorder	0.000	-0.239	-0.239	**0.027**	**0.019**	**0.062**	0.000	0.078	0.078	0.725	0.753	0.929
Oppositional Defiant Problems	0.000	-0.224	-0.223	**0.036**	**0.059**	0.103	0.000	0.113	0.113	0.822	0.873	0.873
Child Behavior Checklist (Year 6) - Factor Scores												
Affective Problems	0.000	-0.063	-0.063	0.318	0.436	0.712	0.000	-0.107	-0.108	0.193	0.176	0.530
Anxiety Problems	0.000	-0.085	-0.085	0.241	0.473	0.669	0.000	0.108	0.108	0.787	0.871	0.871
Somatic Problems	0.000	0.083	0.083	0.732	0.856	0.856	0.000	0.061	0.061	0.674	0.463	0.679
Attention Deficit Hyperactivity Problems	0.000	-0.269	-0.269	**0.017**	**0.057**	0.200	0.000	-0.053	-0.053	0.341	0.293	0.654
Oppositional Defiant Problems	0.000	-0.032	-0.032	0.402	0.339	0.679	0.000	-0.103	-0.103	0.222	0.271	0.658
Conduct Problems	0.000	-0.269	-0.269	**0.013**	**0.002**	**0.012**	0.000	-0.017	-0.017	0.447	0.435	0.758
MacArthur Story Stem Battery (MSSB) (Year 6) - Factor Scores												
Dysregulated Aggression	0.000	-0.040	-0.040	0.372	**0.099**	0.197	0.000	-0.179	-0.179	0.109	**0.056**	0.228
Warmth and Empathy	0.000	0.360	0.360	**0.004**	**0.006**	**0.025**	0.000	-0.097	-0.097	0.767	0.557	0.936
Emotional Integration	0.000	-0.045	-0.045	0.633	0.813	0.813	0.000	0.022	0.022	0.437	0.566	0.856
Performance Anxiety	0.000	-0.037	-0.037	0.381	**0.075**	0.210	0.000	0.051	0.051	0.637	0.811	0.811
Aggression	0.000	-0.182	-0.182	**0.066**	**0.001**	**0.006**	0.000	-0.151	-0.151	0.147	**0.082**	0.297
Internalizing, Externalizing, Absences (Year 12)												
Presence of Internalizing Disorders	0.239	-0.037	-0.088	0.253	0.310	0.649	0.403	-0.098	-0.205	**0.062**	**0.037**	**0.071**
Presence of Externalizing Disorders	0.182	-0.025	-0.065	0.310	0.592	0.831	0.187	0.085	0.224	0.943	0.840	0.840
Average # of Absences (School Years 1 - 5)	10.186	0.399	0.054	0.652	0.681	0.681	11.803	-1.798	-0.234	**0.034**	**0.026**	**0.074**

Notes: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Cntr. Mean) of each result set shows the mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Cd. Diff. Mn.) gives the conditional difference in means between the treatment group and the control group. The third column (Cd. Eff. Size) calculates the conditional effect size for the respective group. The fourth column (Ass. P-val.) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the difference in means between treatment and control groups. 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 fifth column (Block Perm. FL/Single P-val.) 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 4. 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 Perm. FL/Stepdown) provides p-values that accounts for multiple-hypothesis testing based on the stepdown algorithm of Romano and Wolf (2005a). 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.

7 Decomposition of Treatment Effects

We consider two types of mechanisms underlying the NFP treatment effects. The first type considers maternal investments and home environment as mediators, which tackles the question of how parental investments respond to interventions and how such responses affect children's skill formation. The second set of mechanisms considers children's skills, which addresses the question of how program-induced enhancement of early skills translates to improvements on later ones.

Section 7.1 provides a statistical description of the mediators we analyze. In Section 7.2, we examine if program improvement of children's skills at age 6 were mediated by the effects on health at birth and home investments at age 2. In Section 7.3, we assess if the enhancement of child outcomes at age 12 were mediated by the effects on cognitive and socio-emotional skills at age 6.

7.1 NFP Effects on Traits at Age 2 and 6

Figures 1–2 present the kernel densities of the factor scores used as mediators in our analysis. We also display the p-value for the single hypothesis inference of no treatment effect as described in Section 4.

Figure 1 shows the density of mediators at birth and age 2. It shows that NFP participation significantly increased birth weight of treated boys. At age 2, the NFP intervention had significantly improved the quality of home environment for both boys and girls (p-values: 0.003 and 0.07, respectively). It also significantly reduced maternal abusive/neglecting attitudes (p-values: 0.01 for females and 0.00 for males). The maternal characteristics for mothers of females also improved: maternal anxiety decreased (p-value: 0.03) and maternal self-esteem increased (p-value: 0.00). Maternal mastery skills for mothers of both boys and girls were positively influenced as well.

Figure 2 shows the density of mediators at age 6. It shows that the NFP intervention enhanced cognition for boys (p-values: 0.02). For girls, it reduced attention deficit, conduct problems, and aggression issues (p-values: 0.06, 0.00, and 0.00, respectively). NFP also increased warmth or empathy skills (pro-social skills) for girls (p-value: 0.01).

7.2 Decomposition of Treatment Effects at Age 6

A large body of evidence corroborates the importance of early parental investments on promoting cognitive and socio-emotional development later in life (Almond and Currie, 2010; Cunha and Heckman, 2007a; Heckman, 2008). We investigate this premise by examining how early variables impact later skills at age 6. By later skills we mean cognition, warmth/empathy, attention problems, conduct problems and aggression. We only decompose skills whose treatment effect is statistically significant according to the inference method

explained in Section 6. Therefore, the eligible variables for decomposition differ by gender. Figures 3 and 4 decompose the treatment effect on these variables into components associated with changes on factors evaluated at birth and age 2. We compute p-values using the bootstrap method.

We find that cognition at age 6 was enhanced through home environment and parenting practices for boys and girls. Specifically, the program improvement of home investments at age 2 explains 35% of the treatment effect on cognition at age 6 for girls (p-value 0.03). The respective number for boys is 22% (p-value 0.05). Similarly, the enhancement of parenting practices explains 14% of the treatment effect on cognition for girls (p-value 0.08). For boys, parenting practices explain 11% (p-value 0.05). Other contributions to cognition were birth weight and maternal anxiety.[12] Birth weight gains explain 14% of the cognitive treatment effect for boys (p-value: 0.06). For girls, maternal anxiety explains 25% (p-value: 0.08).

We find that the intervention-induced change on home environment accounts for 21% of the treatment effect on female warmth/empathy (p-value: 0.01). Home environment also accounts for 16% of the reduction in female aggression problems (p-value: 0.09). Program enhancement of parenting practices at age 2 explains 9% of the reduction in attention problems (p-value: 0.05) and 11% of the improvement in female warmth/empathy (p-value: 0.02). For boys, parenting practices explains 8% of the reduction in aggression (p-value: 0.09).

Maternal skills also influenced child development. For girls, the decrease in maternal anxiety explains 14% of the reduction in conduct problems (p-value: 0.06). The contribution of maternal self-esteem is negative and accounted for 29% of the treatment effect on warmth and empathy. This result is consistent with research that argues that improvements in self-esteem may increase selfishness (Burr and Christensen, 1992). Conversely, the mother's maternal mastery explains 29% of the improvement in female empathy/warmth (p-value: 0.09).

7.3 Decomposition of Treatment Effects at Age 12

Figures 5–7 decompose statistically significant treatment effects of variables at age 12 into treatment-induced changes of skills at age 6.

We find that changes in male cognitive abilities at age 6 play a substantial role in explaining achievement scores at age 12. Male cognitive effects at age 6 lead to an increase in the average percentile of the Tennessee Comprehensive Assessment Program (TCAP) by 2.1 percentage points for language and 2.7 percentage points for math (p-value: 0.07).[13] The improvement of cognitive skills also explains 46% of the treatment effect on PPVT reading comprehension (p-value 0.08) and 51% of the treatment effect on PPVT math scores

[12] Birth weight, a measure that summarizes prenatal investments and fetal development, is usually associated with childhood development (Breslau et al., 1994; Currie and Moretti, 2005).

[13] This test was taken for grades 3 and above. The control group had a 38 percentile mean in reading and a 40.3 percentile mean in math.

(p-value 0.04) for boys. Remarkably, cognition explains 66% of the treatment effect in average math GPA for boys (p-value: 0.07).

