HEDG
HEALTH, ECONOMETRICS AND DATA GROUP
THE UNIVERSITY of York

WP 13/18

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

Orla Doyle

August 2013

york.ac.uk/res/herc/hedgwp

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 Hyeok Moon (University of Chicago)

Abstract

The literature on skill formation and human capital development clearly demonstrates 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, five-year home visiting program in Ireland targeting disadvantaged families - Preparing for Life (PFL). We examine the impact of investment between utero to 18 months of age on a range of parental and child outcomes. Using the methodology of Heckman et al. (2010a), permutation testing methods and a stepdown procedure are applied to account for the 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 and the home environment, with little impact on child development at this early stage. 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. While correcting for attrition bias leads to some changes in the precision of estimates, overall the results are quite similar.

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

JEL Classification: C12, C93, J13, J24.

This Version: 8th May, 2013

* We thank the European Research Council (ERC) for the Advanced Investigator Award to James J. 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. Helpful comments from seminar participants at ZEW Mannheim are gratefully acknowledged. The usual disclaimer applies.

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 implemented in childhood, but with long term follow-up, suggests 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) present 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 evidence on the nature of skill formation in the early years based on an experimentally designed, home visiting program in Ireland targeting disadvantaged families known as Preparing for Life (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 (Nelson, 2000; Knudsen et al., 2006). Based on a rich and extensive data set including both child and parental outcomes, we investigate the early impact of the program on participating families. This allows us to determine whether treatment effects from targeted intervention programs manifest early in the lifecycle, and to identify the mechanisms involved in generating this process.

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. This is

important as adopting a naive evaluation strategy (which examines each outcome individually, and calculates the proportion of outcomes for which a significant difference is found) would result in a higher number of significant treatment effects than a more conservative approach that accounts for the testing of multiple hypotheses simultaneously. Specifically, when we adopt a naïve approach we find a significant effect for 25 percent of outcomes (6/24) at six months, 7 percent of outcomes (1/14) at 12 months, and 16 percent of outcomes (5/30) at 18 months. While this could be interpreted as an overall positive program effect, when a more rigorous method is applied, such that the *p*-values are adjusted to account for the increased likelihood of a Type I error in a multiple hypotheses setting, we find significantly fewer program effects. The results using this more rigorous approach indicate that the program effects are concentrated on parenting outcomes, while the joint null hypothesis of no effect on child development outcomes fails to be rejected at six months, 12 months and 18 months.

The paper is structured as follows. Section 2 reviews the early-years findings from studies of other home visiting programs on child development and parenting. Section 3 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 4. The results are provided in Section 5 and Section 6 concludes our discussion.

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). We document the evidence on the impact of a range of home visiting programs on child development and parenting outcomes in Table A1 in Appendix A. The primary source of information for this review was the Home Visiting Evidence of Effectiveness website (`http://homvee.acf.hhs.gov`, U.S. Department of Health and

Human Services, 2009)[1]. We limited our assessment to programs that have conducted follow-up assessments before and up to 18 months of age. 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. We also limited our scope to results from studies that were rated as either 'High' (random assignment studies with low attrition of sample members and no reassignment of sample members after the original random assignment; and single case and regression discontinuity designs that meet the What Works Clearinghouse (WWC) design standards) or 'Moderate' (random assignment studies that due to flaws in the study design or analysis (e.g. high sample attrition) do not meet the criteria for the high rating; matched comparison group designs; and single case and regression discontinuity designs that meet WWC design standards with reservations). In addition, we conducted an extensive literature search according to the criteria outlined by HomVee and added any additional relevant studies.

Overall, there is little evidence of a treatment effect on child development outcomes as early as 18 months. The results for parenting outcomes are more mixed. Many studies find a positive effect of home visiting programs on the quality of the home environment up to 18 months (Culp et al. (2004) with the Community-based Family Resource Service Programs; Wagner et al. (1996) with Parents as Teachers, Black et al. (1994) with a home intervention for drug abusing mothers, Mackenzie et al. (2004) with Starting Well; and Larson (1980) with Pre/Post-Natal Home Visiting). However, other studies find no effect of home visiting on the quality of the home environment (Mitchell-Herzfeld et al. (2005) and Duggan, McFarlane et

[1] This site was launched by the U.S. Department of Health and Human Services to conduct a thorough and transparent review of the home visiting research literature and provide an assessment of the evidence of effectiveness for home visiting programme models that target families with pregnant women and children from birth to age five. Trained reviewers evaluated randomised controlled trials and quasi-experimental designs for each model and authors were given the opportunity to respond to missing information.

al. (2005) with Healthy Families America; Shute and Judge (2005) with Starting Well; and Caughy et al. (2004) with Healthy Steps).

None of these studies have been evaluated using methods that address sample size limitations. Some studies have the advantage of larger samples (Olds et al. (2002) and 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 and LeCroy and Krysik (2011) with Healthy Families America). The problems associated with hypothesis testing of multiple outcomes are largely ignored in this literature 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. This study, which addresses these methodological issues, will thus enhance the home visiting literature.

3. Preparing for Life – Program Design and Impact Data

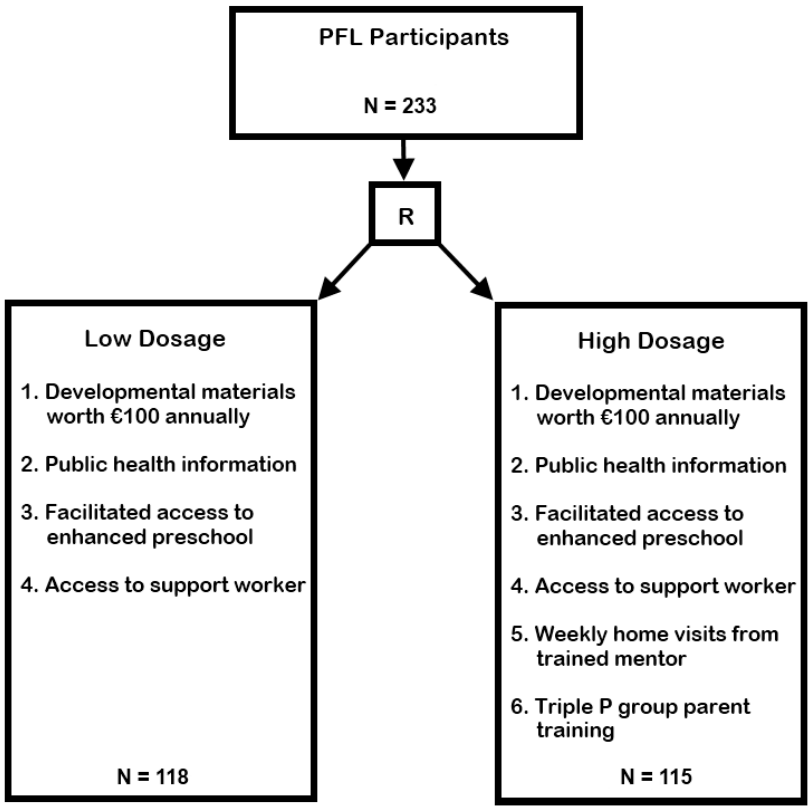
3.1 Description of the Intervention

PFL is a five-year program that was developed to address the problems of socioeconomic disadvantage in a multi-generation, suburban community classified by welfare authorities as disadvantaged and consisting mainly of low-density welfare provided (or social) housing and in Dublin, Ireland.[2] The program was initiated and developed by community representatives and local health and education service providers to improve on the documented low levels of

[2] Census data from 2006, before the recent Irish economic crisis, show that 62 percent of residents lived in social housing (the national average was 9 percent), the unemployment rate was three times the Irish national average at 12 percent, and just five percent of residents had received postsecondary education while the national average was 29 percent. (Census 2006).

school readiness in the catchment area.[3] The intervention begins during pregnancy and will continue until the child starts school at age 4/5. The program was available to all pregnant women residing in the community and participation was voluntary. The program is being evaluated using a randomized control trial (RCT) design in which all families who consented to take part are randomly assigned to either a low level of treatment or a high level of treatment. Figure 1 describes the structure of the program.

