Measuring Investment in Human Capital Formation: An Experimental Analysis of Early Life Outcomes*

Orla Doyle (University College Dublin)
Colm Harmon (University of Sydney and IZA)
James J Heckman (University of Chicago, University College Dublin, NBER and IZA)
Caitriona Logue (University College Dublin)
Seong H Moon (University of Chicago)

Abstract

The literature on skill formation and human capital development indicates that early investment in children is an equitable and efficient policy with large returns in adulthood. Yet little is known about the mechanisms involved in producing these long-term effects. This paper presents early evidence on the nature of skill formation based on an experimentally designed, home visiting program in Ireland targeting disadvantaged families - Preparing for Life (or PFL). We examine the impact of investment in utero and up to six months of age on a range of both parental and child outcomes. Using the methodology of Heckman et al. (2010a), permutation testing methods and a stepdown procedure are applied to account for small sample size and the increased likelihood of false discoveries when examining multiple outcomes. The results show that the program impact is concentrated on parental behaviors with little impact on key domains such as birth weight or early child development. This indicates that home visiting programs can be effective at offsetting deficits in parenting skills within a relatively short timeframe, yet continued investment may be required to observe direct effects on child development.

Keywords: Early childhood intervention; randomized control trial; multiple hypotheses; permutation testing.

JEL Classification: C12, C93, J13, J24.

This Version: 2.02, 5th June 2012

* We thank the European Research Council (ERC) for the Advanced Investigator Award to James Heckman, and the TCD/UCD Innovation Academy for their Bursary awarded to Caitriona Logue. The Northside Partnership (through the Irish Government Department of Children and Youth Affairs and The Atlantic Philanthropies) funds the evaluation of the Preparing for Life program. We would like to thank all those who supported this research including the PFL intervention staff and the UCD Geary Institute evaluation team. The UCD Human Research Ethics Committee, the Rotunda Hospital Ethics Committee and the National Maternity Hospital Ethics Committee granted ethical approval for this study.

1. Introduction

Investment in early childhood is increasingly recognized as a key policy mechanism for ameliorating social disadvantage. Evidence from the few experimentally designed programs with long term follow-up demonstrate positive effects into adulthood across multiple domains, including fewer behavioral problems and criminal convictions, lower dependency on welfare, and increased employment (Olds et al., 1998; Heckman et al., 2010b). Cunha and Heckman (2007) presents a model of skill formation demonstrating that early skills facilitate the accumulation of more advanced skills and these higher levels of skills, early in life, make further investment throughout the lifecycle more productive, through a process of dynamic complementarity. These processes form the theoretical basis of why early investment generates high returns in adulthood, yet little is known about the mechanisms involved in producing these long-term effects.

In this paper we present early evidence on the nature of skill formation based on an experimentally designed, home visiting program in Ireland targeting disadvantaged families - Preparing for Life (or PFL). The program begins in utero and continues until age 5 and thus has the potential to influence skill formation during a period when brain development is at its most malleable (Knudsen et al., 2006). Based on a rich and extensive dataset including both child and parental outcomes, we investigate the early outcomes for families participating in the program. This allows us to determine whether treatment effects from targeted intervention programs can manifest early in the lifecycle, and will allow us to identify the mechanisms involved in generating this process.

The paper is structured as follows. Section 2 describes the PFL program and experimental design, including a description of the recruitment and randomization procedure and the data used in our analysis. The econometric framework in presented in Section 3 -

using the methodology of Romano and Wolf (2005) and Heckman et al. (2010a), we apply permutation testing and a stepdown procedure to account for the small sample size and the increased likelihood of false discoveries when examining multiple outcomes. The results from our analyses are provided in Section 4. Finally, in Section 5, we conclude our discussion.

2. Preparing for Life – Program Design and Impact Data

2.1 Description of the Intervention

PFL is a five-year home visiting program in Dublin, Ireland, which was developed to address the problems of socioeconomic disadvantage in a multi-generation, suburban community consisting mainly of low-density welfare provided (or social) housing. The area is classified by the Irish welfare authorities as disadvantaged. Census data from 2006, before the recent Irish economic crisis, show that 62 percent of residents in this community lived in social housing, the unemployment rate was three times the Irish national average at 12 percent, and just five percent of residents had received some form of third level education (Census 2006). Such socioeconomic disadvantage is also evident in the low level of cognitive and non-cognitive skills of children residing in the community. Doyle and McNamara (2011) find that children from this community were rated below the norm at school entry by teachers with respect to their physical health as well as their cognitive and non-cognitive skills.

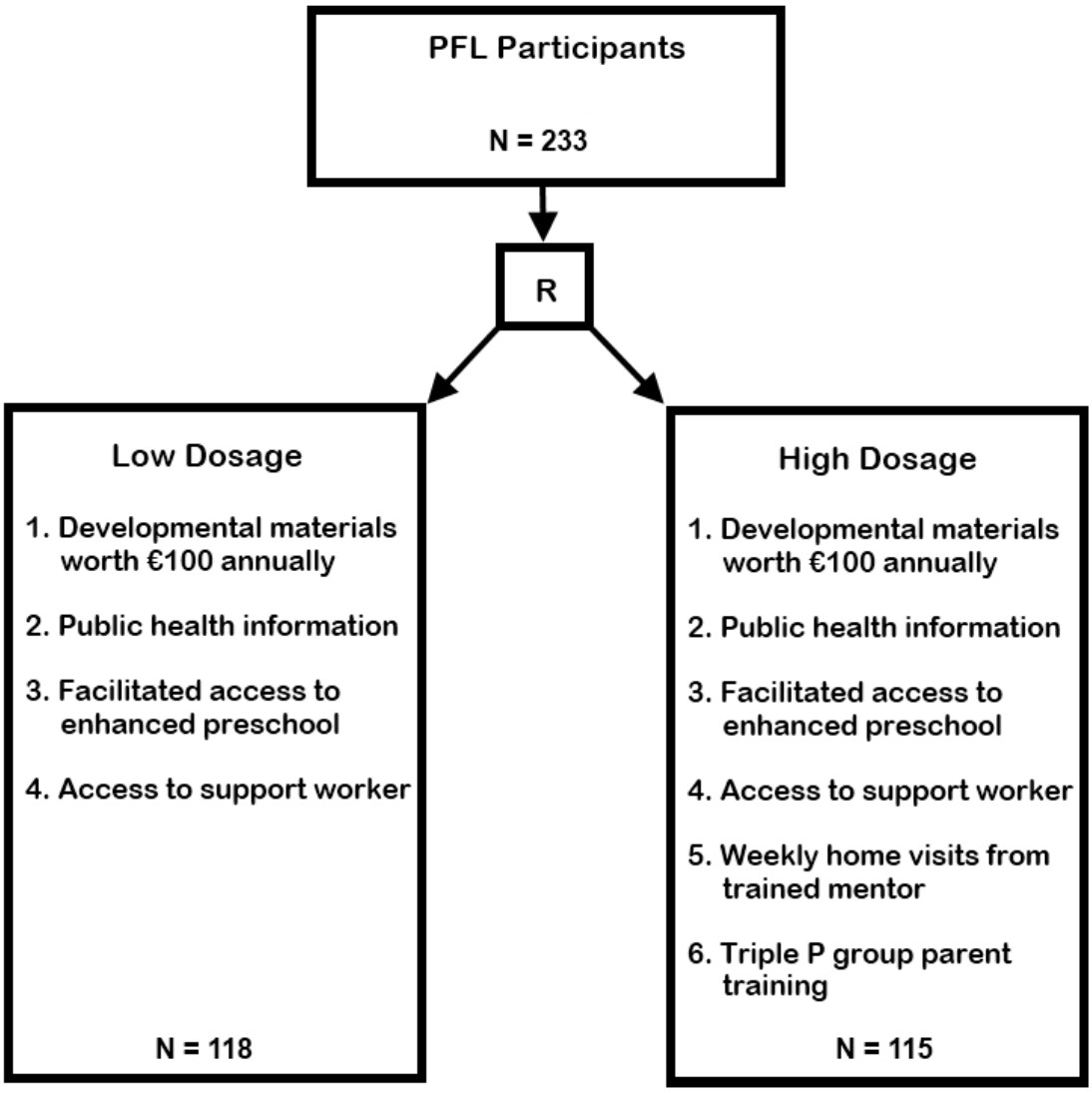
The PFL program was initiated and developed by community representatives and local health and education service providers to improve these documented low levels of school readiness. The intervention begins during pregnancy and continues until the child starts school (a duration of approximately five years). The program is evaluated using a randomized control trial (RCT) based on a dosage experiment whereby all participating families receive some low level of treatment, and a group randomly allocated into the high treatment group receive a much higher intensity treatment. Recruitment into the PFL program took place

between 2008 and 2010. All pregnant women residing in the PFL catchment area were eligible to participate regardless of income or family background. Eligible candidates were identified using hospital records and community referral. After consenting to take part in the program, the participant was assigned to their level of treatment using an unconditional randomization procedure, such that each participant had an equal chance of being allocated to a high or low treatment group. A total of 233 pregnant women consented to participate. This represents a recruitment rate of 52 percent, according to public health nurse records on the number of live births in the community during the recruitment window. Twenty-two percent of potential participants in the area were not identified for recruitment and 26 percent indicated that they did not want to participate in the program[1]. Figure 1 summarizes the program flow.