Male cognitive effects also contribute to explain the treatment effects on class absenteeism and internalizing problems. The inmprovement in cognition decreases the average number of day absences between the first and the fifth years of schooling by .36 (p-value .06).[14] Also, they explain 17% of the male treatment effect on internalizing behavior (p-value: 0.06) and 14% of the results on male anxiety/depression at age 12 (p-value: 0.03). The intervention also reduced aggression at age 6, which accounts for 12% of the reduction in the likelihood of being anxious/depressed (see Figure 6).[15]

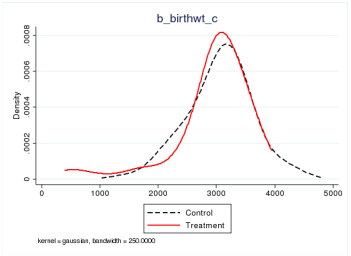
The program decreased the likelihood of being overweight for females. The reduction in conduct problems fostered by the program at age 6 explains 19% of the treatment effect on standardized female BMI (p-value: 0.06). We also decompose the treatment effects on female risky behavior at age 12. However, this effect is largely unexplained by our mediators.

The remarkable role that gains in cognition play in explaining the effects on achievement at age 12 for males is likely a consequence of the program characteristics. The NFP treatment started at prenatal stages and continued until age 2. The timing coincided with critical periods for language, speech, and early cognitive skills formation (Thompson and Nelson, 2001). The differential treatment effects on cognition and achievement by gender follow the pattern of treatment effects on health at birth, which are stronger for boys. This goes in line with evidence that males are more sensitive to changes during the prenatal period (Kraemer, 2000).

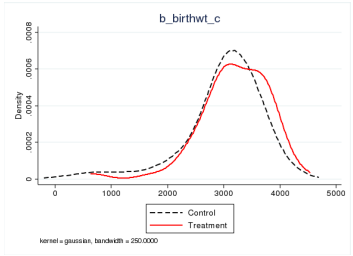
[14] The average number of absences in those years was 11.8 days for the control group.
[15] Other related results are shown in Tables G.8 - G.11 in the Appendix.

Figure 1: Kernel Densities of Factor Scores - 2 years skills

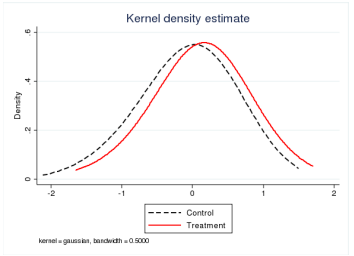
(a) Birth weight, females
$p = .893$



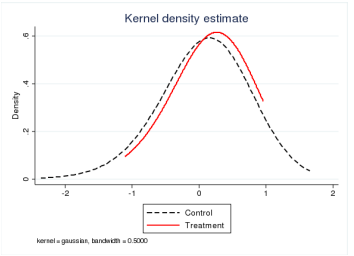
(b) Birth weight, males
$p = .03$



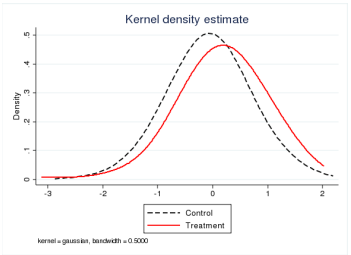
(c) Home environment, females
$p = .00$



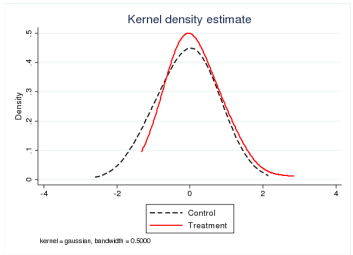
(d) Home environment, males
$p = .07$



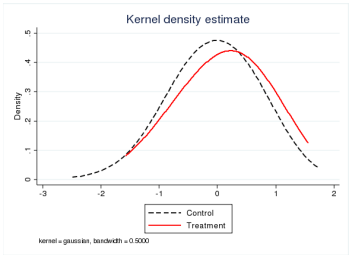
(e) Non-abusive parenting attitudes, females
$p = .01$



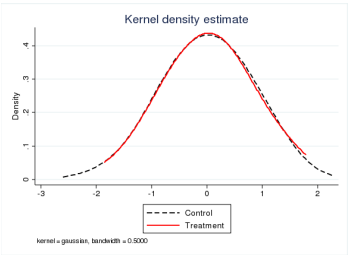
(f) Non-abusive parenting attitudes, males
$p = .00$



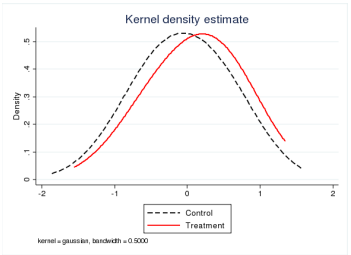
(g) Maternal less anxiety, females
$p = .03$



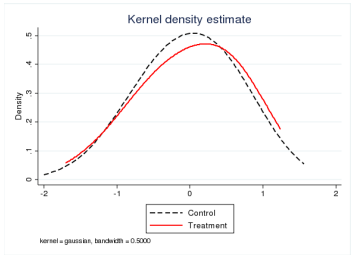
(h) Maternal less anxiety, male
$p = .46$



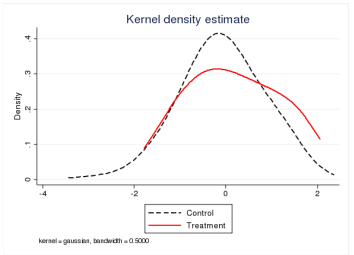
(i) Maternal self-esteem , females
$p = .00$



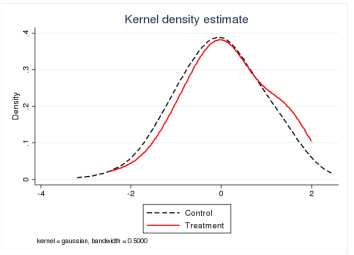
(j) Maternal self-esteem, males
$p = .45$



(k) Maternal mastery, females
$p = .01$



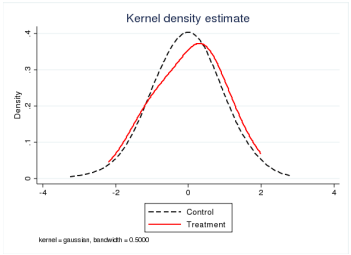
(l) Maternal mastery, males
$p = .08$



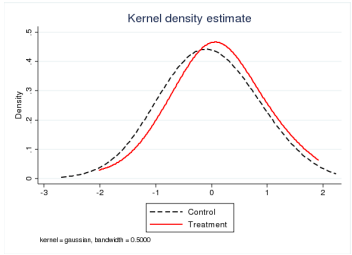
Notes: Kernel density functions based on Bartlett (1937) using a normal kernel. Numbers on the top of each figure correspond to one-sided FL permutation p-values testing the equality of factor scores means as explained in Section 4 and presented in table 7.

Figure 2: Kernel Densities of Factor Scores - 6 years skills

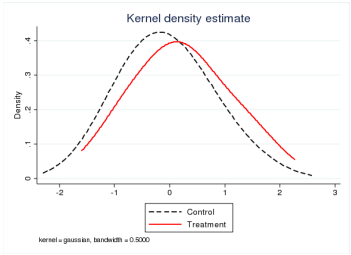
(a) Cognition, females
$p = .08$



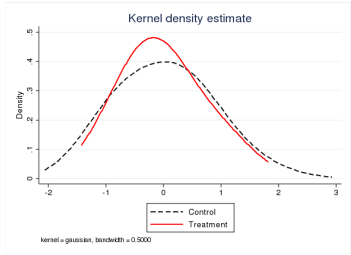
(b) Cognition, males
$p = .02$



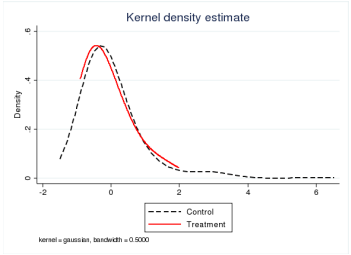
(c) Warmth/empathy, females
$p = .01$



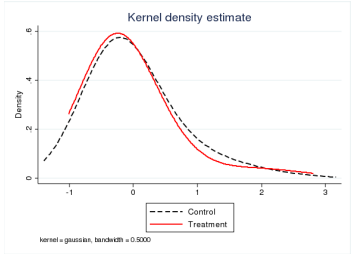
(d) Warmth/empathy, males
$p = .55$



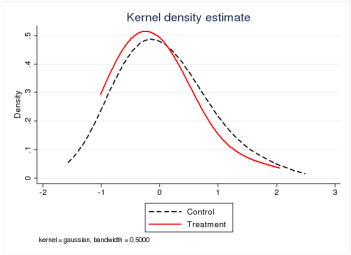
(e) Aggression, females
$p = .00$



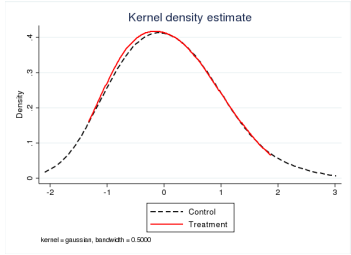
(f) Aggression, males (g) Attention problems, females
$p = .08$