Figure 1 Program Evaluation Structure – Preparing for Life



All participating families are provided with developmental toys for each year they are in the program, facilitated access to preschool in the year before starting school, and are

[3] Doyle and McNamara (2011) find that children from this community were rated below the norm (Canadian) at school entry by teachers across all five domains on the Short Early Development Instrument (S-EDI; Janus, Duku, & Stat, 2005). These domains relate to children's *physical health and wellbeing*, *social competence*, *emotional maturity*, *language and cognitive development*, *communication and general knowledge*. Table B1 in Appendix B presents the results from this analysis.

encouraged to attend public health workshops on stress control and healthy eating. The participants in the low treatment group also have access to an information officer (to provide, for example, details about public "services as usual" in the area, such as housing services and childcare services), yet they may not receive any information on parenting or child development. Participants in the high treatment group receive the additional provision of a home visiting service for five years, whereby 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 originally prescribed weekly visits, yet the majority of families received fortnightly visits while some only participated in monthly visits (Doyle, 2013). Thus, PFL is an intention-to-treat program as the actual dosage received by each participant may be less than prescribed. The home visits start in the prenatal period, as soon as the participant joins the program, and finish when the children enter school at age 4/5.

The home visiting mentors, from various professional backgrounds, act as advisors to the participating families. 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[4]. High treatment families also receive group parent training using the Triple P Positive Parenting programme (Sanders, Markie-Dadds, and Turner, 2003) which begins when the PFL child is two years of age. As this paper examines child outcomes up to 18 months of age, our comparison of the high and low treatment groups will focus solely on the impact of the home visiting component. A full

[4] 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. Two examples of the Tip Sheets are presented in Appendix B3.

description of the PFL curriculum is available in Appendix B2 and Doyle (2013) discusses the PFL program and evaluation design in greater detail.

3.2 PFL Recruitment & Randomization

Recruitment into the PFL program took place between 2008 and 2010. All pregnant women residing in the PFL study area were eligible to participate regardless of income or family background. Eligible candidates were identified using hospital records and community referral. After voluntarily consenting to take part in the program, participants were assigned to their level of treatment using an unconditional randomization procedure[5]. Each participant had an equal chance of being allocated to the high or low treatment group. A total of 233 pregnant women consented to participate. This represents a recruitment rate of 52 percent based on 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[6].

To test the validity of the randomization procedure, a baseline survey was administered to 205 (low = 101; high =104) participants post-randomization, yet before

[5] *PFL* participants were randomised after informed consent was obtained. To ensure randomisation was not compromised an unconditional probability computerized randomisation procedure was used whereby the participant pressed a key on a computer which randomly allocated her treatment group assignment. Once assignment was completed, an automatic email was generated which included the participant's unique ID number and assignment condition. This email was automatically sent to the *PFL* programme manager and the evaluation manager. This ensures the recruiter had no influence on the treatment assignment given the evidence that the experimental design in some of the most influential early childhood interventions from the US (such as the Perry Preschool Program) were compromised (Heckman et al., 2010). If there were any attempts to reassign participants from one group to another group, by either directly changing the database or repeating the randomisation procedure, a second email would automatically highlight this intentional subversion.

[6] Socio-demographic data for these eligible non-participants are not currently available. However, data collection on this group is on-going. Specifically, we are conducting direct cognitive assessments with all non-participating children when they are 4 years old. In addition, participant parents are asked to complete a questionnaire on their socio-demographics when they were pregnant with the non-participating child and when their current demographics. This will allow us to determine whether the non-participants differ with respect to their socio-demographics compared to the PFL participants, addressing the selection issue.

treatment begins.[7] 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 between the high treatment and low treatment groups for 119 (97 percent) of the measures, indicating the randomization process was successful. Full descriptive tables including all the measures included in the baseline analysis are available in Appendix C and a more detailed discussion of the baseline analysis is available in Doyle (2013).

Table 1: Proportion of Measures Not Significantly Different at Baseline

Category	***PFL* Low – *PFL* High**
Parental Demographics & SES Indicators	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%)

The evaluation collects data at eight points during program implementation: baseline, six months, 12 months, 18 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. This paper uses data from the baseline, six month, 12 month, and 18 month assessments.

[7] 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, however the sample is too small to make any formal inference on this group.

3.3 Stylized Facts and Participant Profiles

3.3.1 Description of Participants

Table 2 provides baseline descriptive statistics for the estimation sample available at each outcome wave[8]. The participating mothers were 26 years old on average, and 21 weeks pregnant when they joined the program. Approximately 40 percent were employed, over 80 percent had a partner, and almost half were first time mothers. A high proportion indicated that they had a mental health condition (approximately 26 percent). With respect to substance use, one half of participants smoked during pregnancy, just over a quarter drank alcohol at some stage during pregnancy, and just 1 percent of respondents indicated that they had used an illegal drug during pregnancy.

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 (an Irish statewide examination which is completed at 15 to 16 years of age following approximately three years of secondary (high) school) or lower, which is effectively minimum compulsory schooling. This compares with an age-cohort completion rate of high school for comparable females of 74 percent. Thus, the dropout rates from high school are almost three times the national average. Using a more refined measure of cognitive capacities, the average level of cognitive resources was approximately 82 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.

[8] Note that although the sample size for the high treatment group is 82 at both six months and 12 months, the composition of the sample is not identical at each point as individuals who missed a survey at one data collection point could reengage at later waves.

A number of other important psychometric measures are reported at baseline. A measure of the parent's ability to interact and form attachment with others was measured 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 (~18). 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) falls within the 'normal' range for this scale indicating a moderate to small risk of abusive behavior. 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 (~2.9) was below the average score of 3.14 found for a representative American sample (The Panel Study of Income Dynamics, 2010). Normative scores are not available for the final two psychometric scales but 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 can 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 five years older on average than the PFL parents, with education levels in line with expected national averages. Fewer than 11 percent of parents in the GUI sample report either a physical or mental health condition, which is considerably lower than the PFL sample. A

much smaller proportion of the GUI sample indicated that they smoked during pregnancy (18 percent compared with approximately 50 percent in PFL), yet the proportion of respondents who drank alcohol during pregnancy was similar to PFL. A much higher proportion of the GUI sample were married (68 percent compared with approximately 16 percent in PFL), while the percentage that indicated having *either* a partner or spouse was similar to the PFL sample (88 percent compared with approximately 82 percent in PFL). Overall, this comparison highlights that the PFL cohort reflects a relatively disadvantaged sample 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.[9]