The evaluation collects data at seven points during program implementation: baseline, six months, 12 months, 24 months, 36 months, 48 months, and school entry. Trained interviewers, who are blinded to the treatment condition, collect data through face-to-face interviews conducted primarily in the participant's home using computer-assisted personal interviewing (CAPI). This paper uses data from the baseline and six month assessment.

[1] Socio-demographic data for these eligible non-participants are not currently available.

Figure 1 Program Evaluation Structure – Preparing for Life



To test the validity of the randomization, a baseline survey was administered to 205 (low = 101; high =104) participants post-randomization yet before they began the program[2]. Table 1 provides a summary of the measures that were tested. One hundred and twenty-three variables were analyzed using permutation testing and no significant differences were found

[2] Note that 19 participants (low=13; high=6) dropped out before the program began, two participants (low= 1; high=1) miscarried before completing the baseline interview, and seven (low = 3; high = 4) missed the baseline interview. An analysis of a subset (N=12) of these early program exits suggests they did not differ on age, education, employment, financial status and support from family and friends, but the sample is too small to draw any formal inference on this group.

between the high treatment and low treatment groups for 119 (97 percent) of the measures. A more detailed discussion of the baseline analysis is available in Doyle et al. (2010).

Table 1: Proportion of Measures Not Significantly Different at Baseline

Category	***PFL* Low – *PFL* High**
Parental Demographics & SES Indicator	33/33
Maternal Well-being & Personality	24/24
Maternal Health & Pregnancy	35/35
Thoughts About Parenting	10/13
Social Support	17/18
Total	119/123 (97%)

On joining the program and completing the baseline interview, all participating families are provided with developmental toys, facilitated access to preschool, and public health information. The participants in the low treatment group have access to an information officer (to provide, for example, details about services in the area and upcoming PFL events), yet they may not receive information on parenting or child development. Participants in the high treatment group receive the additional provision of a home visiting service, where an assigned mentor visits the home up to once a week for between 30 minutes and two hours for the duration of the program. The PFL manual prescribed weekly visits, yet the majority of families engage in fortnightly visits while some engage monthly (Doyle et al., 2011).

The home visiting mentors, from various professional backgrounds, act as advisors to the participating mothers. They have been trained to support and educate parents about child development through structured home visits using "Tip Sheets" - colorful handouts succinctly presenting best-practice information relating to child development which are given to the

participant and serve as an on-going parenting resource[3]. High treatment families also receive group parent training which begins when the child is two years of age. In this paper, we examine child outcomes at six months, thus a comparison of the high and low treatment groups in this paper will focus solely on the impact of the home visiting component. Doyle (2012) discusses the PFL program in more detail.

2.2 Comparison with Existing Home Visiting programs

Family-focused approaches to early intervention have become increasingly popular due to a strong belief that parental outcomes serve a mediating role in child development (Brooks-Gunn et al. 2000). However, as Shonkoff and Phillips (2000) state, changing parenting behavior is difficult. Kahn and Moore (2010) synthesize the findings from 66 home visiting program that were evaluated using experimental designs and found that just one intervention was effective at reducing substance abuse by parents, while seven interventions had a positive and significant effect on parenting practices.

There are many other home visiting programs with a similar program design to PFL yet which differ in terms of their target population, duration, and intensity. Table 2 presents a summary of existing home visiting programs that start during pregnancy and have been evaluated at the six-month stage using a RCT. All programs focus on similar mechanisms that promote child success: educating parents about child development and child health, encouraging a healthy lifestyle, affirming maternal perceptions of self-efficacy in the parenting role, and encouraging positive parenting practices.

The U.S. Nurse Family Partnership (NFP) program bears the closest resemblance to the PFL initiative. Both start during pregnancy and provide participants with home visits from

[3] The Tip Sheets were designed at a reading level of a 12 year-old to make them as accessible as possible. Various Tip Sheets are delivered to participants depending on their child's developmental stage and their family's needs. It is required that all participants must have received the full set of Tip Sheets by the end of the program.

a trained professional who provides parenting education. The trained visitors in both programs discuss factors relating to birth outcomes and child development such as nutrition, substance use, and breastfeeding. Public health nurses deliver NFP while the PFL mentors come from various professional backgrounds including social care, youth studies, psychology, and early childcare and education. Unlike PFL, NFP is only available to primiparous women as first-time mothers are usually considered a higher-risk group with a greater need for an intervention[4]. PFL is also a somewhat more intensive program than NFP in that it works with families up until the child reaches approximately five years of age (as compared with NFP which ends when the program child is 24 months old). In addition, PFL prescribes weekly home visits while the frequency is more staggered in NFP (see Table 2). Finally, PFL also offers participants the additional provisions of social events and baby massage classes to supplement the home visiting element.

None of the studies presented in Table 2 have been evaluated using methods that address sample size limitations. Some studies have the advantage of larger samples (Olds et al. (2002); Kitzman et al. (1997) with NFP; Lee et al. (2009) with Health Families America), while others acknowledge the issue of small samples yet do not adapt their statistical approach (Jungmann et al. (2009) with Pro Kind, LeCroy and Krysik (2011) with Healthy Families America). The problems associated with hypothesis testing of multiple outcomes is largely ignored with the exception of LeCroy and Krysik (2011) who reduce the number of outcome variables, and Culp et al. (2004) where multivariate analysis of variance (MANOVA) methods are used for two outcome clusters. We will return to the comparative aspects of these studies to our PFL outcomes in Section 4.

[4] Harwood et al. (2007) note that expectant mothers are often optimistic about the transition to parenthood. It could be argued that multiparous mothers have more realistic expectations of parenthood and, therefore, may recognize a *greater* need for parenting assistance. Stolk et al. (2008) discuss the possibility that it may be easier to influence the behavior of first-time parents through early intervention.

Table 2: Comparison of Early Childhood Home Visiting Programs which Start During Pregnancy and Follow-Up at Six Months of Age

Program	***Target***	***Duration***	***Visit Frequency***
Preparing for Life	Pregnant women residing in a disadvantaged community	Pregnancy – School Entry	Weekly
Nurse Family Partnership[a,b]	Low income, primiparous, pregnant women. Especially target pregnant teenagers	Pregnancy (1st or 2nd Trimester) – 24 months age	Registration – Month 2: Weekly Month 2 – Child Birth: Biweekly Child Birth – 6 Weeks Age: Weekly 6 Weeks Age – 20 Months Age: Biweekly 20 Months Age – 24 Months Age: Monthly
Pro Kind[c]	Low income, primiparous, pregnant women. Especially targets pregnant teenagers, women with a history of substance use or exposure to violence.	Pregnancy (2nd Trimester) – 24 Months Age	Biweekly
Healthy Families America[a,b]	Pregnant women who have low-income or in their teenage years, and at risk of endorsing child abuse/neglect.	Pregnancy – School Entry	Registration – 6 Months Age: At least weekly 6 Months Age – School Entry: Less frequently
Healthy Start[a]	Pregnant women who are single, or have less than high school qualifications, or Medicaid eligible, or considered "at risk" by health care provider.	Pregnancy – 24 Months Age	Frequency varies by site
Early Intervention Program for Adolescent Mum[a]	Ethnic minority, primiparous, pregnant women aged 14 – 19. Women with serious health issues or drug addiction were excluded.	Pregnancy (1st or 2nd Trimester) – 12 Months Age	Registration – Child Birth: 2 visits Child Birth – 12 Months: 15 visits
Community-based Family Resource Service Programs[d]	Primiparous pregnant women.	Pregnancy (1st or 2nd Trimester) – 36 Months Age	Registration – Month 2: Weekly Month 2 – Child Birth: Biweekly Child Birth – 3 Months Age: Weekly 3 Months Age – 21 Months Age: Biweekly 21 Months Age – 36 Months Age: Monthly

[a] Source: U.S. Department of Health and Human Services (2009). [b] Source: Kahn and Moore (2010). [c] Source: Jungmann et al. (2009). [d] Source: Culp et al. (2004).

2.3 Data and Stylized Facts

2.3.1 Description of Participants

Table 3 provides descriptive statistics for the participants who completed a six-month interview as part of the evaluation[5]. The mothers were 26 years old on average, and 21 weeks pregnant when they joined the program. Approximately 40 percent were employed, some 82 percent had a partner, and almost half were first time mothers. A high proportion indicated that they had a mental health condition (26 percent). With respect to substance use, 50 percent of participants smoked during pregnancy, 26 percent stated that they had drank alcohol at some stage during pregnancy, and 16 percent of respondents indicated that they had used an illegal drug at least once in their lives.