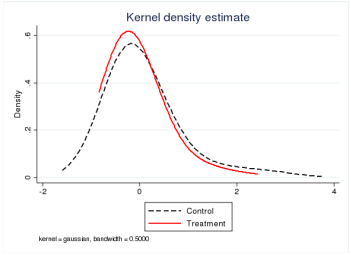
$p = .0.06$



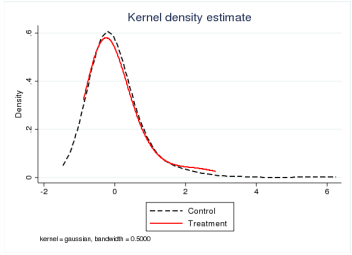
(h) Attention problems, males
$p = .29$



(i) Conduct problems, females
$p = .00$



(j) Conduct problems, males
$p = .43$



Notes: Kernel density functions based on Bartlett (1937) using a normal kernel. Numbers on the top of each figure correspond to one-sided FL permutation p-values testing the equality of factor scores means as explained in 4 and presented in table 8 and 9.

Figure 3: Decomposition of Treatment Effects on Female Outcomes - Age 6

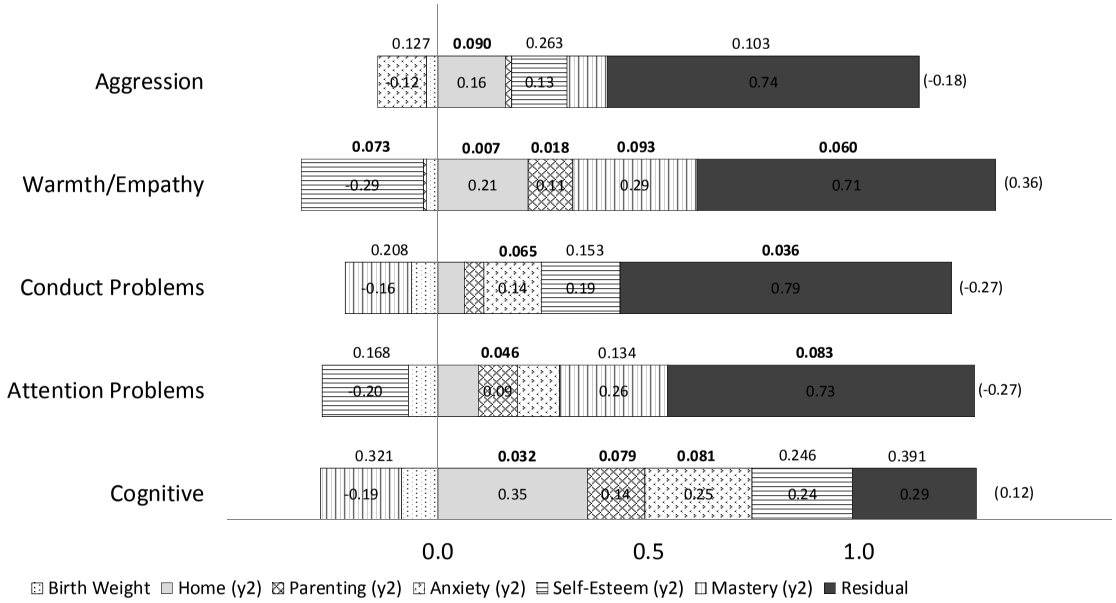
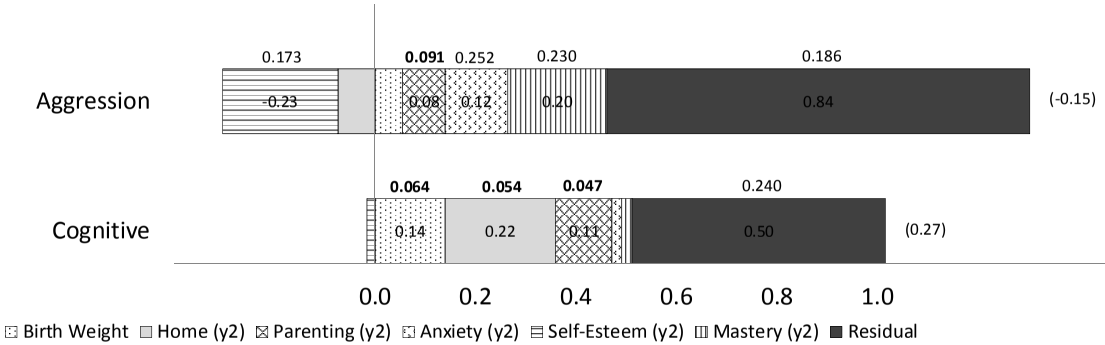


Figure 4: Decomposition of Treatment Effects on Male Outcomes - Age 6



Notes: These bar graphs show the fractions of the overall treatment effect explained by changes in the mediators as well as the unexplained portion (residual). The names of the outcomes are located on the left side of the bar graphs. Note that the fractions do not add up to 1 because some of the mediators contribute such a small and statistically insignificant fraction that they have not been shown. Values to the left of the horizontal line are negative and values to the right of the horizontal line are positive; bold values are significant at the 10% level. Values above the bars are the p-values associated with the mediation fraction, and values inside the bars are the mediation fractions themselves. Numbers in parentheses to the right of the bars are the total treatment effects. Negative values correspond to the mediator working in the opposite direction of the overall treatment effect, while positive values correspond to the mediator working in the same direction as the overall treatment effect. The mediators are the home environment, parenting attitudes, maternal anxiety, maternal self-esteem, and maternal mastery, all of which were measured when the child was 6 years old, as well as birth weight. The following controls were used: 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.

Figure 5: Decomposition of Treatment Effects on Male Achievement Outcomes - Age 12