Table 2: Baseline comparison of high/low treatment participants

	High Treatment Mean (SD)			**Low Treatment Mean (SD)**		
Variables	***6 Month Sample***	***12 Month Sample***	***18 Month Sample***	***6 Month Sample***	***12 Month Sample***	***18 Month Sample***
Weeks in pregnancy at program entry	21.78 (7.83)	21.84 (7.88)	21.93 (7.93)	21.18 (6.87)	21.17 (7.02)	21.32 (6.62)
Mother's age at baseline interview	25.67 (5.76)	25.87 (6.01)	25.93 (5.91)	25.69 (6.04)	25.13 (6.02)	25.56 (6.10)
Married	0.16 (0.37)	0.16 (0.37)	0.16 (0.37)	0.17 (0.38)	0.16 (0.37)	0.15 (0.36)
Has partner (including married)	0.80 (0.40)	0.82 (0.39)	0.79 (0.41)	0.83 (0.38)	0.83 (0.38)	0.82 (0.39)
Living with parent(s)	0.55 (0.50)	0.54 (0.50)	0.54 (0.50)	0.45 (0.50)	0.48 (0.50)	0.47 (0.50)
First time mother	0.52 (0.50)	0.51 (0.50)	0.53 (0.50)	0.46 (0.50)	0.49 (0.50)	0.47 (0.50)
Low education	0.29 (0.46)	0.30 (0.46)	0.30 (0.46)	0.36 (0.48)	0.30 (0.46)	0.34 (0.48)
Mother employed	0.43 (0.50)	0.43 (0.50)	0.43 (0.50)	0.40 (0.49)	0.43 (0.50)	0.41 (0.50)
Saves regularly	0.50	0.50	0.49	0.53	0.55	0.53

[9] The GUI data are collected when children are aged 9/10 months and 36 months. Currently we have PFL data at 6 months, 12 months, and 18 months. We will carry out an outcomes comparison with GUI when the PFL 36 month surveys are completed.

	High Treatment Mean (SD)			Low Treatment Mean (SD)		
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
Social housing	0.54 (0.50	0.54 (0.50)	0.55 (0.50)	0.56 (0.50)	0.55 (0.50)	0.53 (0.50)
IQ	82.52 (12.94)	83.11 (12.60)	83.32 (12.35)	80.60 (13.14)	81.54 (12.75)	82.04 (12.16)
Physical Health Condition	0.76 (0.43)	0.76 (0.43)	0.75 (0.44)	0.64 (0.48)	0.65 (0.48)	0.63 (0.49)
Mental Health Condition	0.27 (0.45)	0.28 (0.45)	0.26 (0.44)	0.26 (0.44)	0.26 (0.44)	0.26 (0.44)
Smoking during pregnancy	0.51 (0.50)	0.51 (0.50)	0.51 (0.50)	0.49 (0.50)	0.46 (0.50)	0.47 (0.50)
Alcohol during pregnancy	0.27 (0.45)	0.29 (0.46)	0.29 (0.46)	0.25 (043)	0.26 (0.44)	0.27 (0.45)
Drugs during pregnancy	0.01 (0.11)	0.01 (0.11)	0.01 (0.11)	0.03 (0.18)	0.01 (0.11)	0.01 (0.12)
Vulnerable attachment (VASQ)	18.05 (3.67)	18.00 (3.87)	18.00 (3.92)	17.89 (4.04)	17.54 (3.86)	17.21 (3.54)
Positive parenting attitudes (AAPI)	5.23 (1.24)	5.25 (1.23)	5.28 (1.22)	5.20 (1.38)	5.33 (1.34)	5.27 (1.30)
Self-efficacy (Pearlin)	2.82 (0.60)	2.79 (0.60)	2.80 (0.60)	2.89 (0.63)	2.91 (0.62)	2.96 (0.62)
Self-esteem (Rosenberg)	13.06 (2.60)	12.98 (2.63)	12.96 (2.58)	12.75 (2.95)	12.78 (2.84)	12.82 (2.97)
Knowledge of infant development (KIDI)	72.40 (7.10)	72.33 (7.04)	72.46 (7.16)	70.51 (8.29)	70.70 (8.30)	70.92 (8.37)
N	82	82	80	89	82	73

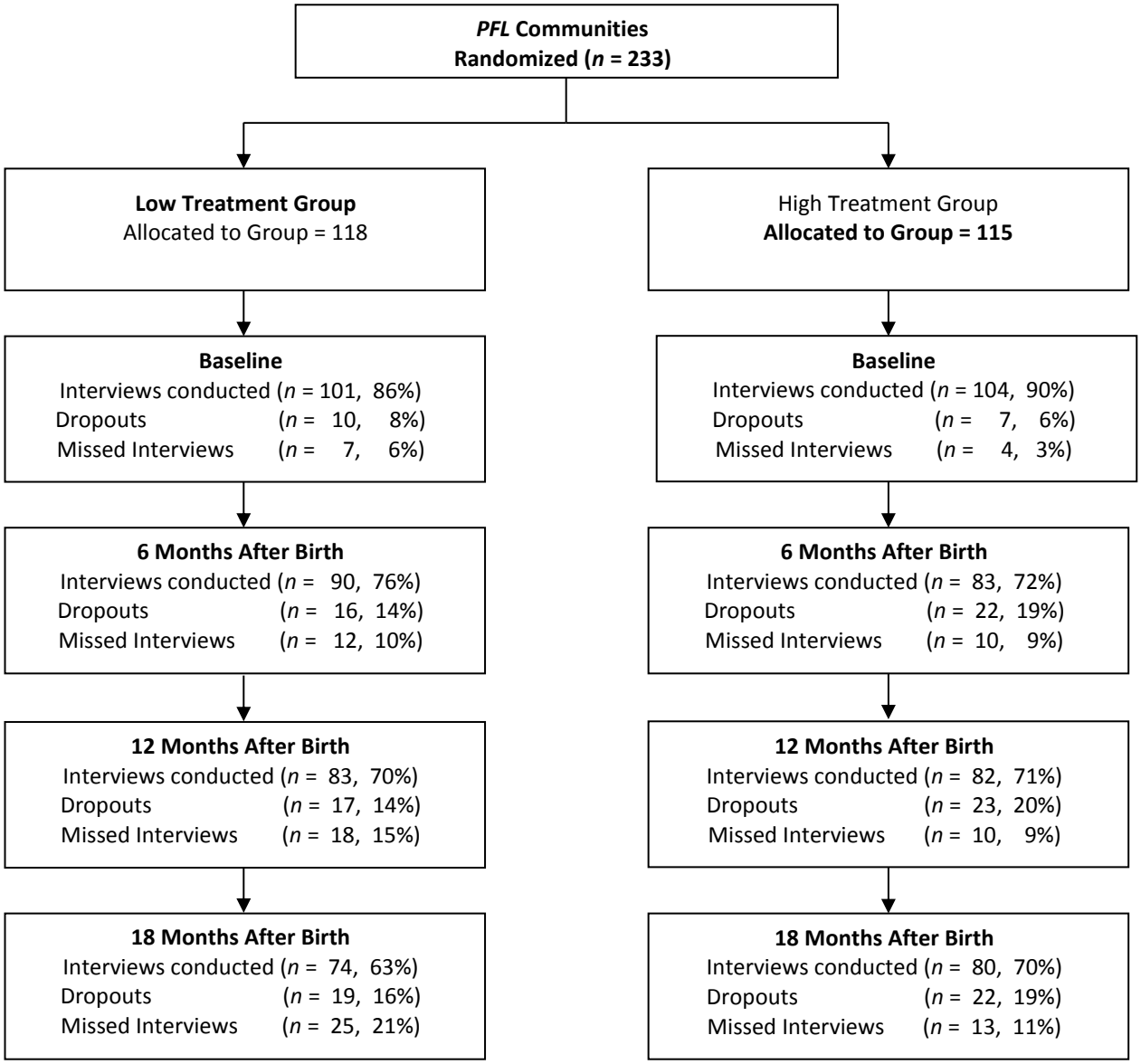
The Weschler Abbreviated Scale of Intelligence (WASI) was used to measure IQ at 3 months postpartum. The Vulnerable Attachment Style Questionnaire (VASQ) measures the respondents' interactions and dependence on other people. Scores above 15 are indicative of depressive disorders. The Adult Adolescent 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.