Table 3: Baseline comparison of high/low treatment participants in six month sample

	High Treatment			**Low Treatment**		
Variables	***N***	***Mean***	***SD***	***N***	***Mean***	***SD***
Weeks in pregnancy at program entry	82	21.78	7.83	89	21.18	6.87
Mother's age	82	25.67	5.76	89	25.69	6.04
Partnered	82	0.80	0.40	89	0.83	0.38
Married	82	0.16	0.37	89	0.17	0.38
Living with parent(s)	82	0.55	0.50	89	0.45	0.50
First time mother	82	0.52	0.50	89	0.46	0.50
Low education	82	0.29	0.46	89	0.36	0.48
Mother employed	82	0.43	0.5	89	0.40	0.49
Saves regularly	82	0.50	0.50	89	0.53	0.50
Social housing	82	0.54	0.50	89	0.56	0.50
ŀCognitive resources (WASI)	81	82.49	13.01	88	80.66	13.20

[5] While there was some attrition between the baseline and six-month interviews, no statistically significant differences between the low and high treatment groups were found in terms of socio-demographics, cognitive and noncongitive resources, and health behavior.

	High Treatment			Low Treatment		
Physical Health Condition	82	0.76	0.43	89	0.64	0.48
Mental Health Condition	82	0.27	0.45	89	0.26	0.44
Smoking during pregnancy	82	0.51	0.50	89	0.49	0.50
Alcohol during pregnancy	82	0.25	0.43	89	0.25	0.43
Drug ever used	82	0.01	0.11	89	0.03	0.18
Vulnerable attachment (VASQ)	82	18.05	3.76	89	17.89	4.04
Positive parenting attitudes (AAPI)	82	5.23	1.24	89	5.20	1.38
Self-efficacy (Pearlin)	82	2.82	0.60	89	2.89	0.63
Self-esteem (Rosenberg)	82	13.06	2.60	89	12.75	2.95
Knowledge of infant development (KIDI)	82	72.4	7.10	89	70.51	8.29

The Weschler Abbreviated Scale of Intelligence (WASI) was used to measure cognitive resources. The Vulnerable Attachment Style Questionnaire (VASQ) measures the respondents' interactions and dependence on other people. Scores above 15 are indicative of depressive disorders. The Adult Adolescent Parenting Inventory (AAPI) measures approaches to parenting and provides an indicator of the endorsement of abuse/neglect. Scores range from 1 to 10 with scores below 4 indicating a low risk of abusive/neglect and scores above 8 indicating a high risk of abuse/neglect. The Pearlin Self-Efficacy scale ranges from zero to four with higher scores indicating higher self-efficacy. The Rosenberg scale ranges from zero to 18 with higher scores indicating more maternal self-esteem. The Knowledge of Infant Development (KIDI) score represents the percentage of correct responses to questions relating to child development milestones. Scores range from zero to 100 and higher scores indicate more knowledge of infant development. ł Measured at 3 months postpartum.

The participants have a low level of formal education compared to the national average. Approximately 30 percent indicate that their highest level of education was the Junior Certificate or lower, which is effectively minimum compulsory schooling[6]. This compares with an age-cohort completion rate of high school for comparable females of 74 percent. Thus the drop out rate from high school is almost three times the national average. Using a more refined measure of educational skill, the average level of cognitive resources was 81.54 measured using the *Wechsler Abbreviated Scale of Intelligence* (WASI, Wechsler, 1999), which is below the lower bound on the expected population average range for this measure of between 85 and 115.

[6] The Junior Certificate is an Irish state exam which is completed at 15 to 16 years of age, or after three years of secondary (high) school.

A number of other important psychometric measures are also reported at baseline. Measures of the parent's ability to interact and form attachment with others were obtained using the *Vulnerable Attachment Style Questionnaire* (VASQ; Bifulco et al., 2003). A score above 15 indicates vulnerability for depressive disorders and our sample mean was above this threshold (17.96). Approaches to parenting prior to the intervention were measured using the *Adult Adolescent Parenting Inventory 2* (AAPI-2 ; Bavolek and Keene, 1999) which indicates a parent's tendency towards abuse and neglect. The mean score in the PFL cohort (5.21) falls within the 'normal' range for this scale indicating a moderate to small risk of abusive behavior in this sample. The *Pearlin Self-Efficacy* scale (Pearlin and Schooler, 1978) ranges from zero to four with higher scores indicating the respondent had a stronger feeling of control over her life. The mean score for the PFL sample (2.85) was below the average score found for a representative American sample (3.14; The Panel Study of Income Dynamics). Normative scores are not available for the final two psychometric scales but do allow us to compare the underlying characteristics of the low and high treatment groups. The *Rosenberg* measure is used to compare levels of self-esteem among the participants - scores range from 1 to 18 with higher scores indicating higher levels of self-esteem. The *Knowledge of Infant Development* (KIDI; MacPhee, 1981) shows the percentage of correct responses to questions relating to child development milestones.

To place the PFL cohort in context, we compare our sample with the nationally representative *Growing up in Ireland (GUI) - Nine Month Cohort Study*, which was administered to 11,134 households (or one third of all nine-month old infants living in Ireland) during the period September 2008 to April 2009. The GUI parental sample was noticeably five years older on average then the PFL parents, with education levels in line with expected national averages. Fewer than 11 percent of parents report either a physical or mental health condition. A much smaller proportion of the GUI sample indicated that they

smoked during pregnancy (18 percent compared with 51 percent in PFL), yet the proportion of respondents who drank alcohol during pregnancy was similar to PFL, as was the proportion who reported having a partner (0.88; SD = 0.33). Overall, this comparison highlights that the PFL cohort reflects a relatively disadvantaged community when compared with national averages, with significant differences in self-reported health and objective health behaviors such as smoking, yet there are some similarities such as presence of husband/partner.

3 Econometric Framework

3.1 Estimation Model and Six Month Outcomes

The PFL program is evaluated using an RCT expressed simply as

$$y_i = \alpha + \beta D_i + \varepsilon_i \quad (1)$$

where y_i represents the observed outcome for individual i, and D is a binary variable where a value of one indicates assignment to the high treatment group. We estimate this model for multiple y_i. Specifically, we examine 64 outcome measures related to child development, child health, and parenting. Table 4 summarizes the standardized scales used, while a number of individual non-scale items are also included.

The level of child development at six months of age is measured using three standardized measures: the *Ages and Stage Questionnaire* (ASQ; Squires et al., 1999), the *Ages and Stages Questionnaire: Social-Emotional* (ASQ:SE; Squires et al., 2003) and an assessment of temperament based on the *Infant Characteristics Questionnaire* (Bates et al., 1979). Child health outcomes are measured in three main domains. *Birth Outcomes* relates to birth weight, breathing problems, and health problems that require hospital or GP care. *Mother's Health Decisions for her Infant* relates to parental behavior which directly impacts upon the child's health such as the decision to get her child vaccinated, her feeding choices,

and knowledge of her baby's weight (a proxy for the mother's interest in the health of her child). Finally, we examine *Sleep Routines* - sleep impacts upon child development and parental failure to establish a bedtime routine is cited a major cause of sleep problems in children (Jaffa et al., 1993).

Table 4: Standardized Scales Uses to Measure Child and Adult Outcomes

Domain	***Instrument***	***Scale***	***Higher Scales Indicate***
Child Development	Ages and Stages Questionnaire (ASQ; Squires et al., 1999) Subdomains: *communication*, *gross motor*, *fine motor*, *problem solving*, and *personal-social*	0 – 60	Favorable
	Ages and Stages Questionnaire: Social Emotional (ASQ-SE; Squires et al., 2003)	0 – 285	Unfavorable
	Infant Characteristics Questionnaire (Bates et al., 1979) Difficult temperament	0 – 42	Unfavorable
Parenting	Parental Locus of Control (PLOC; Campis et al., 1986) Subdomains: *parental efficacy*, *parental responsibility*, *child control of parent's life, parental belief in fate,* and *parental control of child's behaviour*	4 – 20	Unfavorable
	Condon Maternal Attachment Scale (CMAS; Condon and Corkindale, 1998) Subdomains: *quality of attachment*, *pleasure in interaction*, and *absence of hostility*	1 – 5	Favorable
	Parenting Stress Index (PSI; Abidin, 1995) Subdomains: *difficult child*, *parenting distress*, and *parent-child dysfunctional interactions*	12 – 60	Unfavorable
	Parental Cognition and Conduct Toward the Infant Scale (PACTOIS; Boivin et al., 2005) Subdomains: *parental self-efficacy*, *perceived parental impact*, and *parental warmth* Subdomains: *parental hostile-reactive behavior* and *parental overprotection*	1 – 11 1 – 11	Favorable Unfavorable
	HOME Observation for Measurement of the Environment (HOME; Caldwell and Bradley, 1999) and The Supplement to the HOME Scale for Impoverished Families (SHIF; Ertem et al., 1996) Subdomains: *responsivity*, *acceptance*, *organization*, *learning materials*, *involvement*, *variety*, *daily routines*, *child care*, *outings*, *toys and books*, *play*, *physical environment*, and *interaction* Subdomain: *restriction*	0 – 1 0 – 1	Favorable Unfavorable

Parenting behavior is examined using six standardized measures: *Parental Locus of Control* (PLOC ; Campis et al., 1986); *Condon Maternal Attachment Scale* (CMAS; Condon and Corkindale, 1998); *Parenting Stress Index* (PSI; Abidin, 1995); *Parental Cognition and Conduct Toward the Infant Scale* (PACTOIS; Boivin et al., 2005); the Infant-Toddler version of the *Home Observation for Measurement of the Environment* (HOME; Caldwell and Bradley, 1999); and the *Supplement to the HOME Scale for Impoverished Families* (SHIF; Ertem et al., 1996). We also examine three simple indicators of whether the parent smoked tobacco, consumed alcohol, or took an illegal drug during pregnancy.