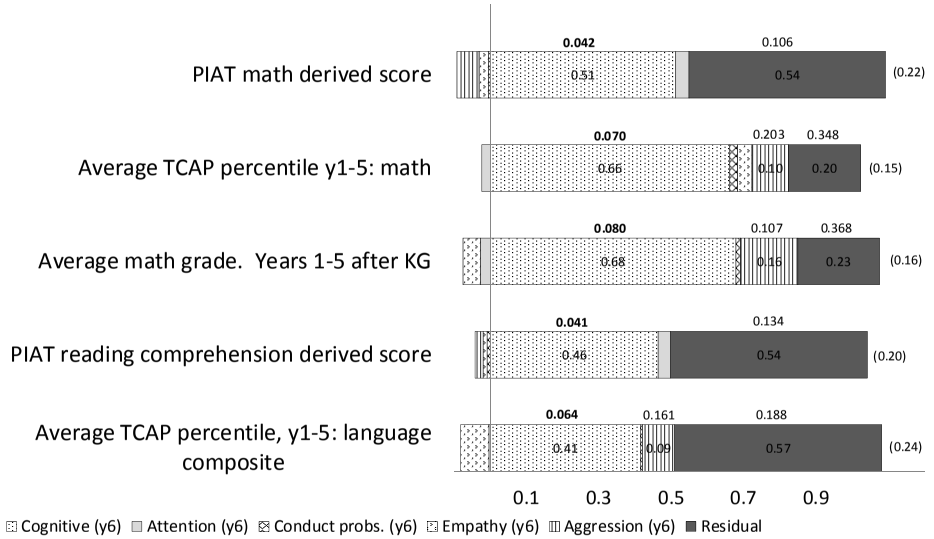
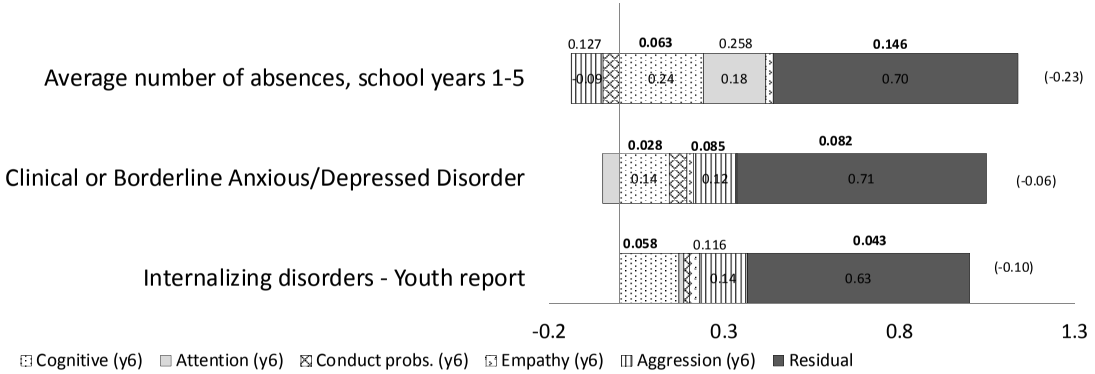
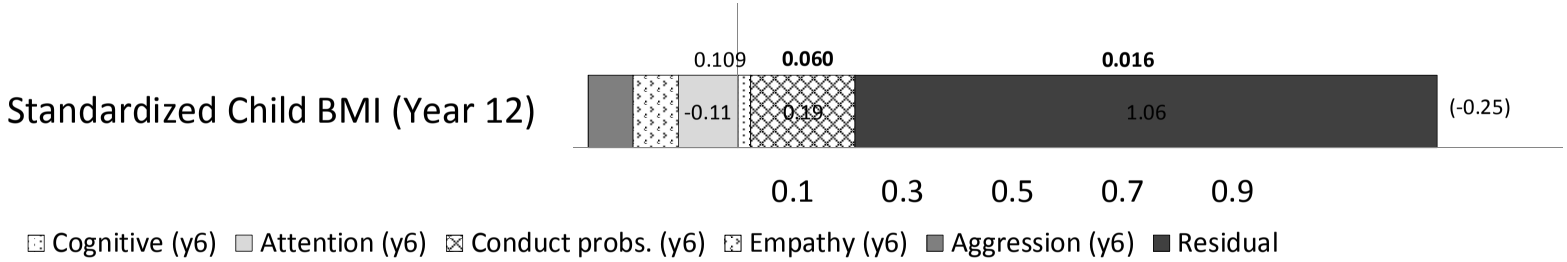


Figure 6: Decomposition of Treatment Effects on Male Non-cognitive Outcomes - Age 12



Notes: These bar graphs show the fractions of the overall treatment effect explained by changes in the mediators as well as the unexplained portion (residual). The names of the outcomes are located on the left side of the bar graphs. Note that the fractions do not add up to 1 because some of the mediators contribute such a small and statistically insignificant fraction that they have not been shown. Values to the left of the horizontal line are negative and values to the right of the horizontal line are positive; bold values are significant at the 10% level. Values above the bars are the p-values associated with the mediation fraction, and values inside the bars are the mediation fractions themselves. Numbers in parentheses to the right of the bars are the total treatment effects. Negative values correspond to the mediator working in the opposite direction of the overall treatment effect, while positive values correspond to the mediator working in the same direction as the overall treatment effect. The mediators are cognition, attention problems, conduct problems, aggression and warmth and empathy, all of which were measured when the child was 6 years old. The following controls were used: 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, maternal parenting attitudes and the age at the 12 year interview.

Figure 7: Decomposition of Treatment Effects on Female Outcomes - Age 12



Notes: These bar graphs show the fractions of the overall treatment effect explained by changes in the mediators as well as the unexplained portion (residual). The names of the outcomes are located on the left side of the bar graphs. Note that the fractions do not add up to 1 because some of the mediators contribute such a small and statistically insignificant fraction that they have not been shown. Values to the left of the horizontal line are negative and values to the right of the horizontal line are positive; bold values are significant at the 10% level. Values above the bars are the p-values associated with the mediation fraction, and values inside the bars are the mediation fractions themselves. Numbers in parentheses to the right of the bars are the total treatment effects. Negative values correspond to the mediator working in the opposite direction of the overall treatment effect, while positive values correspond to the mediator working in the same direction as the overall treatment effect. The mediators are cognition, attention problems, conduct problems, aggression and warmth and empathy, all of which were measured when the child was 6 years old. The following controls were used: 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, maternal parenting attitudes and the age at the 12 year interview.

8 Conclusion

A vast literature on the economics of psychology and education provides evidence of the importance of early childhood investment in enhancing later life outcomes (see Carneiro et al. (2003); Shonkoff and Phillips (2000) for a survey). Our paper adds to this discussion by enlightening the understanding of how early childhood interventions that rely on home visitations during infancy impact skill formation.

The Nurse-Family Partnership (NFP) is the most cited home visiting program in the US (Howard and Brooks-Gunn, 2009). NFP consists of nurse visits to disadvantaged, first time, poor, young and unmarried mothers. It fosters child development by enhancing parenting skills towards healthier, more responsive, and competent care for children. Currently, NFP surrogates provide home visiting service for more than $20,000$ families in 31 states across the U.S. In this paper we analyse the data of the NFP randomized controlled trial performed in Memphis, TN, 1990.

We investigate outcomes up to age 12 and we improve upon previous evaluations by performing a small-sample inference that accounts for the characteristics of the NFP randomization protocol. We address the problem of selective reporting of statistically significant outcomes (i.e., "cherry picking") by implementing a multiple-hypothesis testing that relies on the stepdown procedure of Romano and Wolf (2005a). We also examine the underlying mechanisms generating the treatment effects. This enables us to interpret treatment effects at early stages as building blocks of treatment effects at later stages. We decompose significant treatment effects into interpretable components that can be associated with program-induced changes in children's early skills and parental investments.

We find statistically significant effects on maternal mental health, home environment and parenting skills. For boys, we find that the treatment fostered health status at birth and resulted in better cognitive abilities by age 6. We also find that treatment enhanced achievement scores for boys at age 12. Treated girls experienced an improvement in cognitive and socio-emotional skills at age 6. At the same age, treated girls had significant lower body-mass index (BMI) than the control counterpart.

The decomposition of statistically significant treatment effects allow us to investigate the channels through which the intervention operated. We find that maternal investment at birth and child skills at age 2 boosted middle childhood outcomes. Specifically, we find that NFP improved home investments, parenting attitudes and maternal mental health for boys and girls at age 2. At age 6, the program enhanced cognitive skills for both boys and girls while improving early socio-emotional skills for girls. We find that male enhanced cognition at age 6 plays a substantial role in explaining treatment effects on achievement scores at age 12. Cognitive ability gains at age 6 explain between 40% and 60% of the treatment effects in those achievement scores.

References

Almond, D. and J. Currie (2010). Human capital development before age five. Technical report, National Bureau of Economic Research.

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

Association, A. P. and A. P. A. T. F. on DSM-IV. (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR.* American Psychiatric Publishing, Inc.

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.

Breslau, N., J. DelDotto, G. Brown, S. Kumar, S. Ezhuthachan, K. Hufnagle, and E. Peterson (1994). A gradient relationship between low birth weight and iq at age 6 years. *Archives of Pediatrics and Adolescent Medicine 148*(4), 377.

Burr, W. and C. Christensen (1992). Undesirable side effects of enhancing self-esteem. *Family Relations*, 460–464.

Carneiro, P., F. Cunha, and J. J. Heckman (2003, October 17). Interpreting the evidence of family influence on child development. Paper presented at the conference "The Economics of Early Childhood Development: Lessons for Economic Policy," Federal Reserve Bank of Minneapolis, Minneapolis, MN, October 17, 2003.

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 (2007a). Identifying and estimating the distributions of *Ex Post* and *Ex Ante* returns to schooling: A survey of recent developments. *Labour Economics 14*(6), 870–893.

Cunha, F. and J. J. Heckman (2007b, 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.

Currie, J. and E. Moretti (2005). Biology as destiny? short and long-run determinants of intergenerational transmission of birth weight.

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.

Gorsuch, R. (1983). *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Heckman, J. and R. Pinto (2014a). Causal analysis after haavelmo: Definitions and a unified analysis of identification. *Theoretical Economics*.

Heckman, J. and R. Pinto (2014b). Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs. *Econometric Reviews*.

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.

Heckman, J. J. (2008). Role of income and family influence on child outcomes. *Annals of the New York Academy of Sciences 1136*(Reducing the Impact of Poverty on Health and Human Development: Scientific Approaches), 307–323.