3.3.2 Attrition and Non-response

Figure 2 describes the progression of the participants from programme entry until eighteen months. The 18 month assessment captured 70 percent of the original high treatment group and 63 percent of the original low treatment group. Dropouts are defined as those who

actively told the *PFL* program staff or the evaluation team that they wanted to leave the program.

Figure 2: Flowchart of Program Participation and Dropping Out



On average, 19 percent of the high treatment group and 16 percent of the low treatment group were classified as official 'dropouts' between baseline and eighteen months, with the majority of dropout occurring before 6 months. Participant who 'missed interviews' are those who have not officially dropped out of the program, but were difficult to engage at

the assessment point. The proportion of missed interviews across the high and low treatment groups were 11 percent and 21 percent respectively between baseline and eighteen months. Due to attrition and non-response, the estimation samples differ at each of the data collection points. To account for the potential bias that this may introduce, we used an inverse probability weighting technique as a robustness test. This method is described in detail in Section 4.2.

4. Econometric Framework

4.1 Estimation Model and Outcome Measures

The PFL program is evaluated using an RCT. The standard model of program evaluation describes the observed outcome Y_i of participant $i \in I$ by

$$Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0) \quad (1)$$

where $I = \{1 \ldots N\}$ denotes the sample space, D_i denotes the treatment assignment for participant i ($D_i = 1$ if treatment occurs, $D_i = 0$ otherwise) and $(Y_i(0), Y_i(1))$ are potential outcomes for participant i. We test the null hypothesis of no treatment effect. This hypothesis is equivalent to the statement that the counterfactual outcome vectors share the same distribution H-1: $Y(1) \overset{d}{=} Y(0)$ where $\overset{d}{=}$ denotes equality in distribution.

Specifically, we examine 31 outcome measures related to child development and parenting. Table 3 summarizes each of the standardized scales. Various standardized psychometric scales were administered at the six month, 12 month and 18 month data collection waves. We restrict our analysis to standardized measures that are repeated in at least two of the three time points. This results in six child development instruments: the *Ages and Stage Questionnaire* (ASQ; Squires et al., 1999); the *Ages and Stages Questionnaire: Social-Emotional* (ASQ:SE; Squires et al., 2003); an assessment of temperament based on the

Infant Characteristics Questionnaire (Bates et al., 1979); the *MacArthur-Bates Communicative Development Inventories: Words and Gestures, Short Form (CDI-WG:* Fenson et al., 2000*),* the *Brief Infant-Toddler Social and Emotional Assessment* (*BITSEA*; Briggs-Gowan and Carter, 2006); and finally the *Developmental Profile 3, Cognitive Section* (*DP-3*; Alpern, 2007).

Parenting behavior is examined using three standardized scales: a measure of parental interactions with the child based on the *Community Support Inventory* (Centres for the Prevention of Child Neglect, 2000); the *Framingham Safety Survey* (FSS; American Academy of Pediatrics, 1991); and the Infant-Toddler version of the *Home Observation for Measurement of the Environment* (HOME; Caldwell and Bradley, 1984) combined with the *Supplement to the HOME Scale for Impoverished Families* (SHIF; Ertem et al., 1996). Further information on each of these scales and the subdomains listed in Table 3 can be found in Appendix D.

Table 3: Standardized Scales Measuring Child and Parent 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
	Difficult temperament: Infant Characteristics Questionnaire (Bates et al., 1979)	0 – 42	Unfavorable
	MacArthur-Bates Communicative Development Inventories: Words and Gestures, Short Form *(CDI-WG:* Fenson et al., 2000*)* Subdomains: *First Signs of Understanding*, *First Communicative Gestures*, *Words Understood*, *Words Produced*	 0 – 3 0 – 12 5 – 99 5 – 99	 Favorable

	Brief Infant-Toddler Social and Emotional Assessment (BITSEA; Briggs-Gowan & Carter, 2006) Subdomains: *Competence,* *Problem Solving*	 0 – 44 0 – 22	 Favorable Unfavorable
	Developmental Profile 3, Cognitive Section (DP-3; Alpern, 2007)	70 – 140	Favorable
Parenting	Interaction With Baby: Community Support Inventory (Centres for the Prevention of Child Neglect, 2000)	0 – 5	Favorable
	Framingham Safety Survey (FSS; American Academy of Pediatrics, 1991)	0 – 10	Favorable
	HOME Observation for Measurement of the Environment (HOME; Caldwell and Bradley, 1984) Subdomains: *variety ,* *organization, involvement,* *acceptance,* *learning materials,* *responsivity* HOME Observation for Measurement of the Environment and The Supplement to the HOME Scale for Impoverished Families (SHIF; Ertem et al., 1996) Subdomains: *child care, outings,* *restriction,* *acceptance,* *physical environment,* *toys and books, play, daily routines,* *interaction*	 0 – 5 0 – 6 0 – 8 0 – 9 0 – 11 0 – 20 0 – 5 0 – 6 0 – 8 0 – 9 0 – 10 0 – 14	Favorable Favorable Favorable

4.2 Permutation Testing

Although the RCT design in (1) is a simple specification, the use of traditional *t* tests for hypothesis testing is not viable given the small sample size and the likely non-normality of the data. Permutation methods do not depend on distributional assumptions and thus facilitate the estimation of treatment effects in small samples. While our analysis replicates one recent study of an early childhood intervention using this approach (Heckman et al., 2010a), it is not yet extensively used in the policy evaluation literature.

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.[10] 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 one sided (right tailed) *p*-values in order to test whether the high level treatment is having a positive effect on child and parenting outcomes compared to the low level treatment. We adopt a 10% p-value to assess statistical significance.

4.3 The Stepdown Procedure

Conducting permutation tests for each of the 31 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 (Romano, Shaikh, and Wolf, 2010). This procedure adjusts the *p*-values associated with individual tests to account for the effect of testing multiple outcomes.

The stepdown procedure involves firstly calculating a test statistic for each null hypothesis in a family of outcomes - we use the *t*-statistic. The test statistics are then placed in descending order. Using the permutation testing method described above, the largest observed *t*-statistic is compared with the distribution of the maximal permuted *t*-statistics. If the

[10] 100,000 replications are permuted using Monte Carlo simulations in our analyses.

probability of observing this statistic by chance is high ($p \geq 0.1$) we fail to reject the joint null hypothesis that the high treatment has no impact on any outcome in the cluster of hypotheses 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 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) – monotonicity with respect to the critical values. This ensures that the largest unadjusted *p*-value corresponds to the largest adjusted *p*-value (Heckman et al., 2010a).