3.2 Permutation Testing

Although the RCT design in (1) is a simple specification, the use of traditional OLS for estimation and *t*-tests for hypothesis testing is not appropriate given the small sample size. Permutation methods do not depend on distributional assumptions and thus facilitate the estimation of treatment effects in small samples (Heckman et al., 2010a).

A permutation test relies on the assumption of exchangeability under the null hypothesis (see Good, 2005). In this paper, the observed *t*-statistic is recorded and compared to the distribution of *t*-statistics that result from multiple, random permutations of the treatment label (1,000 replications are permuted using Monte Carlo simulations in our analyses). Upton (1992) reviews the literature, which shows that the mid-*p*-value is more suitable when dealing with discrete data, therefore we report the right-sided, mid-*p*-value, which is calculated as:

$$MP(T) = P(T^* > T) + 0.5P(T^* > T),$$

where P is the probability distribution, T is the randomly permuted *t*-statistic, and T is the observed *t*-statistic. We use right-sided testing in order to test whether the high treatment is

having a positive effect on child and parenting outcomes. If $p < 0.1$ the effect is statistically significant.

3.3 The Stepdown Procedure

Conducting permutation tests for each of the 64 outcomes increases the likelihood of a Type I error (rejecting a null hypothesis when it is in fact true) and studies of RCTs have been criticized for overstating treatment effects as a result of this 'multiplicity' effect (Pocock et al., 1987). To address this problem, methods have been developed which control the Family-Wise Error Rate (FWER), the probability of rejecting at least one true null hypothesis at a pre-determined level, α. Testing for this involves adjusting the *p*-values associated with individual tests to account for the effect of multiple outcomes.

We employ the stepdown procedure described in Romano and Wolf (2005) and applied in Heckman et al. (2010a). The stepdown procedure involves firstly calculating a test statistic for each null hypothesis in a family of outcomes - we use the *t*-statistic. Using the permutation testing method described above, the appropriate *t*-distribution under the null hypothesis is constructed. The stepdown procedure starts by extracting the largest observed *t*-statistic and comparing it with the distribution of maximum statistics for the joint hypotheses. If the probability of observing this statistic by chance is high ($p \geq 0.1$) we accept the joint hypothesis that the high treatment does not have a statistically significant impact on the family of outcomes being tested..

On the other hand, if the probability of observing this t-statistic is low ($p < 0.1$) we reject the joint null hypothesis and proceed by excluding the most significant hypothesis and testing the subset of hypotheses that remain for joint significance. This process of dropping the most significant hypothesis continues until the resulting subset of hypotheses is accepted, or only one hypothesis remains. 'Stepping down' through the hypotheses in this manner

allows us to isolate the hypotheses that lead to a rejection of the null. This method is superior to the well-known Bonferroni adjustment methods as it accounts for interdependence across the outcomes. The Romano and Wolf (2005) method uses a weaker assumption than other established stepwise methods (Benjamini and Hochberg,1995; Westfall and Wolfinger, 1997) that of monotonicity with respect to the critical values. This ensures that the largest unadjusted p-values correspond to the largest adjusted p-values (Heckman et al., 2010a).

4 Results

4.1 Six Month Analysis

We divide our results into the three main outcome categories: child development, child health, and parenting. We present the mean outcome scores by group, the p-values that result from individual permutation testing, the adjusted *p*-values calculated using the stepdown procedure, and Cohen's *d*-effect sizes[7]. Note that in order to implement the stepdown method, outcomes must be strictly increasing in score such that measures where increasing values are associated with a negative outcome are inverted.

To aid interpretation of the "Stepdown (Adj.)" column, the outcomes within a category are ordered from largest to smallest observed t-statistic. Each adjusted *p*-value represents the likelihood of rejecting the joint null hypothesis when the variables in the rows above are excluded.[8]

[7] Cohen's *d* = mean difference/pooled standard deviation. *d* = 0.2 indicates a small effect, *d* = 0.5 indicates a medium effect, *d* = 0.8 indicates a large effect (Gravetter and Wallnau, 2008)

[8] For example, in Table 5, the first adjusted *p*-value (0.490) in the *ASQ Scores & Difficult Temperament* category is the result of jointly testing all seven outcomes in that category. The next adjusted *p*-value (0.511) is the result of excluding the *ASQ Gross Motor Score* variable from the joint-hypothesis test. The adjusted p-value of 0.712 is the result of excluding both the *ASQ Gross Motor Score* and the *ASQ Communication Score*. Thus, as we step down through the hypotheses, the most statically significant variables are excluded.

4.1.1 Child Development

Table 5 presents the results for the child development outcomes. The results from the individual tests indicate that none of the test statistics are statistically significant at the 10 percent level, and the stepdown procedure fails to reject the null hypothesis of no treatment effect on child development outcomes at the six months stage. In addition, the effect sizes for each outcome are small.

The comparable literature on the impact of interventions on child development at 6 months is limited. Anisfeld et al. (2004) report that the Healthy Families America program has no impact on cognitive development at 6 months of age yet do not report statistical significance. Similarly the German Pro Kind program (Jungmann et al., 2009) does not have an effect on cognitive functioning at the same milestone based on simple t-tests on a small sample (N = 76). Using a Logit model on a relatively large sample (N= 543) Olds et al. (2002) find that the NFP program is effective at reducing emotional vulnerability in response to fear stimuli. Jungmann et al. (2009) find that Pro Kind reduces the presence of symptoms of a difficult temperament at six months of age. By comparison we do not identify a precisely determined treatment effect with respect to the non-cognitive development measures (*Difficult Temperament, ASQ Personal Social Score, ASQ Social-Emotional Score*), yet the mean differences indicate that the high treatment group is performing better on average.

Table 5: Comparison of High and Low Treatment Outcomes: Child Development

Variable	N (N_{HIGH}/ N_{LOW})	M_{HIGH} (SD)	M_{LOW} (SD)	*p*-values Individual Test[1]	*p*-values Stepdown (Adj.)[2]	Effect Size (d)
ASQ Scores & Difficult Temperament						
ASQ Gross Motor Score	173 (83/90)	40.78 (11.93)	38.50 (12.99)	0.124	0.49	0.18
ASQ Communication Score	173 (83/90)	53.07 (7.84)	51.78 (8.49)	0.154	0.511	0.16
ŀDifficult Temperament	173 (83/90)	11.70 (5.71)	12.21 (5.50)	0.278	0.712	0.09
ASQ Personal Social Score	173 (83/90)	46.69 (12.10)	45.94 (13.57)	0.361	0.755	0.06

łASQ Social-Emotional Score	173 (83/90)	14.76 (10.68)	15.17 (13.75)	0.385	0.746	0.03
ASQ Fine Motor Score	173 (83/90)	50.78 (9.48)	51.39 (10.17)	0.671	0.854	0.06
ASQ Problem Solving Score	173 (83/90)	51.87 (9.39)	52.56 (9.92)	0.710	0.710	0.07

Notes: `N' indicates the sample size. `M' indicates the mean. `SD' indicates the standard deviation. [1] one-tailed (right-sided) p-value from an individual permutation test with 1000 replications. [2] one-tailed (right-sided) p-value from a stepdown permutation test with 1000 replications. ** Significant at the 5 percent level * Significant at the 10 percent level. ł Represents a negative outcome and was reverse scored for the stepdown analyses. The sample sizes reported are those used in the individual tests and may differ from the sample size used in the stepdown procedure. For the stepdown procedure, any observations missing data for measures within the stepdown family are dropped. The variables are reported in the order in which they are dropped from the stepdown procedure.