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

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

Heckman, J. J., R. Pinto, and P. A. Savelyev (2012). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. Unpublished manuscript, University of Chicago, Department of Economics (first draft, 2008). Under revision, *American Economic Review*.

Howard, K. and J. Brooks-Gunn (2009). The role of home-visiting programs in preventing child abuse and neglect. *The Future of Children*, 119–146.

Imai, K., L. Keele, and T. Yamamoto (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science 25*(1), 51–71.

Jo, B. (2008). Causal inference in randomized experiments with mediational processes. *Psychological Methods 13*(4), 314.

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: follow-up of a randomized trial among children at age 12 years. *Archives of Pediatrics and Adolescent Medicine 164*(5), 412.

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.

Kraemer, S. (2000). The fragile male. *Clinical Medicine NetPrints*, 1.

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

MacKinnon, D., A. Fairchild, and M. Fritz (2006). Mediation analysis.

Olds, D. (2002). Prenatal and infancy home visiting by nurses: From randomized trials to community replication. *Prevention Science 3*(3), 153–172.

Olds, D., P. Hill, J. Robinson, N. Song, and C. Little (2000). Update on home visiting for pregnant women and parents of young children. *Current Problems in Pediatrics 30*(4), 107.

Olds, D., H. Kitzman, R. Cole, and J. Robinson (1997). Theoretical foundations of a program of home visitation for pregnant women and parents of young children. *Journal of Community Psychology 25*(1), 9–25.

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.

Pinto, R. (2012). Evaluation of small-sample compromised randomization: Long-term effects of early childhood intervention on health and addictive behavior. *Brazilian Review of Econometrics* (2).

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

Romano, J. P. and M. Wolf (2005b). Stepwise multiple testing as formalized data snooping. *Econometrica 73*(4), 1237–1282.

Shonkoff, J. P. and D. Phillips (2000). *From Neurons to Neighborhoods: The Science of Early Child Development*. Washington, DC: National Academy Press.

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

Thompson, R. A. and C. A. Nelson (2001, January). Developmental science and the media: Early brain development. *American Psychologist 56*(1), 5–15.

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*. John Wiley and Sons.

Appendix

A NFP Randomization Protocol

Pregnant women were enrolled as they visited the Medical Center during pregnancy. The randomization protocol was sequential, that is to say that each participant was randomized according to the order of enrollment. Pregnant women who accepted 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 and
5. 4 geographic regions of residence.[16]

Within strata, randomization was perform 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 of the elder sibling.
2. Else, the participant was randomized according in to control (C) or treatment group (T). However, if the sample size difference between treatment and control group is larger that a threshold, the participant is deterministically assigned to the treatment status that has fewer participants.
3. Next the participant was randomized again into her final treatment status. If in step 2 the participant was assigned to the control, she is next randomized in to:

   - *Group 1*: control women that received free transportation to and from their prenatal appointments (sample size: 166).
   - *Group 2*: control women that 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 that had 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)

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

As in the previous step, if the absolute difference in group size exceed 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.

B Instruments

B.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 (Bradley and Caldwell (1984)). The test is based on the child and is carried on during a visit of 45 to 90 minutes. A more in depth explanation of the HOME inventory is in Bradley and Caldwell (1984).

There are several versions of the inventory. The initial version, Infant/Toddler HOME, is for kids aged 0 to 3 years old. It consists of 45 binary-choice items grouped into 6 subscales. The Early Childhood HOME is for kids 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 kids 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 environments.*
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.

B.2 KABC

The Kafuman Assessment Battery for Children (KABC) 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 intelligence. 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 sequenced.*
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 focus 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.

B.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, a 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.

B.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.

B.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.

B.6 C-DISC

The Computerized Diagnostic Interview Schedule for Children (C-DISC) is a comprehensive and structured interview that studies mental health disorders such as general anxiety, panic, eating, elimination, depression, ADHD, and conduct. The majority of the questions are "yes and no". Some questions have additional options such as "sometimes". The C-DISC has about $3,000$ questions with 358 core questions that the child must answer. About 1300 questions are asked depending on the answer to the core questions. There are 732 questions that ask about age of onset and treatment for reported symptoms. Lastly, there are about 700 questions from the optional whole-life module. The NFP Program used C-DISC when the child was 9 years old.

B.7 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 where 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 child's inner workings of the 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.

C Method for Permutation-based Inference

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

$$Y_i = D_i Y_{i,1} + (1 - D_i) Y_{i,0}, \tag{16}$$

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 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-2. $Y(\mathbf{d}) \perp\!\!\!\perp D \mid X;\ \mathbf{d} \in \mathrm{supp}(D)$,

where variables $X = (X_i;\ i \in J), D = (D_i;\ i \in J)$ are N-dimensional vectors of treatment assignments and pre-program variables, and $Y(\mathbf{d}) = (Y_{i,d_i}; i \in J,\ d_i \in \{0,1\})$ and $\mathbf{d} \in \mathrm{supp}(D) = \{0,1\}^{|J|}$ denotes the vector of counterfactual outcomes. In the same fashion, we represent the vector of observed outcomes of Equation (16) 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-2. $Y(\mathbf{d}) \overset{d}{=} Y(\mathbf{d}') \mid X \; ; \; \mathbf{d}, \mathbf{d}' \in \text{supp}(D)$,

Hypothesis **H-2** can be restated in more tractable form:

Hypothesis H-1′. Under Assumption **A-2** and Hypothesis **H-2**, we have that $Y \perp\!\!\!\perp D \mid X$.

Testing Hypothesis **H-1′** poses some statistical challenges. First, small sample sizes cast doubt on inference that rely on the asymptotic behavior of test statistics. We address the problem of small sample size by generating the exact of a test statistic conditioned on data. Second, the presence of multiple outcomes allows for the arbitrary selection of statistically significant outcomes. The selectively reporting statistically significant outcomes is often termed *cherry picking* and generates a downward biased inference with smaller p-values. We solve the problem of cherry picking by implementing a multiple-hypothesis testing based on the stepdown procedure of (Romano and Wolf, 2005a). They explain that the stepdown procedure strongly controls for Family-wise error rate (FWER), while classical tests do not. Also, Romano and Wolf (2005a) shows that the strong FWER control can be obtained by ensuring 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).

In summary, our method is based on three steps. First, we seek to characterise the exact conditional distribution of $D|X$. Specifically we characterize the a multiset $D_x(\mathbf{d})$, defined by:

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

such that the distribution of D conditioned on realised data is uniform among elements of $D_x(\mathbf{d})$. Next we use the assumption of null hypothesis of no treatment effects, i.e. $H_0 : Y \perp\!\!\!\perp D|X$, to generate the exact conditional distribution of a test statistic $T(Y, D)|X$. Upon it, 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 necessary to describe 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 X by

$$H_L \; : \; Y_L \; \perp\!\!\!\perp \; D|X, \text{ where } Y_L = (Y_k \; : \; k \; \in \; L). \tag{17}$$

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

Our goal is to tests a single (joint) null hypothesis controlling for the probability of a 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 Hypothesis (17),

$$(Y_L, D)|X \stackrel{d}{=} (Y_L, gD)|X \ \forall \ g \in \mathbf{G}_X, \tag{18}$$

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

$$\mathbf{G}_X = \{g \ ; \ g : J \to J \text{ is a bijection and } \ g(j) = j' \Rightarrow (X_j) = (X_{j'})\},$$

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 (18) to generate a statistical test whose exact distribution the test statistic $T_L(Y_L, gD)$ is obtained by re-evaluating $T_L(Y_L, gD)$ as g varies in $\mathbf{G}_X$. Note that the inference on Hypothesis 17 is depend on the choice of statistics. That is to say that even though any statistics $T_L(Y_L, D)$ whose value provide 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,

$$T_L(Y_L, D) = \max_{k \in L} T_k(Y_k, D), \tag{19}$$

where $T_k(Y_k, D)$ is the t-statistics for outcome Y_k. Relation (18) implies that $T_L(Y_L, D)|X \stackrel{d}{=} T_L(Y_L, gD)|X$ for any $g \in \mathbf{G}_X$. Moreover, let $\mathbf{d} \in \{0,1\}^{|J|}$ such that $\text{P}(D = \mathbf{d}|X = x) > 0$, then the distribution of D conditioned on $X = x$ is uniform across elements of $\mathbf{D}_x(\mathbf{d})$ (see Lehmann and Romano (2005), Chapter 15). Thus, a critical value $c_{L,x}(Y_L, \mathbf{d}, \alpha)$ such that $\text{P}(T_L(Y_L, D) > c_{L,x}(Y_L, \mathbf{d}, \alpha)|X = x, H_L \text{ is true}) \leq \alpha$ can be computed as:

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

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

$$c_{L,x}(Y_L, \mathbf{d}, \alpha) = T_{L,x}^{(\lceil (1-\alpha)|\mathbf{D}_x| \rceil)} \tag{20}$$

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 $T_L(Y_L, \mathbf{d})$ actually observed, i.e. the p-value, is given by:

$$p_{L,x}(\mathbf{d}) = \inf_{\alpha \in [0,1]} \Big\{ c_{L,x}(Y_L, \mathbf{d}, \alpha) \leq T_L(Y_L, \mathbf{d}) \Big\}. \tag{21}$$

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

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

Thus:

$$T_{L,x}^{(r_{L,x})} = T_L(Y_L, \mathbf{d}). \tag{22}$$

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

$$p_{L,x}(\mathbf{d}) = 1 - \frac{r_{L,x}}{|\mathbf{D}_x(\mathbf{d})|}. \tag{23}$$

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

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

We implemented a inference method that tests for the multiple null hypothesis that each outcome $Y_k; k \in L$ is independent of treatment status D conditional on pre-program variables X. The representation of these multiple hypothesis in the same fashion as the single (joint) null hypothesis, namely:

$$H_L = \cap_{k \in L} H_{\{k\}} \; ; \; H_{\{k\}} : Y_k \perp\!\!\!\perp D | (Z, U).$$

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 familywise 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 P(reject $H_L|H_L$ 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 step-down procedure of Romano and Wolf (2005a) is less conservative as it accounts for the information about the dependence structure of p-values. The method is based on the monotonicity assumption, which, in our case, can be stated as:

$$c_{K,x}(Y_K, \mathbf{d}, \alpha) \geq c_{L_0,x}(Y_{L_0}, \mathbf{d}, \alpha) \text{ for any subset } K \text{ of } L \text{ containing } L_0 \text{ i.e. } L_0 \subset K \subset L. \tag{24}$$

Assumption (24) is satisfied by our choice of test statistic (19) and the fact that $L_0 \subset K$.

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

Algorithm 1.

Step 1: Set $L_1 = L$. If

$$\max_{k \in L_1} T_k(Y_k, \mathbf{d}) \leq c_{L,x}(Y_{L_1}, \mathbf{d}, \alpha) \ , \tag{25}$$

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

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

and go to Step 2.

$\vdots$

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

$$\max_{k \in L_j} T_k(Y_k, \mathbf{d}) \leq c_{L,x}(Y_{L_j}, \mathbf{d}, \alpha), \tag{26}$$

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

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

and go to Step $j+1$.

$\vdots$

We can compute the multiplicity-adjusted p-values of Equations(25)–(26) in the same fashion described by Equations (21)–(23).

C.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 a statistical inference by conditioning on those pre-program variables. Let Z be the pre-program variables that were not used in the randomization protocol and we ought to control for.

Variables Z precede the treatment intervention and therefore $Z \perp\!\!\!\perp D \mid X$ holds due to randomization. Under the hypothesis of no-treatment, $Y \perp\!\!\!\perp D \mid X$ also holds. These two relations imply that $Y \perp\!\!\!\perp D \mid (X, Z)$. Likewise Section 4.1, we can use this relation to to generate a permutation test that consider the strata formed by values of covariates X and Z. This way we can generates an inference method that non-parametrically condition on variables X and 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) tested a range of permutation methods for linear models. They found that Freedman and Lane (1983) generated 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 Z not used in the randomization protocol. According to Freedman and Lane (1983) method, our approach can be summarized by the following steps: (1) compute the residuals $Y - Z\hat{\boldsymbol{\beta}}$ such that $\hat{\boldsymbol{\beta}} = (Z'Z)^{-1}Z'Y$; (2) permute these residuals according to permutations $g \in \mathscr{G}_X$. (3) add these permuted residuals to $Z\hat{\boldsymbol{\beta}}$, call it $\tilde{Y}$; (4) regress $\tilde{Y}$ on Z and the treatment statuses D. (5) we then use the t-statistic associated with covariate D of the last regression as test statistic.

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

$$Y = Z\beta + D\delta + \epsilon,$$

where 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 Z and D.

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

$$M_Z = M_Z' = M_Z M_Z = M_Z' M_Z. \tag{27}$$

The estimated residuals of Y generated by the the regression

$$Y = Z\beta + \epsilon$$

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

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

$$\tilde{Y} = \hat{Y} + B_g \hat{e}. \tag{28}$$

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

$$\tilde{Y} = Z\beta + D\delta + \tilde{\epsilon}. \tag{29}$$

We now examine the δ estimate on Equation (29). This estimate ia actually the same as the one computed by the following regression:

$$M_Z \tilde{Y} = M_Z D\delta + \tilde{\epsilon}. \tag{30}$$

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

$$\hat{\delta}_g = (D' M_Z' M_Z D)^{-1} D' M_Z' M_Z \tilde{Y}. \tag{31}$$

We now use previous equations to transform Equation (31) into a more general formula:

$$
\begin{aligned}
\hat{\delta}_g &= (D'M_Z'M_ZD)^{-1}D'M_Z'M_Z\tilde{Y} && \text{by (31)},\\
&= (D'M_ZDM_ZD'M_Z\tilde{Y} && \text{by (27)},\\
&= (D'M_ZDM_ZD'M_Z(Y + B_g\hat{e}) && \text{by (28)} ,\\
&= (D'M_ZDM_ZD'M_Z((I - M_Z)Y + B_g\hat{e}) && \text{because } M_Z = I - Z(Z'Z)^{-1}Z',\\
&= (D'M_ZDM_ZD'((M_Z - M_Z)Y + M_ZB_g\hat{e}),\\
&= (D'M_ZDM_ZD'(M_ZB_g\hat{e}),\\
&= (D'M_ZDM_ZD'(M_ZB_gM_ZY) && \text{because } \hat{e} = M_ZY. \qquad (32)
\end{aligned}
$$

Kennedy (1995) pointed out that Freedman and Lane (1983) algorithm is summarized by Equation 32. Notationally, we can use $T_Z(Y, gD); g \in \mathscr{G}_X$ (instead of $T(Y, gD); g \in \mathscr{G}_X$) to represent the distribution of the test statistic associated with the t-statsitic of the D covariate in the Freedman and Lane (1983) regression just described. Using this notation, the analysis of the previous sections holds unaltered.

D Additional Baseline Tables

These tables explore differences in pre-program variables between treatment and control groups at each follow up wave. These results are presented in order to highlight that there were no differential patterns of attrition across waves.

Table D.1: 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 Miarjuana (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	0.517
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 D.2: 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 Miarjuana (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.

E Additional Inference Results: Addressing Attrition using Inverse Propensity Weights

One aspect of the NFP that may cause concern is attrition. In order to address this we proceed as follows. We estimate by OLS a linear regression model which dependent variable is binary and equal to one if the mother represents an attrition case. We choose the model as to maximize the R^2 and incorporate pre-program variables with very few missing values. This permits us to establish the set of variables that provides the maximum amount of information when it comes to estimate the probability of attrition. Then, we predict the probability of attrition through a Logit model and apply an inverse probability weighting scheme to our estimations. The results do not change much after this correction. Tables E.4–9 show these results. The tables can be read in the same way as Tables 6–9 in the paper.