The 31 outcomes measures were placed into a number of Stepdown families for the purposes of analysis. The outcomes included in each family should be correlated and represent an underlying construct, however outcomes which are derived from the same measure should not be included in the same Stepdown family. For the child development outcomes, we include one Stepdown family at 6 months (ASQ) and 4 at 12 and 18 months (ASQ, CDI, CDI NORM, & BITSEA). For the parenting outcomes, we include two Stepdown families at 6 and 12 months (HOME & HOME and SHIF).

5. Results

5.1 Analysis of Treatment Effects at Six, Twelve and Eighteen Months[11]

The impact of the program on child development and parenting are presented in Tables 4 and 5. We present the mean outcome scores by group, the *p*-values that result from individual permutation testing (*p(i)*), and the adjusted *p*-values calculated using the stepdown procedure (*p(ii)*). These results are presented for each of three waves. Note that in order to implement the stepdown method, all measures included in a stepdown category must be scored in a consistent direction given that we employ one-tailed tests.

Superscripts are presented for the *p(ii)* values, indicating their relative magnitude within the block. Thus superscript 1 indicates the measure which corresponds to the largest *t*-statistic. Each adjusted *p(ii)*-value represents the likelihood of rejecting the joint null hypothesis when the variables of higher ordering are excluded. For example, in Table 4, the first adjusted *p(ii)*-value (0.441) in the *ASQ Scores* category is the result of jointly testing all six outcomes in that category. The next adjusted *p(ii)*-value (0.483) is the result of excluding the *ASQ Gross Motor Score* variable from the joint-hypothesis test. The adjusted *p(ii)*-value of 0.800 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. We order this stepdown reporting in line with the 6 month data in our tables.

[11] As the potential for contamination in PFL is high given the geographical proximity of the participants, a number of strategies were devised to measure cross-talk and information flows between the two treatment groups (information on these strategies can be found in Doyle & Hickey, 2013). Results of this analysis find that while the conditions for contamination or spillover effects is quite high as participants are regularly in contact with each and share material, the blue-dye analysis suggests that these practices do not translate into improved parenting knowledge among the low treatment group suggesting that contamination from the high to low treatment group is quite low (see Doyle & PFL Evaluation Team, 2013).

5.1.1 Child Development

Table 4 presents the results for the child development outcomes. The results indicate that there is little evidence of a treatment effect on development measures at the early stages of the children's lives. Only three of the individual test statistics are statistically significant (*ASQ Fine Motor* at 12 months, *ASQ Gross Motor* at 18 months, and *DP-3 Cognitive Development* at 18 months). The ASQ fine scores measure the child's ability to engage in developmentally appropriate finger and hand movements, while the ASQ gross motor scores measure the child's ability to display developmentally appropriate movement skills (e.g. walking and kicking). The DP-3 Cognitive Development score measures the child's general cognitive abilities. However, the stepdown procedure fails to reject the null hypothesis of no treatment effect on child development outcomes at six months, 12 months and 18 months.

The available literature on the impact of interventions on early child development is consistent with our results. With respect to cognitive development, Anisfeld et al. (2004) report that the Healthy Families America has no impact at six or 12 months. Similarly, the German Pro Kind program (Jungmann et al., 2009) does not have an effect on cognitive functioning at six month based on simple *t*-tests on a small sample (N = 76).

With respect to noncognitive skills, in contrast to our results, Olds et al. (2002) use a Logit model on a relatively large sample (N= 543) and find that the Nurse Family Partnership program is effective at reducing emotional vulnerability in response to fear stimuli at six months. While Jungmann et al. (2009) find that Pro Kind reduces the presence of symptoms of a difficult temperament at six months. Overall, we do not identify a precisely determined treatment effect with respect to any non-cognitive development measures (*Difficult Temperament, ASQ Personal Social Score, ASQ Social-Emotional Score, BITSEA*).

Table 4: **Estimated Treatment Effects for Child Development Outcomes as 6, 12, 18 Months**

	6 Months					*12 Months*					*18 Months*				
Instrument	*N*	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$p^{(i)}$	$p^{(ii)}$	*N*	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$p^{(i)}$	$p^{(ii)}$	*N*	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$p^{(i)}$	$p^{(ii)}$
ASQ Scores															
ASQ Gross Motor Score	173	40.78 (11.93)	38.50 (12.99)	0.117	0.441[1]	165	42.07 (18.34)	40.72 (18.27)	0.319	0.721[3]	154	56.31 (5.44)	53.72 (12.02)	**0.047****	0.204[1]
ASQ Communication	173	53.07 (7.84)	51.78 (8.49)	0.154	0.483[2]	165	49.88 (10.74)	50.18 (10.55)	0.575	0.783[5]	154	45.69 (13.16)	45.34 (13.96)	0.437	0.636[5]
(-)ASQ Social-Emotional Score	173	14.76 (10.68)	15.17 (13.75)	0.403	0.800[3]	165	23.48 (21.51)	21.14 (16.05)	0.779	0.779[6]	154	29.13 (19.92)	29.05 (31.84)	0.506	0.506[6]
ASQ Personal Social Score	171	46.36 (12.07)	45.94 (13.57)	0.418	0.709[4]	165	49.88 (8.82)	48.55 (10.46)	0.190	0.559[2]	154	50.88 (7.91)	49.46 (9.24)	0.160	0.475[2]
ASQ Fine Motor Score	172	50.85 (9.52)	51.39 (10.17)	0.638	0.816[5]	165	54.33 (8.63)	51.87 (10.29)	**0.050***	0.219[1]	154	54.13 (8.26)	53.38 (8.28)	0.291	0.644[3]
ASQ Problem Solving Score	173	51.87 (9.39)	52.56 (9.92)	0.679	0.679[6]	164	46.40 (11.71)	46.40 (13.13)	0.499	0.826[4]	153	45.69 (11.60)	45.07 (10.69)	0.369	0.669[4]
Communicative Development Inventory (CDI)															
First Signs of Understanding		-	-	-	-	151	2.97 (0.16)	2.96 (0.20)	0.308	0.508[1]	148	2.99 (0.11)	2.94 (0.37)	0.178	0.300[1]
First Communicative Gestures		-	-	-	-	147	9.01 (2.23)	9.78 (1.96)	0.986	0.986[2]	149	11.27 (1.37)	11.41 (1.26)	0.740	0.740[2]
Communicative Development Inventory (CDI) NORMS															
Vocabulary Words Produced NORM		-	-	-	-	80	57.34 (33.90)	55.08 (33.71)	0.383	0.561[1]	85	53.18 (29.97)	58.61 (26.50)	0.811	0.912[1]
Vocabulary Words Understood NORM		-	-	-	-	80	71.71 (26.61)	82.49 (17.00)	0.984	0.984[2]	85	64.89 (31.20)	73.51 (24.13)	0.923	0.923[2]
Brief Infant-Toddler Social and Emotional Assessment (BITSEA)															
BITSEA Competence Score		-	-	-	-	162	15.44 (3.41)	14.88 (3.57)	0.154	0.274[1]	154	17.85 (2.61)	17.59 (3.45)	0.305	0.461[1]
(-)BITSEA Problem Score		-	-	-	-	165	8.82 (5.74)	8.90 (6.49)	0.466	0.466[2]	154	9.44 (6.63)	9.14 (7.18)	0.606	0.606[2]
Non Step-down Measures															

(-)Difficult Temperament	173	11.70 (5.71)	12.21 (5.50)	0.275	-	164	12.60 (5.54)	13.30 (5.76)	0.216	-		-	-	-	-
DP3: Cognitive development standardised score		-	-	-	-	165	116.20 (13.66)	115.13 (16.03)	0.323	-	153	119.01 (15.83)	114.53 (17.94)	**0.053***	-

Notes: '*N*' indicates the sample size. '*M*' indicates the mean. 'SD' indicates the standard deviation. [(i)] one-tailed (right-sided) *p*-value from an individual permutation test with 100,000 replications. [(ii)] one-tailed (right-sided) *p*-value from a Step-down permutation test with 100,000 replications and the superscripts indicate the ordering in which the variables are dropped in the Step-down analysis from the largest to smallest T statistic. (-) indicates the variable was reverse coded for the testing procedure. ***, **, * indicate that the test is statistically significant at the 1%, 5%, and 10% level respectively.