4.1.2 Child Health

To examine child health, we divided the outcomes into three clusters of variables. Table 6 displays twenty-three individual permutation test *p*-values and the results from the three separate stepdown procedures. With respect to the first cluster (*Birth Outcomes*) no significant differences were found between the low and high treatment group. Of particular interest is the lack of impact on birth weight[9]. Participants joined PFL when they were 22 weeks pregnant on average (half way into a normal term pregnancy) therefore the treatment may have started too late in pregnancy to significantly impact upon birth weight through improvements in maternal nutrition, or reducing alcohol consumption and smoking behavior. Moreover, birth weight in our sample is parent-reported and may be susceptible to measurement error[10]. We find no differences in the number of days the child spent in hospital following birth, which differs from studies such as Koniak-Griffin et al. (2000) who show that the Early Intervention Program was effective at reducing the duration of birth related

[9] This is consistent with comparable programs (Stabile and Graham, 2000; Koniak-Griffin et al., 2000); Kitzman et al., 1997; Jungmann et al., 2009). However, some studies have found effects - Lee et al. (2009) using a larger sample (500) from Healthy Families America find a reduced likelihood of low weight births.

[10] One participant in the low treatment group reported giving birth to a baby weighing 5273 grams (much higher than the national average of 3470 grams), whilst one participant in the high treatment group reported giving birth to a baby weighing 1588 grams (the World Health Organization classifies babies weighing less than 2500 grams as low birth weight). If these two participants are excluded, the high treatment group had a higher average birth weight but the difference remains statistically insignificant.

hospitalization, albeit among teenage mothers aged 14-19 which represent a higher risk group than the participants in PFL.

Table 6: Comparison of High and Low Treatment Outcomes: Child Health

Variable	N (N_{HIGH}/ N_{LOW})	M_{HIGH} (SD)	M_{LOW} (SD)	*p*-values Individual Test[1]	*p*-values Stepdown (Adj.)[2]	Effect Size (d)
Birth Outcomes						
ŀAge (in days) left hospital	173 (83/90)	3.23 (7.03)	3.16 (3.72)	0.564	0.965	0.01
Birth weight (grams)	170 (80/90)	3319 (589)	3338 (613)	0.587	0.948	0.03
Good health since birth	173 (83/90)	0.93 (0.26)	0.93 (0.25)	0.576	0.905	0.02
ŀStayed in hospital during first 6 months	173 (83/90)	0.10 (0.30)	0.09 (0.29)	0.591	0.891	0.03
ŀNo. of health problems taken to medical centre	173 (83/90)	1.37 (1.62)	1.28 (1.09)	0.669	0.827	0.07
ŀProblem breathing	173 (83/90)	0.22 (0.41)	0.14 (0.35)	0.910	0.910	0.19
Mothers' Health Decisions for her Infant						
Baby eats appropriate food	173 (83/90)	0.87 (0.34)	0.77 (0.43)	0.013**	0.129	0.26
Necessary immunizations at 4 months	172 (82/90)	0.96 (0.19)	0.88 (0.33)	0.029**	0.149	0.32
Appropriate frequency of eating	173 (83/90)	0.77 (0.42)	0.63 (0.48)	0.023**	0.135	0.30
ŀLeave baby to cry	173 (83/90)	0.41 (0.49)	0.46 (0.50)	0.303	0.885	0.09
Necessary immunizations at 6 months	172 (82/90)	0.35 (0.48)	0.31 (0.47)	0.370	0.877	0.09
Mother breastfed as a baby	171 (81/90)	0.15 (0.36)	0.12 (0.33)	0.400	0.846	0.08
ŀBaby's crying a problem	173 (83/90)	0.12 (0.33)	0.11 (0.32)	0.414	0.797	0.03
Attempted breastfeeding	173 (83/90)	0.24 (0.43)	0.22 (0.42)	0.482	0.743	0.04
Knows baby's weight	173 (83/90)	0.41 (0.49)	0.48 (0.50)	0.807	0.807	0.14
Sleep Routines						
Appropriate sleep preparation	173 (83/90)	0.48 (0.50)	0.39 (0.49)	0.114	0.575	0.19
ŀTime to sleep (>15 mins)	172 (82/90)	0.29 (0.46)	0.33 (0.47)	0.355	0.878	0.09
ŀBaby awakening a problem	173 (83/90)	0.24 (0.43)	0.23 (0.43)	0.515	0.933	0.02
Sleeps more than 8 hrs per night	171 (83/88)	0.76 (0.43)	0.78 (0.41)	0.609	0.958	0.06
Sleeps undisturbed through the	173	0.75	0.77	0.607	0.952	0.05

night	(83/90)	(0.44)	(0.43)			
łDifficulty falling asleep	173 (83/90)	0.45 (0.50)	0.38 (0.49)	0.841	0.994	0.14
Sleeps undisturbed by 3 months	173 (83/90)	0.36 (0.48)	0.46 (0.50)	0.914	0.988	0.19
Appropriateness of sleeping location	173 (83/90)	0.90 (0.30)	0.99 (0.11)	0.998	0.998	0.39

Notes: `N' indicates the sample size. `M' indicates the mean. `SD' indicates the standard deviation. [1] one-tailed (right-sided) p-value from an individual permutation test with 1000 replications. [2] one-tailed (right-sided) p-value from a stepdown permutation test with 1000 replications. ** Significant at the 5 percent level * Significant at the 10 percent level. ł Represents a negative outcome and was reverse scored for the stepdown analyses. The sample sizes reported are those used in the individual tests and may differ from the sample size used in the stepdown procedure. For the stepdown procedure, any observations missing data for measures within the stepdown family are dropped. The variables are reported in the order in which they are dropped from the stepdown procedure.

With respect to our second cluster of variables (*Mothers' Health Decisions for Her Infant*), we fail to reject the joint hypothesis of no treatment effect on all variables in this cluster. However, the first adjusted p-value is close to the 10 percent cutoff and three of the individual permutation tests are significant for immunization and feeding patterns. Using a large sample (N =1950), Guyer et al. (2003) reports that the Healthy Steps program, which targets mothers with newborn children, significantly increased the likelihood that children received appropriate immunizations. No impact was found for comparable programs that start during pregnancy. In regards to the final cluster of variables relating to sleep routines, we find no significant treatment effects.

4.1.3 Parenting

The parenting outcomes are divided into five clusters and the results are presented in Table 7. We find that two clusters, measuring parental stress and the home environment, remain statistically significant after joint hypothesis adjustment. Moving through the list of variables in the parental stress cluster, we find that the *Parent-Child Dysfunction Interactions* subdomain is driving the joint rejection at the 10 percent level. This measure relates to the respondent's perception that their child is a negative element of their life (for example, whether the parent feels rejected/abused by their child, or feels that the child does not meet their expectations).

Table 7: Comparison of High and Low Treatment Outcomes: Parenting

Variable	N (N_{HIGH}/ N_{LOW})	M_{HIGH} (SD)	M_{LOW} (SD)	*p*-values Individual Test[1]	*p*-values Stepdown (Adj.)[2]	Effect Size (d)
Parental Locus of Control (PLOC)						
ŀParental Control of Child's Behavior	173 (83/90)	6.92 (2.82)	7.22 (2.64)	0.273	0.713	0.11
ŀChild Control of Parent's Life	173 (83/90)	8.43 (3.36)	8.74 (3.11)	0.281	0.705	0.10
ŀParental Responsibility	173 (83/90)	12.57 (3.18)	12.86 (3.02)	0.253	0.608	0.09
ŀParental Belief in Fate	173 (83/90)	9.70 (3.65)	9.97 (3.32)	0.330	0.502	0.08
ŀParental Efficacy	173 (83/90)	6.65 (2.43)	6.76 (2.43)	0.399	0.399	0.04
Maternal Attachment (CMAS)						
Quality of Attachment	173 (83/90)	4.69 (0.30)	4.68 (0.37)	0.459	0.772	0.03
Pleasure in Interaction	173 (83/90)	4.33 (0.38)	4.34 (0.43)	0.578	0.805	0.02
Absence of Hostility	173 (83/90)	4.39 (0.53)	4.41 (0.53)	0.613	0.613	0.04
Parenting Stress Inventory (PSI)						
ŀParent-Child Dysfunctional Interactions	173 (83/90)	16.94 (4.81)	18.40 (5.71)	0.041**	0.082*	0.28
ŀDifficult Child	173 (83/90)	19.45 (5.00)	20.19 (5.50)	0.174	0.277	0.14
ŀParental Distress	173 (83/90)	26.02 (7.98)	25.71 (7.47)	0.603	0.603	0.04
Parental Cognition and Conduct Toward the Infant Scale (PACTOIS)						
Baby Comparison Score	173 (83/90)	7.52 (1.92)	7.04 (1.90)	0.047**	0.228	0.26
ŀParental Hostile-Reactive Behavior	173 (83/90)	0.80 (1.13)	1.04 (1.21)	0.077*	0.335	0.20
Parental Self-Efficacy	173 (83/90)	8.80 (1.11)	8.67 (1.24)	0.255	0.665	0.10
Parental Impact	173 (83/90)	7.25 (2.00)	7.07 (2.23)	0.304	0.664	0.08
ŀParental Over-Protection	173 (83/90)	6.18 (2.19)	6.14 (1.99)	0.535	0.835	0.02
Parental Warmth	173 (83/90)	9.18 (1.17)	9.24 (1.27)	0.649	0.649	0.06
Home Observation for Measurement of the Environment (HOME + SHIF)						
Variety	170 (81/89)	3.54 (1.12)	3.10 (1.01)	0.003***	0.058*	0.42
Childcare	170 (81/89)	4.19 (0.59)	3.93 (0.83)	0.007***	0.114	0.36
Toys and Books	170 (81/89)	7.75 (1.75)	7.28 (1.80)	0.041**	0.453	0.27
Physical Environment	169 (80/89)	6.49 (1.17)	6.21 (1.20)	0.058*	0.560	0.23