Table E.3: Using Logit to Obtain Probabilities

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

Notes: 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 E.4: Child Health Outcomes

	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

Notes: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Cntr. Mean) of each result set shows the mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Cd. Diff. Mn.) gives the conditional difference in means between the treatment group and the control group. The third column (Cd. Eff. Size) calculates the conditional effect size for the respective group. The fourth column (Ass. P-val.) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the difference in means between treatment and control groups. 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 fifth column (Block Perm. FL/Single P-val.) 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 4. 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 Perm. FL/Stepdown) provides p-values that accounts for multiple-hypothesis testing based on the stepdown algorithm of Romano and Wolf (2005a). 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 E.5: Family Environment

	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
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

Notes: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Cntr. Mean) of each result set shows the mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Cd. Diff. Mn.) gives the conditional difference in means between the treatment group and the control group. The third column (Cd. Eff. Size) calculates the conditional effect size for the respective group. The fourth column (Ass. P-val.) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the difference in means between treatment and control groups. 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 fifth column (Block Perm. FL/Single P-val.) 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 4. 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 Perm. FL/Stepdown) provides p-values that accounts for multiple-hypothesis testing based on the stepdown algorithm of Romano and Wolf (2005a). 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 E.6: Cognitive Abilities and Achievement Outcomes

	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**

Notes: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Cntr. Mean) of each result set shows the mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Cd. Diff. Mn.) gives the conditional difference in means between the treatment group and the control group. The third column (Cd. Eff. Size) calculates the conditional effect size for the respective group. The fourth column (Ass. P-val.) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the difference in means between treatment and control groups. 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 fifth column (Block Perm. FL/Single P-val.) 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 4. 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 Perm. FL/Stepdown) provides p-values that accounts for multiple-hypothesis testing based on the stepdown algorithm of Romano and Wolf (2005a). 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 E.7: Socio- Emotional Abilities

	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**

Notes: The first column provides the outcome description. Our results are presented in six columns for each gender. The first column (Cntr. Mean) of each result set shows the mean for the control group. When factor scores were computed, we set the mean in the control group to zero. The second column (Cd. Diff. Mn.) gives the conditional difference in means between the treatment group and the control group. The third column (Cd. Eff. Size) calculates the conditional effect size for the respective group. The fourth column (Ass. P-val.) provides the asymptotic p-value for the one-sided single hypothesis test associated with the t-statistic for the difference in means between treatment and control groups. 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 fifth column (Block Perm. FL/Single P-val.) 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 4. 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 Perm. FL/Stepdown) provides p-values that accounts for multiple-hypothesis testing based on the stepdown algorithm of Romano and Wolf (2005a). 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.

F Theoretical Framework for Mediation Analysis

This section develops a theoretical framework that helps to interpret the estimates of a standard mediation model for early childhood interventions. Our model is based on a technology of skill formation according to (Cunha and Heckman, 2007b). In it, later skills build on previous ones 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 $\mathbf{I}_{i,t}$ represent investments at the same period. We use X_i for family background characteristics and $\upsilon_{i,t}$ for an exogenous error term independent of $\boldsymbol{\theta}_{i,t}, \mathbf{I}_{i,t}$ and X_i. The structural equation that governs the evolution of skills is given by:

$$\boldsymbol{\theta}_{i,t+1} = q_{t+1}(\boldsymbol{\theta}_{i,t}, \mathbf{I}_{i,t+1}, X_i, \upsilon_{i,t+1}); t \in \{0, 1, \ldots, T-1\}. \tag{33}$$

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

$$\mathbf{I}_{i,t+1} = h_{t+1}(\boldsymbol{\theta}_{i,t}, X_i, \varepsilon_{i,t+1}); t \in \{0, 1, \ldots, T-1\}, \tag{34}$$

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

$$Y_i = g_T(\boldsymbol{\theta}_{i,T}, X_i, \xi_{i,T}). \tag{35}$$

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

We can use a recursive substitution of investments and skills of Equations (33)–(34) into (35) to generate the following equation:

$$Y_i = f_{t'}(\boldsymbol{\theta}_{i,t'}, X_i, \{\upsilon_{i,\tilde{t}}\}_{\tilde{t}=t'}^{T}, \{\varepsilon_{i,\tilde{t}}\}_{\tilde{t}=t'}^{T}, \xi_{i,T}), \tag{36}$$

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 (34) at period t' by:

$$\mathbf{I}_{i,t'} = h_{t'}(\boldsymbol{\theta}_{i,t'-1}, D_i, X_i, \varepsilon_{i,t'}); \text{ for some } t' \in \{0, 1, \ldots, T-1\}, \tag{37}$$

The counterfactual values investment $\mathbf{I}_{i,t'}$ are defined by the value $\mathbf{I}_{i,t'}$ 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 (37). That is:

$$\mathbf{I}_{i,t',d} = h_{t'}(\boldsymbol{\theta}_{i,t'-1}, d, X_i, \varepsilon_{i,t'}); d \in \{0,1\} \text{ for some } t' \in \{0,1,\ldots,T-1\}. \tag{38}$$

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

$$\boldsymbol{\theta}_{i,t',d} = q_{t'}(\boldsymbol{\theta}_{i,t'-1}, \mathbf{I}_{i,t',d}, X_i, \upsilon_{i,t'}).$$

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

$$\begin{aligned}
\mathbf{I}_{i,t+1,d} &= h_{t+1}(\boldsymbol{\theta}_{i,t,d}, X_i, \varepsilon_{i,t+1}), \text{ and} \\
\boldsymbol{\theta}_{i,t+1,d} &= q_{t+1}(\boldsymbol{\theta}_{i,t,d}, \mathbf{I}_{i,t+1,d}, X_i, \upsilon_{i,t+1}); t > t'.
\end{aligned}$$

We can also define the counterfactual outcomes by:

$$Y_{i,d} = f_{t'}(\boldsymbol{\theta}_{i,t',d}, X_i, \{\epsilon_{i,\tilde{t}}\}_{\tilde{t}=t'}^{T}, \{\varepsilon_{i,\tilde{t}}\}_{\tilde{t}=t'}^{T}, \xi_{i,T}), \tag{39}$$

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

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

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

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

We now cast on Equation (39) to generate a tractable equation to examine mediation effects. Note that Equation (39) holds ont onlt for t' but for any $t \geq t'$.

$$Y_{i,d} = f_t(\boldsymbol{\theta}_{i,t,d}, X_i, \{\upsilon_{i,\tilde{t}}\}_{\tilde{t}=t}^{T}, \{\varepsilon_{i,\tilde{t}}\}_{\tilde{t}=t}^{T}, \xi_{i,T}), \text{ for any } t \in \{t', t'+1, \ldots, T\}. \tag{40}$$

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 X_i. For sake of notational simplicity, we

can substitute those error terms by ζ_t without loss of generality. Equation (40) then becomes:

$$Y_{i,d} = f_t(\boldsymbol{\theta}_{i,t,d}, X_i, \zeta_{i,t}). \tag{41}$$

We achieve a linear form of Equation (41) by approximating it through a Maclaurin expansion. This generates the following equation:

$$Y_{i,d} = \kappa_t + \boldsymbol{\alpha}_{t,d}\boldsymbol{\theta}_{i,t,d} + \boldsymbol{\beta}_{t,d}X_i + \epsilon_{i,t,d}, \quad d \in \{0,1\}. \tag{42}$$

where $\epsilon_{t,d}$ accounts for the approximation error. Equations (41)–(42) are used in our mediation analysis in Section 5.

G Mediation Methodology

G.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 highlighted in the paper, we perform two sets of analysis. First, we study if 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 if 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.[17] Thus, our linear measurement system looks as follows:[18]

$$M^j_{m^j,d} = \nu^j_{m^j} + \varphi^j_{m^j}\theta^j_d + \eta^j_{m^j,d}, \tag{43}$$

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

[17] This follows the same notation as Heckman et al. (2010)
[18] We control for pre-program variables X but we keep it implicit to shorten notation.

idiosyncratic error term which, by assumption, is independent of $\boldsymbol{\theta}_d^j \; \forall j \in \mathcal{J}$. We normalize the loading factor associated with the first measure of each factor to 1 to set a scale, otherwise the scale is arbitrary.[19] 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: $E[\theta^j(d)] = \mu_d^j$, $Var[\theta_d] = \Sigma_{\theta_d}$, $\nu_{m^j}^j$, $\varphi_{m^j}^j$, $Var[\eta_{m^j}^j]$). Heckman et al. (2010) show that the existence of at least three measures for each latent skill guarantees identification.[20] 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 by 45 items.

The maternal skills selected correspond to anxiety, assessed by the Rand Mental Health Inventory, self-esteem, measured by the Rosemberg scale, and mastery, approximated by the Pearlin scale. Similarly, for age 6, we select as plausible mediators children's skills influenced by the NFP. Children 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 B 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 order to do this, we assume that the latent skills and the error terms, , $\theta^j, \eta_{m^j}^j$, are normal and i.i.d. We use full-information maximum likelihood to deal with the missing values in the measures for some individuals. Recent work by citation shows that 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.[21]

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

[19] 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.