5.1.2 Parenting

The results for parenting outcomes are presented in Table 5. The format of the table is the same as that of Table 4. For the HOME instrument, the *variety* measure (which measures the child's frequency of interaction with individuals other than their mother along with the variety of the child's play environment) and the *learning* measure (which relates to the number of toys and books in the home which are designed to facilitate child development) were found to be statistically significant at 6 months using the individual permutation testing method. The first stepdown *p*-value is also statistically significant, indicating a rejection of the joint-null hypothesis of no effect across all subdomains. In the next step, when the *variety* measure is excluded and the remaining subdomains are tested, the joint-null hypothesis is also rejected. This reinforces the evidence of a positive treatment effect on each of these home environment measures. At 18 months a statistically significant effect was found for the *acceptance* measure (which represents parental acceptance of negative behaviour from the child and avoidance of unnecessary punishment) using the individual testing method. However, this effect was no longer precisely determined when adjusted for joint-hypothesis testing. Together these results suggest that the PFL mentoring program successfully prompted parents to provide variety and quality learning materials to their children and there is some evidence that it also encouraged parents to accept their child's behavior. Overall, the results for the home environment are weaker at 18 months than those estimated at 6 months.

For the HOME and SHIF combined instrument, a statistically significant treatment effect was estimated for the *childcare* subdomain (which relates to the range and adequacy of the care arrangements) and the *toys and books* subdomain (which counts the total number of stimulating play materials and books in the home) when the individual permutation testing method was applied. An examination of the stepdown results indicates that the joint-null hypothesis is also rejected giving strong evidence of a true impact of the program on the

home environment. In the second step, when the *childcare* measure is excluded and the remaining measures are jointly tested, the joint-null hypothesis can no longer be rejected. This indicates that the impact of the program on the *childcare* subdomain is driving the rejection of the joint-null hypothesis. At 18 months, a statistically significant treatment effect was found with respect to the *restriction* subdomain (which measures the level of restraint the parent places on the child during the visit, in the form of physical punishment and scolding, as well as inappropriate handling by older children) when the individual testing method was used. However, this effect is no longer precisely determined when adjusted for joint-hypothesis testing. Together the results for the HOME and SHIF instrument indicate that there is strong evidence that the treatment successfully impacted upon the mother's choice to use appropriate childcare for their baby. There is some weaker evidence to also indicate that mothers in the high treatment group were more likely to have appropriate toys and books to aid the child's development and they were less likely to employ inappropriate punishment approaches. The total score on the SHIF instrument was also tested separately and the high treatment group was found to score significantly higher than the low treatment group at 6 months. This provides further evidence of a significant treatment effect with respect to the home environment.

The interaction with baby scale contains no subdomains and therefore it was not included in the stepdown procedure. However, the individual permutation testing result is statistically significant at both six months and 18 months. Mothers in the high treatment group were more likely to engage in activities such as peek-a-boo games, storytelling, and taking their child shopping.

Many other home visiting programs have used the HOME inventory as an outcome measure at six, 12, and 18 months. In line with our results, many of these studies find positive effects on the quality of the home environment. Using the MANOVA testing method Culp

(2004) reports that the Community-based Family Resource Service Program had a positive impact on the *acceptance* and *organization* subscales at both six months and 12 months. Similarly, Wagner (1996) finds a positive effect of the Parents as Teachers program on the total HOME score, the *responsivity* domain, and the *play* domain at 12 months. Also, at the 18 month milestone, Black et al. (1994) find a positive effect of a home intervention for drug abusing mothers on the *variety* and *responsivity* domains, as well as the overall HOME score.

While the interaction with baby scale used in the PFL evaluation has not been used in other evaluations of home visiting programs, similar measures such as the Nurse Child Assessment Satellite Training (NCAST, Sumner and Spietz, 1994) have been used to measure parent-child interactions and activities with the child during structured play. The literature examined shows no evidence of a treatment effect on the NCAST measure at the six months, 12 months or 18 months (Anisfeld et al. (2004) with Healthy Families America at 6 and 12 months; Caughy et al. (2004) with Healthy Steps at 18 months; Duggan et al. (1999) with Hawaii's Healthy Start Program; Koniak-Griffen et al. (2002) with The Early Intervention program at 12 months). Using similar measures to examine the level of stimulation that parents provide for their children, other studies have also found no significant impact of home visiting by 18 months (Siegel et al. (1980) with the Hospital and Home Support Intervention During Infancy at 4 and 12 months; Schuler et al. (2002) at 18 months for an early intervention for drug-using mothers). In contrast, we find that PFL has a robust effect on parent-child interactions and this result is statistically significant at both six months and 18 months postpartum.

Table 5: **Estimated Treatment Effects for Parenting Outcomes as 6, 12, 18 Months**

	6 Months					*12 Months*					*18 Months*			
Instrument	*N*	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	*N*	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$
Home Observation for Measurement of the Environment (HOME)														
Variety	169	3.54 (1.12)	3.11 (1.01)	**0.005*** **	**0.030** *[1]**	-	-	-	-	154	4.08 (1.00)	3.99 (1.05)	0.309	0.493[5]
Learning Materials	125	6.88 (1.65)	6.26 (1.72)	**0.021** **	**0.097** *[2]**	-	-	-	-	96	8.24 (0.97)	8.04 (1.12)	0.176	0.467[3]
Responsivity	74	8.83 (1.73)	8.55 (2.32)	0.276	0.690[3]	-	-	-	-	88	9.50 (1.59)	9.07 (2.08)	0.144	0.455[2]
Acceptance	119	6.36 (0.56)	6.36 (0.60)	0.484	0.856[4]	-	-	-	-	86	6.12 (0.80)	5.66 (1.45)	**0.035** **	0.176[1]
Organisation	140	5.57 (0.64)	5.58 (0.66)	0.543	0.768[5]	-	-	-	-	125	5.52 (0.69)	5.45 (0.78)	0.290	0.590[4]
Involvement	125	4.28 (1.25)	4.40 (1.25)	0.697	0.697[6]	-	-	-	-	97	3.88 (1.47)	4.23 (1.56)	0.872	0.872[6]
Home Observation for Measurement of the Environment & Supplement to the HOME for Impoverished Families (HOME and SHIF)														
Childcare	169	4.19 (0.59)	3.94 (0.82)	**0.013** **	**0.095** *[1]**	-	-	-	-	154	3.84 (0.77)	3.77 (0.79)	0.299	0.692[5]
Toys and Books	170	7.75 (1.75)	7.28 (1.80)	**0.042** **	0.314[2]	-	-	-	-	155	9.36 (0.92)	9.32 (1.03)	0.399	0.603[7]
Daily Routines	168	7.36 (1.40)	7.13 (1.23)	0.129	0.527[3]	-	-	-	-	154	8.14 (1.31)	8.11 (1.20)	0.437	0.437[8]
Play	153	7.24 (1.62)	7.03 (1.44)	0.191	0.584[4]	-	-	-	-	142	7.22 (1.60)	7.13 (1.75)	0.369	0.687[6]
Interaction	71	11.50 (1.99)	11.26 (2.75)	0.338	0.822[5]	-	-	-	-	88	12.13 (2.08)	11.43 (2.91)	0.101	0.453[2]
Physical Environment	123	6.16 (1.11)	6.08 (1.13)	0.343	0.777[6]	-	-	-	-	91	6.39 (1.35)	6.02 (1.44)	0.107	0.439[3]
Outings	168	4.76 (0.46)	4.80 (0.43)	0.683	0.866[7]	-	-	-	-	154	4.78 (0.53)	4.69 (0.60)	0.172	0.557[4]