Learning Materials	168 (81/87)	6.80 (1.66)	6.42 (1.63)	0.067*	0.535	0.23
Daily Routines	170 (81/89)	7.36 (1.40)	7.14 (1.22)	0.148	0.622	0.17
Play	170 (81/89)	7.32 (1.61)	7.05 (1.44)	0.126	0.591	0.18
Responsivity	165 (80/85)	9.09 (1.73)	8.83 (1.95)	0.184	0.636	0.14
Interaction	165 (80/85)	11.62 (2.03)	11.33 (2.48)	0.21	0.678	0.13
Acceptance	168 (80/88)	6.39 (0.60)	6.34 (0.60)	0.295	0.816	0.09
Involvement	168 (81/87)	4.36 (1.30)	4.35 (1.22)	0.483	0.868	0.01
Organization	169 (80/89)	5.57 (0.65)	5.59 (0.68)	0.59	0.888	0.04
Outings	170 (81/89)	4.77 (0.45)	4.80 (0.43)	0.713	0.860	0.07
ł Restriction	169 (81/88)	5.96 (0.16)	5.98 (0.11)	0.869	0.861	0.17

Notes: `N' indicates the sample size. `M' indicates the mean. `SD' indicates the standard deviation. [1] one-tailed (right-sided) p-value from an individual permutation test with 1000 replications. [2] one-tailed (right-sided) p-value from a stepdown permutation test with 1000 replications. *** Significant at the 1 percent level ** Significant at the 5 percent level * Significant at the 10 percent level. ł Represents a negative outcome and was reverse scored for the stepdown analyses. The sample sizes reported are those used in the individual tests and may differ from the sample size used in the stepdown procedure. For the stepdown procedure, any observations missing data for measures within the stepdown family are dropped. The variables are reported in the order in which they are dropped from the stepdown procedure.

For the home observation cluster which measures the quality of the home environment, we can reject the null hypothesis of no impact in this category, indicating a positive treatment effect. Stepping down through the variables reveals that the measure of variety, which relates to the child's frequency of interaction with individuals other than their own mother, is driving the rejection of the joint null. We fail to reject the null hypotheses of no effect on the *Parental Locus of Control*, *Condon Maternal Attachment*, and *Parental Cognition and Conduct Toward the Infant* clusters.

Finally, there is no evidence of a significant treatment effect on participants' smoking or alcohol consumption during pregnancy. However it is worth noting that this is not inconsistent with the literature. Jungmann et al. (2009) found that the Pro Kind program did not exhibit an effect on smoking during pregnancy (measured using a MANOVA method on a small sample (N = 46)), and using a quasi-experimental method on a larger sample (N = 439),

Johnston et al. (2004) indicate that the Healthy Steps program, which targets mothers with newborn babies, has no effect on a participating household's smoking or alcohol consumption at three months postpartum. Although public health initiatives tend to focus on reducing risky behaviors during pregnancy (such as tobacco and alcohol consumption), it appears that this may not be the most effective mechanism for improving child outcomes.

4.2 Robustness and Sensitivity

4.2.1 Attrition analysis

The data analyzed in this study was originally collected for 173 (low = 90; high = 83) participants indicating that 16 percent of the original sample did not complete a six-month interview (low = 11 percent; high =20 percent). While the official drop-out rate was 9 percent (low =6 percent, high = 13 percent), the remaining 7 percent were disengaged with the program at the time of the six month interview.

To test whether such attrition and disengagement is systematically related to individual characteristics and thus introducing bias to the six month results, permutation tests were applied to twenty-one key baseline measures presented in Table 3 to determine whether there were significant differences between the participants who completed a six-month interview and those who did not. The results for the high treatment group are presented in Appendix Table A1. Two significant differences were found indicating that mothers who did not complete a six-month interview (in the high treatment group) were less likely to be employed and had a lower level of self-esteem (measured using the Rosenberg scale). The same analysis was conducted for the low treatment group, presented in Appendix Table A2. Statistically significant differences were found on four measures. Specifically, the participants who did not complete a six month interview in the low treatment group were younger, more

likely to be first time mothers, had lower cognitive resources and had lower scores on the *Knowledge of Infant Development* measure.

While these differences between the attrition and non-attrition sample are a cause for concern, we examined the twenty-one key baseline measures presented in Table 3 to test whether the high and low treatment participants, who remained on the program, differed prior to the intervention. In this restricted sample, we found no statistically significant differences between the low and high treatment groups suggesting that the two groups were equivalent at baseline. In summary, the analysis indicates that attrition is not introducing bias in the final results.

4.2.2 Robustness

For robustness, we also test individual hypotheses using traditional *t* tests and permutation tests using 10,000 replications. The results are similar regardless of the technique used, with all 11 of the significant differences from individual permutation testing identified. The stepdown procedure was also conducted using 10,000 replications and again, the *Parenting Stress Index* and the *Home Observation for Measurement of the Environment* were the only clusters that remained statistically significant after adjusting the *p*-values for strong FWER control, with the *Parent-Child Dysfunctional Interactions* and *Variety* measures again driving the joint rejections in the corresponding families. Finally, all outcomes were treated as one cluster including all 64 variables and a joint hypothesis test was applied using the stepdown method. The joint hypothesis was rejected in the first step[11].

[11] Westfall and Young (1993) recommend against including too many variables in one family as it reduces the statistical power.

5 Summary and Conclusion

This paper investigates the effectiveness of investment from *in utero* to six months of age on key indicators of early skill formation, such as child development and health, and a well-known mechanism for influencing such development – parenting skills. We find that at six months postpartum, few significant treatment effects are identified between the low and high treatment group, yet those that are present are centered on positive impacts with respect to the parenting (specifically the *Parenting Stress Index* and *Home Observation for Measurement of the Environment* categories). These findings are consistent with previous evaluations of home visiting programmes which report limited results at six months (Gomby, Curloss, & Behrman, 1999).

This study indicates that home visiting programs can be an effective means of improving deficits in early parenting skills and the home environment within a relatively short timeframe. Improvement in these more malleable dimensions of skills, which emerge early in the intervention period, may activate a permanent change for the family. In home visiting programs such as PFL, parents are the primary mechanism for change, thus the main avenue by which child skills can develop and grow is via changes in parenting skills and abilities. These new strategies and skills, which have been developed through interactions with family mentors and PFL material, may take time to have an impact on infant behaviour and development. Indeed, the majority of studies that calculate high returns to early childhood investment are based on analyses conducted when the participating children have reached the teenage years or adulthood (Olds et al., 1997b; Heckman et al., 2010a). This study suggests that improvements in early parenting skills may be one such mechanism which accounts for these later findings. As discussed in Section 2, the theory on human skill formation points to a skill multiplier effect (Cunha and Heckman, 2007). The lack of effects on child development

and health at six months is also consistent with the finding that developmental advances and delays are extremely difficult to detect in very young children (Smitsman & Corbetta, 2010).

In addition, the absence of effects on key dimensions of child development may be attributed to dosage and timing. Recall that the average PFL participant began engaging with the program half way into their pregnancy (22 weeks) and had received, on average, 14 home visits at the six-month stage. It is possible that this small window did not allow enough time for the participants to adopt the strategies advised by their mentors as the bond between mentor and participant was still being formed (Ammerman et al. 2006). Barlow et al. (2005) highlight that weak effects at the start of an intervention, as a result of low-intensity delivery, can still serve as the prelude to more engagement as trust levels increase.

From a methodological perspective, a naive evaluation strategy (which examines each outcome measure individually, and calculates the proportion of measures for which a significant difference is found) would indicate a significant effect for 17 percent of measures (11/64) and could therefore be seen as an overall significant effect. Indeed, this strategy of examining the proportion of results which are statistically significant is employed in Kahn and Moore (2010) to define programs that are "found to work"[12]. Similarly, Avellar and Paulsell (2011) note that few of the studies examined as part of the Home Visiting Evidence of Effectiveness (HomVee) review make corrections for multiple outcomes and advise caution when interpreting the significance of the findings presented.