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

[21] Missing at random means that the probability of missing data in a variable x can depend on other observed variables but not on the values of x itself

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 (Bartlett, 1937; Thomson, 1934) and the Bartlett method, which resembles GLS. 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.[22] 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:

$$\boldsymbol{\theta}_{S,i} = \boldsymbol{L}'\boldsymbol{M}_i \tag{44}$$

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:

$$\boldsymbol{\theta}_{S,i} = (\hat{\boldsymbol{\varphi}}'\hat{\boldsymbol{\Omega}}^{-1}\hat{\boldsymbol{\varphi}})^{-1}\hat{\boldsymbol{\varphi}}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{M_i}, \tag{45}$$

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

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

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.[23] 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 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. (2010), can be summarized as follows:

Consider the model following model. To simplify notation, we use W to denote pre-program variables X, treatment indicator and the intercept of equation 10:

$$Y_i = \boldsymbol{\alpha}\boldsymbol{\theta}_i + \boldsymbol{\gamma}\boldsymbol{W}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, N. \tag{46}$$

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,

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

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:

$$Y_i = \boldsymbol{\alpha}\boldsymbol{\theta}_{S,i} + \boldsymbol{\gamma}\boldsymbol{W}_i + \boldsymbol{\epsilon}_i - \alpha\boldsymbol{V}_i. \tag{47}$$

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

The estimation of equation 47 using OLS produces estimates that are biased:

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

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

$$plim \begin{pmatrix} \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} = \underbrace{\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{V},\boldsymbol{V}} & \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{W}} \\ \boldsymbol{\Sigma}_{\boldsymbol{W},\boldsymbol{\theta}} & \boldsymbol{\Sigma}_{\boldsymbol{W},\boldsymbol{W}} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{\theta}} & \boldsymbol{\Sigma}_{\boldsymbol{\theta},\boldsymbol{W}} \\ \boldsymbol{\Sigma}_{\boldsymbol{W},\boldsymbol{\theta}} & \boldsymbol{\Sigma}_{\boldsymbol{W},\boldsymbol{W}} \end{pmatrix}}_{\boldsymbol{A}} \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \end{pmatrix} \tag{48}$$

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:

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

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:

$$p\text{-value} = \frac{1}{B} \sum_{b=1}^{B} 1(t_b^{j,*} > t^j) \text{ with } t^j = \frac{\hat{\alpha}^j}{\hat{\sigma}(\hat{\alpha}^j)} \text{ and } t_b^{j,*} = \frac{(\hat{\alpha}_b^j - \hat{\alpha}^j)}{\hat{\sigma}(\hat{\alpha}_b^j)} \tag{49}$$

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:

$$p\text{-value} = \frac{1}{B} \sum_{b=1}^{B} 1(T_b^{j,*} > T^j) \text{ with } T^j = \frac{\hat{\alpha}^j * E(\theta^j(\widehat{1) - \theta^j}(0))}{\hat{\sigma}(\hat{\alpha}^j * E(\theta^j(\widehat{1) - \theta^j}(0)))} \tag{50}$$

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 G.8 - G.11 shows the parameters of the outcome equations as wells as the decompositions components.

Table G.8: 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 show 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 G.9: 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 show 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 G.10: 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 show 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 G.11: 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 show 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.

H Mediation Specification Test

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 that allows for estimation. We rewrite the equation for potential outcome Y_d as:

$$\begin{aligned}
Y_d &= \kappa_d + \sum_{j \in \mathcal{J}} \alpha_d^j \theta_d^j + \boldsymbol{\beta}_d \boldsymbol{X} + \tilde{\epsilon}_d \\
&= \kappa_d + \underbrace{\sum_{j \in \mathcal{J}_p} \alpha_d^j \theta_d^j}_{\substack{\text{skills that} \\ \text{we measure}}} + \underbrace{\sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha^j \theta_d^j}_{\substack{\text{skills that we} \\ \text{do not measure}}} + \boldsymbol{\beta}_d \boldsymbol{X} + \tilde{\epsilon}_d \\
&= \underbrace{\kappa_d + \sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j \, \mathrm{E}(\theta_d^j)}_{\text{new intercept}} + \underbrace{\sum_{j \in \mathcal{J}_p} \alpha_d^j \theta_d^j}_{\substack{\text{skills that} \\ \text{we measure}}} + \underbrace{\sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j (\theta_d^j - \mathrm{E}(\theta_d^j))}_{\substack{\text{skills that we} \\ \text{do not measure}}} + \boldsymbol{\beta}_d \boldsymbol{X} + \tilde{\epsilon}_d, \\
&= \underbrace{\tau_d}_{\text{new intercept}} + \underbrace{\sum_{j \in \mathcal{J}_p} \alpha_d^j \theta_d^j}_{\substack{\text{skills that} \\ \text{we measure}}} + \boldsymbol{\beta}_d \boldsymbol{X} + \underbrace{\sum_{j \in \mathcal{J} \setminus \mathcal{J}_p} \alpha_d^j (\theta_d^j - \mathrm{E}(\theta_d^j)) + \tilde{\epsilon}_d}_{\text{new error term}}
\end{aligned} \tag{51}$$

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 measures 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:

$$Y_d = \tau_d + \sum_{j \in \mathcal{J}_p} \alpha_d^j \theta_d^j + \boldsymbol{\beta}_d \boldsymbol{X} + \epsilon_d, \tag{52}$$

produces unbiased estimates of parameter $(\alpha_d^j;\ j \in \mathcal{J}_p)$; $d \in \{0,1\}$. Indeed error terms ϵ_d in equation (52) 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

$$Y_0 = \tau_0 + \sum_{j \in \mathcal{J}_p} \alpha^j \theta_0^j + \boldsymbol{\beta}_0 \boldsymbol{X} + \epsilon_0, \tag{53}$$

also produces unbised 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:

$$(\theta_1^j;\ j \in \mathcal{J} \setminus \mathcal{J}_p) \perp\!\!\!\perp (\theta_1^j;\ j \in \mathcal{J}_p)|\boldsymbol{X}, \tag{54}$$

or, alternatively,

$$(\theta_1^j - \theta_0^j;\ j \in \mathcal{J} \setminus \mathcal{J}_p) \perp\!\!\!\perp (\theta_1^j - \theta_0^j;\ j \in \mathcal{J}_p)|\boldsymbol{X}. \tag{55}$$

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 (54)–(55).

While the skill independence assumption in equation (54) 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.

H.1 Skills and the Measurement System

The assumption that the loading factors in the measurement system (equation 43) 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 of the latent skills and not by the map between measures and skills.

Ultimately, we need the decomposition of the treatment effects, (12), 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:

$$H_0 : \boldsymbol{\alpha}_0(\mathrm{E}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)) = \boldsymbol{\alpha}_1(\mathrm{E}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)) \tag{56}$$

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:

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

We would like to test if the contributions to the treatment effects is independent if we use the parameters' from 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 subscript 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 (56).

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

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

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

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

This variance helps us to get the z-statistic:

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

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 (outcome, skill) paired p-values are shown in Tables H.12 and H.13. 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.

H.1.1 Additional Specification Test for the Outcome Equation

In order to clearly interpret the channels through which the NFP affects later outcomes, (5) 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 on the program effect on the 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 H.14 and H.15 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 into interpretable channels.

I 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:[24]

$$\underbrace{E(Y|D=1) - \mathrm{E}(Y|D=0)}_{\text{Treatment Effects}} = \underbrace{(\alpha_1 - \alpha_0)\theta_0}_{\text{Differences unexplained by the skills}} + \underbrace{(\theta_1 - \theta_0)\alpha}_{\text{Explained: differences in skills}} . \quad (57)$$

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 if our decomposition is a plausible specification, we estimate an Oaxaca-Blinder decomposition. The results in Tables I.16 - I.20 evidence that the Oaxaca-Blinder unexplained part that accounts 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.

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

Table H.12: 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: $z = \frac{\hat{\alpha}^0(\bar{\theta}_1^0 - \bar{\theta}_0^0) - \hat{\alpha}^1(\bar{\theta}_1^1 - \bar{\theta}_0^1)}{\sqrt{\text{Var}(\hat{\alpha}^0(\bar{\theta}_1^0 - \bar{\theta}_0^0) - \hat{\alpha}^1(\bar{\theta}_1^1 - \bar{\theta}_0^1))}}$

Table H.13: 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: $z = \frac{\hat{\alpha}^0(\bar{\delta}_1^0 - \bar{\delta}_0^0) - \hat{\alpha}^1(\bar{\delta}_1^1 - \bar{\delta}_0^1)}{\sqrt{\text{Var}(\hat{\alpha}^0(\bar{\delta}_1^0 - \bar{\delta}_0^0) - \hat{\alpha}^1(\bar{\delta}_1^1 - \bar{\delta}_0^1))}}$

Table H.14: 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 H.15: 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 grades 1-5	1.655	**0.034**
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 use alc, mar, tob last 30 days	1.208	0.237
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 I.16: 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 I.17: 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 I.18: 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 I.19: 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 I.20: 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.