Restrictions/Not Items	128	5.97 (0.18)	5.99 (0.12)	0.781	0.781[8]	-	-	-	-	94	5.61 (0.64)	5.33 (1.21)	**0.089***	0.453[1]
Non Step Down Measures														
Total SHIF Score	111	16.94 (1.38)	16.61 (1.31)	**0.099***	-	-	-	-	-	88	17.40 (1.98)	17.04 (2.08)	0.209	-
Framingham Safety Survey	172	7.37 (0.77)	7.46 (0.68)	0.782	-	-	-	-	-	146	8.32 (0.98)	8.33 (0.93)	0.505	-
Interaction With Baby	173	2.79 (0.61)	2.66 (0.53)	**0.082***	-	-	-	-	-	153	3.21 (0.48)	3.05 (0.47)	**0.020****	-

Notes: '*N*' indicates the sample size. '*M*' indicates the mean. 'SD' indicates the standard deviation. [(i)] one-tailed (right-sided) *p*-value from an individual permutation test with 100,000 replications. [(ii)] one-tailed (right-sided) *p*-value from a Step-down permutation test with 100,000 replications and the subscripts indicate the ordering in which the variables are dropped in the Step-down analysis from the largest to smallest T statistic. ***, **, * indicate that the test is statistically significant at the 1%, 5%, and 10% level respectively.

5.2 Robustness Tests

5.2.1 Addressing Attrition and Non-Response

Due to attrition and non-response, the estimation sample sizes differ at each data collection point. To account for the potential bias that attrition and non-response may introduce, we test the robustness of the main analysis using an inverse probability weighting (IPW) technique. This involves two main steps: first a logit model is applied to calculate the probability of completing the research questionnaire at each time point using an exhaustive set of baseline characteristics. Second, the predicted probabilities from the logit models are applied as weights in the permutation analysis so that a larger weight is applied to individuals that are underrepresented in the sample due to missing observations.

With over a hundred baseline measures that could potentially be included as right-hand side variables in the logit model, it was necessary to restrict this variable set. In order to do this, we first ran bivariate analyses in which 133 baseline measures were tested to determine whether a statistically significant difference existed between attrited and non-attrited groups. Note that the attrited group includes participants who officially dropped out of the program and those who did not complete a questionnaire at that particular assessment point, but may engage at another assessment point. The permutation testing method (using 10,000 replications) was employed to individually examine each measure and the testing was carried out separately for each estimation sample (6 months, 12 months and 18 months). In addition, within each estimation sample, the low and high treatment group were examined separately to allow for differential attrition processes in the two dosage groups. This resulted in six different sets of bivariate analyses. Overall, across all six analyses, the number of measures for which statistically significant (p-value < 0.1) differences were found between the attrition and non-attrition samples ranged from 17 to 35 measures out of 133, representing differences on 13-26 percent of the measures analyzed. Overall, the bivariate analyses suggest

that mothers who stay engaged with the program are a more socially advantaged group. For example, in five of the six analyses higher IQ scores were significant predictors of participating in the survey, in three models mothers in employment were more likely to participate, and across all six models mothers with lower education or literacy difficulties were more likely to dropout.

In order to select the baseline predictor set which should be included in each of the corresponding six logit models, we exclude any measure which was not statistical significant in the bivariate analyses. Also, in order to maximize the sample size, we restricted the baseline variable set by excluding any measure for which there were missing observations at baseline. We made one exception, however, with respect to the measure of cognitive resources. This instrument was administered to mothers when their baby had reached three months of ages and, therefore, some participants who completed a baseline interview did not complete this assessment. We included the measure of cognitive resources in our logit model as it was found to be a consistent predictor of attrition in the bivariate analysis[12]. Specifically, mothers who stayed in the program were likely to have higher levels of cognitive resources. Overall, a measure of cognitive resources is available for 88 percent of mothers who completed a baseline interview.

As the sample size for each logit model was quite small (low = 101; high = 90), some additional variables needed to be excluded due to lack of variation[13]. This resulted in six final

[12] The cognitive resources measure is a significant predictor of attrition in five out of six bivariate attrition analyses. Specifically, mothers with lower cognitive resources in the high treatment group were more likely to miss their 6 month, 12 month and 18 month interviews. Similarly mothers with lower cognitive resources in the low treatment group were also more likely to miss their 12 months and 18 month interview. In each of these analyses we found moderate to large Cohen's D effect sizes, ranging from 0.618 (6 months, high treatment group) to 0.995 (12 months, low treatment group). One exception was the low treatment group at 6 months: for this cohort no significant difference was found between mothers who participated and those who did not participate with respect to the measure of cognitive resources.

[13] This only applied to the low treatment group. Many of the binary variables which were included in the logit models at each of the three waves were unbalanced. That is to say that when the dependent variable is tabulated against each of these binary predictor measures there are zero observations in certain cells. This, in turn, implies that at certain values these binary variables are perfect predictors of success in the logit model. This leads to a

logit models in which a dummy variable (1 = non-attrition; 0 = attrition) was regressed on a restricted variable set ranging in size from seven variables to 13 variables. Although the bivariate analyses suggest that mothers who stay engaged with the program are a more socially advantaged group, the results from the multivariate analyses indicate few statistically significant predictors of attrition. For further details of the estimation model employed to calculate the probability weights see Table E1 in Appendix E.

After the probability of non-attrition was calculated using this technique, the weights were merged with the outcome datasets. Applying the inverse of each weight in our inference estimations ensures that a larger weight is applied to the participants that are under-represented in each estimation sample. Tables 6 and 7 show these results and can be read in the same manner as Tables 4 and 5.

Child Development

Table 6 shows that correcting for attrition bias in the domain of child development leads to more precisely determined results. Specifically, the statistically significant difference between the low and high treatment group with respect to the *ASQ fine motor* domain at 12 months rises from the 10% significance level to the 5% significance level. Similarly, at 18 months, the *p*-value associated with the *DP-3 Cognitive Development* measures raises from the 10% level to the 5% level, and a statistically significant difference is also found for the *ASQ Personal Social* domain, which was not found before. Most interestingly, at 18 months, we can now reject the joint null hypothesis of no impact on child development and the *ASQ gross motor* domain is found to be driving this rejection. This suggests that when we correct for misrepresentation due to attrition bias, the original child development estimates which were presented in Table 4 are echoed with more precision.

---

loss of observations if these measures are included in the model. We excluded these variables in order to maximize the sample size. This lead to the exclusion of 7 variables at 6 months, 2 variables at 12 months, and 1 variable at 18 months. Notice that as the number of attritors increase, this problem dissipates.