In our analysis, the *p*-values have been adjusted to account for the increased likelihood of a Type I error in a multiple hypotheses setting. This more rigorous method indicates fewer program effects than a naive approach which examines all outcomes separately. However, the

[12] The authors do no define the cutoff they use but suggest that if 4 of 7 or 5 of 9 measures were found to be statistically significant, the program would be defined as "found to work".

small differences we have identified between the low and high treatment groups could potentially result in large returns over time.

Early childhood interventions have received relatively little attention in Europe. Yet given the social, economic, and cultural differences, especially with respect to the social welfare system, it cannot be assumed that the findings from seminal American studies can be extended to European countries[13]. Further analysis of later waves of outcome data will be examined to understand the true effectiveness of home visiting programs in non-US settings.

[13] The PFL project is part of The European Network on Early Childhood Interventions (ENECI) linking researchers conducting the experimental evaluations of early childhood programs in non-U.S. settings.

References

Abidin, R.R., *Parenting Stress Index* (*PSI*) Psychological Assessment Resources, Inc. 1995.

Ammerman, R., J. Stevens, F. Putnam, M. Altaye, J. Hulsmann, H. Lehmkuhl, J. Monroe,

T. Gannon, and J. Van Ginkel, "Predictors of Early Engagement in Home Visitation," Journal of Family Violence, 2006, 21(2), 105-115.

Anisfeld, E., J. Sandy, and N.B. Guterman, "Best Beginnings: A Randomized Controlled Trial of Paraprofessional Home Visiting Program," Technical Report, Columbia University School of Social Work, New York 2004. Report to the Smith Richardson Foundation and New York State Office of Children and Family Services.

Avelar, S., and D. Paulsell, "Lessons Learned from the Home Visiting Evidence of Effectiveness Review", Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human services, Washington DC 2011. Retrieved from http://homvee.acf.hhs.gov/Lessons_Learned.pdf.

Barlow, J., S. Kirkpatrick, S. Stewart-Brown, and H. Davis, "Hard-to-reach or out-of-reach? Reasons why women refuse to take part in early interventions," *Children & Society*, 2005, *19* (3), 199-210.

Bates, J.E., C.B. Freeland, and M.L. Lounsbury, "Measurement of Infant Diffculties," *Child Development*, 1979.

Bavolek, S. J. and R.G. Keene, *Adult-adolescent parenting inventory – AAPI-2: administration and development handbook* Family Development Resources, Incr 1999.

Benjamini, Y. and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1995, *57* (1), 289-300.

Bifulco, A, J. Mahon, J.-H. Kwon, P. M. Moran, and C. Jacobs, "The Vulnerable Attachment Style Questionnaire (VASQ): an interview-based measure of attachment styles that predict depressive disorder," *Psychological Medicine*, 2003, *33* (06), 1099-1110.

Boivin, M., D. Pérusse, G. Dionne, V. Saysset, M. Zoccolillo, G.M. Tarabulsy, N. Tremblay, and R.E. Tremblay, "The genetic-environmental etiology of parents' perceptions and self-assessed behaviours toward their 5-month-old infants in a large twin and singleton sample," *Journal of Child Psychology and Psychiatry*, 2005, *46* (6), 612-630.

Brooks-Gunn, J., L. Berlin, and A. Fuligni, "Early Childhood Intervention Programs: What About the Family?," in J.P. Shonkoff and S.J. Meisels, eds., *Handbook of Early Childhood Intervention*, 3rd ed., New York: Cambridge University Press, 2000.

Caldwell, B.M. and R.H. Bradley, *HOME Inventory Administration Manual* University of Arkansas 1999.

Campis, L.K., R.D. Lyman, and S Prentics-Dunn, "The Parental Locus of Control Scale: Development and Validation," *Journal of Clinical Child Psychology*, 1986, *15*, 260-267.

Condon, J.T. and C.J. Corkindale, "The Assessment of Parent-to-Infant Attachment: Development of a self-report questionnaire instrument," *Journal of Reproductive and Infant Psychology*, 1998, *16*, 57-77.

Culp, A. McDonald, R. E. Culp, T. Hechtner-Galvin, C.S. Howell, T. Saathoff-Wells, and P. Marr, "First-time mothers in home visitation services utilizing child development specialists," *Infant Mental Health Journal*, 2004, *25* (1), 1-15.

Cunha, F. and J.J. Heckman, "The Technology of Skill Formation," *American Economic Review*, 2007, *97* (2), 31-47.

Doyle, O., "Breaking the Cycle of Deprivation: An Experimental Evaluation of an Early Childhood Intervention," 2012. Forthcoming, *Journal of the Statistical and Social Inquiry Society of Ireland*.

— and K.A. McNamara, "Report on children's profile at school entry 2008-2011: Evaluation of the Preparing for Life early childhood intervention programme," 2011. UCD Geary Institute Working Paper Series, 201108. Retrieved from http://www.ucd.ie/geary/static/publications/workingpapers/gearywp201108.pdf.

— , C.P. Harmon, J.J. Heckman, and R.E. Tremblay, "Investing in early human development: Timing and economic efficiency," *Economics & Human Biology*, 2009, *7* (1), 1-6.

— and K.A. McNamara, C. Cheevers, S. Finnegan, C. Logue, and L. McEntee, "Preparing for Life early childhood intervention impact evaluation report 1: Recruitment and baseline characteristics," 2010. UCD Geary Institute Working Paper Series, 201050. Retrieved from http://www.ucd.ie/geary/static/publications/workingpapers/gearywp201050.pdf.

— , T. Saias, F. Tubach, and T. Brand, "Engagement in Experimental Home Visiting Interventions: A cross-national study on the attributes of parental engagement," 2011. Under submission at *American Journal of Community Psychology*.

Ertem, I.O., A.J. Avni-Singer, and B.W.C. Forsyth, *Supplement to the HOME Scale for Impoverished Families* Yale University 1996.

Gomby, D.S., P.L. Curloss, and R.E. Beherman, "Home visiting: Recent program evaluations: Analysis and recommendations," *The Future of Children,* 1999, *9,* 4-26.

Good, P., *Permutation, Parametric and Bootstrap Tests of Hypotheses*, 3rd ed., New York: Springer, 2005.

Gravetter, F.J and L.B. Wallnau, *Statistics for the Behavioral Sciences*, 8th ed., Belmont, CA: Wadsworth Publishing Co Inc;, 2008.

Guyer, B., M. Barth, D. Bishai, M. Caughy, B. Clark, C. Burkom, and C. Tang, "The Healthy Steps for Young Children Program National Evaluation," Technical Report, Johns Hopkins Bloomberg School of Public Health, Baltimore 2003.

Harwood, K., N. McLean, and K. Durkin, "First-time mothers' expectations of parenthood: What happens when optimistic expectations are not matched by later experiences?," *Developmental Psychology*, 2007, *43* (1), 1-12.

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

—, —, —, —, and —. "The rate of return to the HighScope Perry Preschool Program," *Journal of Public Economics*, 2010, *94* (1-2), 114-128.

Jaffa, T., S. Scott, J.H. Hendriks, and C.M Shapiro, "Sleep Disorders in Children," *British Medical Journal*, 1993, *306*, 640-643.

Johnston, B.D., C.E. Huebner, L.T. Tyll, W.E. Barlow, and R.S. Thompson, "Expanding developmental and behavioral services for newborns in primary care: Effects on parental well-being, practice, and satisfaction," *American Journal of Preventive Medicine*, 2004, *26* (4), 356-366.

Jungmann, T., Y. Ziert, V. Kurtz, and T. Brand, "Preventing adverse developmental outcomes and early onset conduct problems through prenatal and infancy home visitation: The German pilot project `Pro Kind'," *European Journal of Developmental Science*, 2009, 3 (3), 292-298.

Kahn, J. and K.A Moore, "What works for home visiting programs: lessons from experimental evaluations of programs and interventions", 2010. Retrieved from: www.childtrends.org/Files/Child_Trends-2010_7_1_FS_WWHomeVisitpdf.pdf.

Kitzman, H., D.L. Olds, C.R. Henderson, C. Hanks, R. Cole, R. Tatelbaum, K.M. McConnochie, K. Sidora, D.W. Luckey, D. Shaver, K. Engelhardt, D. James, and K. Barnard, "Effect of Prenatal and Infancy Home Visitation by Nurses on Pregnancy Outcomes, Childhood Injuries, and Repeated Childbearing," *JAMA: The Journal of the American Medical Association*, 1997, *278* (8), 644-652.

Knudsen, E.I., J.J. Heckman, J. Cameron, and J.P. Shonkoff, "Economic, Neurobiological, and Behavioral Perspectives on Building America's Future Workforce," *Proceedings of the National Academy of Sciences*, 2006, *103* (27), 10155-10162.

Koniak-Griffin, D., Nancy L.R. Anderson, and M-L Verzemnieks I. nad Brecht, "A Public Health Nursing Early Intervention Program for Adolescent Mothers: Outcomes From Pregnancy Through 6 Weeks Postpartum," *Nursing Research*, 2000, 49 (3), 130-138.

LeCroy, C.W. and J. Krysik, "Randomized trial of the healthy families Arizona home visiting program," *Children and Youth Services Review*, 2011, *33* (10), 1761-1766.

Lee, E., S.D. Mitchell-Herzfeld, A.A. Lowenfels, R. Greene, V. Dorabawila, and K.A. DuMont, "Reducing Low Birth Weight Through Home Visitation: A Randomized Controlled Trial," American Journal of Preventive Medicine, 2009, *36* (2), 154-160.

MacPhee, D., *Knowledge of Infant Development Inventory* University of North Carolina, Department of Psychology 1981. Unpublished questionnaire and manual.

Olds, D.L., C.R. Jr. Henderson, R. Cole, J. Eckenrode, H. Kitzman, and D. Luckey, "Long-term effects of nurse home visitation on children's criminal and antisocial behavior: 15-year follow-up of a randomized trial," *Journal of the American Medical Association*, 1998, p. 1238-1244.

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

— , J. Eckenrode, C.R Henderson Jr, H. Kitzman, J. Powers, R. Cole, K. Sidora, P. Morris, L.M. Pettitt, and D. Luckey, "Long-term Effects of Home Visitation on Maternal Life Course and Child Abuse and Neglect," *Journal of the American Medical Association*, 1997, *278* (8), 637-643.

— , J. Robinson, R. O'Brien, D.W. Luckey, L.M. Pettitt, C.R. Henderson, R.K. Ng, K.L. Sheff, J. Korfmacher, S. Hiatt, and A. Talmi, "Home Visiting by Paraprofessionals and by Nurses: A Randomized, Controlled Trial," *Pediatrics*, 2002, *110* (3), 486-496.

Pearlin, L. I. and C. Schooler, "The Structure of Coping," *Journal of Health and Social Behaviour*, 1978, 2-21.

Pocock, S.J., M.D. Hughes, and R.J. Lee, "Statistical Problems in the Reporting of Clinical Trials," *New England Journal of Medicine*, 1987, *317* (7), 426-432.

Romano, Joseph P. and Michael Wolf, "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing", *Journal of the American Statistical Association*, 2005, *100* (469), 94-108.

Shonkoff, J.P. and D.A. Phillips, *From Neurons to Neighborhoods: The Science of Early Childhood Development*, Washington, DC: National Academies Press, 2000.

Smitsman, A.W., and D. Corbetta, "Action in Infancy- Perspectives, Concepts and Challenges". In J. G. Bremner and T. D. Wachs, eds., *The Wiley-Blackwell Handbook of Infant Development, Volume 1*, 2nd ed., 167-203, West Sussex: Blackwell Publishing, 2010.

Squires, J., D. Bricker, and E. Twombly, *The ASQ:SE user's guide* Brookes Publishing 2003.

Squires, J.K., L. Potter, and D. Bricker, *The Ages and Stages Questionnaire user's guide* Brookes Publishing 1999.

Stabile, I. and M. Graham, "Florida Panhandle Healthy Start: A randomized trial of prenatal home visitation," Technical Report, FSU Center for Prevention and Early Intervention Policy, Tallahassee, FL 2000. Report to the Smith Richardson Foundation and New York State Office of Children and Family Services.

Stolk, M., J. Mesman, J. van Zeijl, L. Alink, M. Bakermans-Kranenburg, M. van IJzendoorn, F. Juffer, and H. Koot, "Early Parenting Intervention: Family Risk and First-time Parenting Related to Intervention Effectiveness," *Journal of Child and Family Studies*, 2008, *17*, 55-83.

The Panel Study of Income Dynamics: Child Development Supplement. User Guide Supplement for CDS-I, 2010. Retrieved from: psidonline.isr.umich.edu/CDS/CDS1_UGSupp.pdf

Upton, G.J.G., "Fisher's Exact Test," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 1992, *155* (3), 395-402.

U.S. Department of Health and Human Services, "Home Visiting Evidence of Effectiveness (HomVee)," 2009. Home Visiting Research Database. Retrieved from http://homvee.acf.hhs.gov/programs.aspx.

Wechsler, D., *Wechsler Abbreviated Scale of Intelligence (WASI)* The Psychological Corporation 1999.

Westfall, P.H. and R.D. Wolfinger, "Multiple Tests with Discrete Distributions", *The American Statistician*, 1997, *51* (1), 3-8.

— , and S.S. Young, *Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment*, 3rd ed., New York: John Wiley & Sons, 1993.

Appendix

Table A1: Comparison of Attritors and Non-Attritors: High Treatment Group

	Attrition			**Non-Attrition**			
Variables	N	Mean	SD	N	Mean	SD	Individual *p*-value
Weeks in pregnancy at program entry	22	20.86	8.04	82	21.78	7.83	0.651
Mother's age	22	24.68	6.23	82	25.67	5.76	0.505
Partnered	22	0.68	0.48	82	0.80	0.40	0.410
Married	22	0.09	0.29	82	0.16	0.37	0.541
Living with parent(s)	22	0.64	0.49	82	0.55	0.5	0.477
First time mother	22	0.59	0.5	82	0.52	0.5	0.629
Low education	22	0.50	0.51	82	0.29	0.46	0.121
Mother employed	22	0.14	0.35	82	0.43	0.50	0.010**
Saves regularly	22	0.36	0.49	82	0.50	0.50	0.337
Social housing	21	0.62	0.50	82	0.54	0.50	0.639
łCognitive resources (WASI)	9	80.67	8.49	81	82.49	13.01	0.570
Physical Health Condition	22	0.73	0.46	82	0.76	0.43	1.000
Mental Health Condition	22	0.32	0.48	82	0.27	0.45	0.760
Smoking during pregnancy	22	0.50	0.51	82	0.51	0.50	1.000
Alcohol during pregnancy	22	0.18	0.39	82	0.27	0.45	0.439
Drug ever used	22	0.05	0.21	82	0.16	0.37	0.159
Vulnerable attachment (VASQ)	22	18.95	3.77	82	18.05	3.76	0.324
Positive parenting attitudes (AAPI)	22	117.91	16.82	82	120.13	12.92	0.565
Self-efficacy (Pearlin)	22	2.62	0.74	82	2.82	0.60	0.263
Self-esteem (Rosenberg)	22	11.91	2.89	82	13.06	2.60	0.096*
Knowledge of infant development (KIDI)	22	71.69	9.39	82	72.4	7.1	0.731

ł Measured at 3 months postpartum. * Significant at the 10 percent level. ** Significant at the 5 percent level. p-values were obtained using permutation based, two-sided t-tests with 1,000 replications.

Table A2: Comparison of Attritors and Non-Attritors: LowTreatment Group

	Attrition			**Non-Attrition**			
Variables	N	Mean	SD	N	Mean	SD	Individual *p*-value
Weeks in pregnancy at program entry	12	22.50	7.70	89	21.18	6.87	0.563
Mother's age	12	22.42	4.94	89	25.69	6.04	0.063*
Partnered	12	0.92	0.29	89	0.83	0.38	0.503
Married	12	0.25	0.45	89	0.17	0.38	0.699
Living with parent(s)	12	0.58	0.51	89	0.45	0.50	0.556
First time mother	12	0.75	0.45	89	0.46	0.50	0.073*
Low education	12	0.67	0.49	89	0.36	0.48	0.131
Mother employed	12	0.33	0.49	89	0.4	0.49	0.772
Saves regularly	12	0.42	0.51	89	0.53	0.50	0.545
Social housing	12	0.50	0.52	89	0.56	0.5	0.759
ǂCognitive resources (WASI)	3	72.67	3.79	88	80.66	13.2	0.097*
Physical Health Condition	12	0.50	0.52	89	0.64	0.48	0.511
Mental Health Condition	12	0.08	0.29	89	0.26	0.44	0.197
Smoking during pregnancy	12	0.33	0.49	89	0.49	0.5	0.367
Alcohol during pregnancy	12	0.42	0.51	89	0.25	0.43	0.308
Drug ever used	12	0.08	0.29	89	0.16	0.37	0.675
Vulnerable attachment (VASQ)	12	17.33	3.63	89	17.89	4.04	0.622
Positive parenting attitudes (AAPI)	11	115.36	15.91	89	117.02	13.55	0.757
Self-efficacy (Pearlin)	12	2.79	0.36	89	2.89	0.63	0.389
Self-esteem (Rosenberg)	12	13.00	2.13	89	12.75	2.95	0.717
Knowledge of infant development (KIDI)	12	64.64	5.21	89	70.51	8.29	0.003**

ǂ Measured at 3 months postpartum. * Significant at the 10 percent level. ** Significant at the 5 percent level. p-values were obtained using permutation based, two-sided t-tests with 1,000 replications.