Table 6: **Estimated Treatment Effects for Child Development Outcomes at 6, 12, 18 Months Correcting for Attrition Bias Using Inverse Probability Weighting.**

	6 Months					*12 Months*					*18 Months*				
Instrument	*N*	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$p^{(i)}$	$p^{(ii)}$	*N*	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$p^{(i)}$	$p^{(ii)}$	*N*	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$p^{(i)}$	$p^{(ii)}$
ASQ Scores															
ASQ Gross Motor Score	173	40.75 (12.02)	38.36 (13.07)	0.115	0.431[1]	165	42.58 (17.75)	40.60 (18.59)	0.249	0.623[3]	154	56.50 (5.33)	53.12 (12.39)	**0.025****	**0.099***[1]
ASQ Communication	173	53.22 (7.35)	51.71 (8.62)	0.112	0.396[2]	165	50.13 (10.70)	50.02 (10.64)	0.476	0.808[4]	154	45.92 (13.02)	45.14 (14.16)	0.367	0.564[5]
ASQ Personal Social Score	171	46.43 (12.29)	45.89 (13.65)	0.398	0.776[3]	165	49.97 (8.74)	48.67 (10.55)	0.199	0.583[2]	154	51.40 (7.56)	49.46 (9.00)	**0.077***	0.299[2]
(-)ASQ Social-Emotional Score	173	15.00 (11.08)	15.14 (13.95)	0.469	0.793[4]	165	23.12 (20.76)	21.22 (15.93)	0.736	0.736[6]	154	29.56 (20.66)	30.50 (32.58)	0.430	0.430[6]
ASQ Fine Motor Score	172	51.13 (9.62)	51.25 (10.18)	0.526	0.725[5]	165	54.37 (8.60)	51.71 (10.39)	**0.041****	0.183[1]	154	54.31 (8.14)	52.99 (8.49)	0.179	0.461[3]
ASQ Problem Solving Score	173	52.29 (9.02)	52.44 (9.98)	0.540	0.540[6]	164	46.47 (11.71)	46.46 (13.17)	0.496	0.743[5]	153	46.17 (11.26)	44.99 (10.63)	0.259	0.538[4]
Communicative Development Inventory (CDI)															
First Signs of Understanding		-	-	-	-	151	2.97 (0.17)	2.96 (0.20)	0.377	0.565[1]	148	2.98 (0.13)	2.94 (0.38)	0.257	0.358[1]
First Communicative Gestures		-	-	-	-	147	9.11 (2.20)	9.79 (1.98)	0.971	0.971[2]	149	11.30 (1.39)	11.42 (1.27)	0.707	0.707[2]
Communicative Development Inventory (CDI) NORMS															
Vocabulary Words Produced NORM		-	-	-	-	80	57.35 (34.21)	55.06 (33.19)	0.382	0.559[1]	85	53.25 (30.47)	61.23 (26.39)	0.889	0.955[1]
Vocabulary Words Understood NORM		-	-	-	-	80	71.93 (27.06)	81.86 (17.94)	0.967	0.967[2]	85	65.79 (31.32)	74.54 (24.17)	0.920	0.920[2]
Brief Infant-Toddler Social and Emotional Assessment (BITSEA)															
BITSEA Competence Score		-	-	-	-	162	15.42 (3.44)	14.86 (3.58)	0.160	0.289[1]	154	17.79 (2.62)	17.31 (3.53)	0.185	0.306[1]
(-)BITSEA Problem Score		-	-	-	-	165	8.95 (5.96)	9.06 (6.44)	0.460	0.460[2]	154	9.77 (7.24)	9.52 (7.43)	0.575	0.575[2]

Non Step-down Measures															
(-)Difficult Temperament	173	11.70 (5.69)	12.30 (5.47)	0.247	-	164	12.64 (5.69)	13.10 (5.70)	0.307	-		-	-	-	-
DP3: Cognitive development standardised score		-	-	-	-	165	116.37 (13.98)	115.13 (16.01)	0.301	-	153	119.43 (16.12)	113.83 (18.19)	**0.030****	-

Notes: '*N*' indicates the sample size. '*M*' indicates the mean. 'SD' indicates the standard deviation. [(i)] one-tailed (right-sided) *p*-value from an individual permutation test with 100,000 replications applying IPW. [(ii)] one-tailed (right-sided) *p*-value from a Step-down permutation test with 100,000 replications applying IPW and the superscripts indicate the ordering in which the variables are dropped in the Step-down analysis from the largest to smallest T statistic. (-) indicates the variable was reverse coded for the testing procedure. ***, **, * indicate that the test is statistically significant at the 1%, 5%, and 10% level respectively.

Parenting

Table 7 shows that while correcting for attrition bias leads to quite similar results for the child development measures, the adjustment leads to less precision when estimating program effects on parenting. Firstly, with respect to the 6 month results, for the HOME category, the same two subdomains (*variety* and *learning*) remain statistically significant when tested individually. Also, the joint null hypothesis can still be rejected, although the stepdown *p*-value is now less precise at the 10% level (as compared with the 5% level in Table 5). When the *variety* measure is excluded from joint testing, we now fail to reject the null hypothesis of no effect on the remaining HOME measures. This indicates that the attrition correction leads to weaker evidence of a treatment effect on the *learning* domain. With respect to the HOME and SHIF combined instrument, the *childcare* and *toys and books* subdomains remain statistically significant when tested individually. However, the joint-null hypothesis can no longer be rejected. Also, the *total SHIF score* and the *interactions with baby* measure are no longer statistically significant. At 18 months, a statistically significant difference remains between the low and high treatment group with respect to the HOME subdomain of *acceptance*. However, the HOME and SHIF combined subdomain of *restriction* is no longer statistically significant. A statistically significant difference remains between the low and high treatment group with respect to the measure of *interactions with baby* at 18 months. Overall, correcting for attrition with respect to parenting measures leads to weaker precision with respect to estimates of the program effect, however quite a similar pattern emerges.

Table 7: **Estimated Treatment Effects for Parenting Outcomes at 6, 12, 18 Months Correcting for Attrition Bias Using Inverse Probability Weighting.**

	6 Months					*12 Months*					*18 Months*				
Instrument	*N*	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	*N*	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$
Home Observation for Measurement of the Environment (HOME)														
Variety	169	3.49 (1.13)	3.12 (1.01)	**0.018****	**0.093***[1]	-	-	-	-	154	4.02 (1.05)	3.95 (1.04)	0.351	0.555[5]
Learning Materials	125	6.78 (1.69)	6.25 (1.72)	**0.049***	0.208[2]	-	-	-	-	96	8.23 (0.97)	8.07 (1.12)	0.236	0.500[4]
Responsivity	74	8.72 (1.79)	8.54 (2.32)	0.362	0.794[3]	-	-	-	-	88	9.43 (1.67)	8.89 (2.18)	0.122	0.375[2]
Acceptance	119	6.35 (0.54)	6.35 (0.60)	0.512	0.871[4]	-	-	-	-	86	6.10 (0.81)	5.65 (1.44)	**0.041****	0.205[1]
Organisation	140	5.56 (0.64)	5.60 (0.66)	0.650	0.855[5]	-	-	-	-	125	5.57 (0.65)	5.44 (0.77)	0.161	0.456[3]
Involvement	125	4.25 (1.23)	4.39 (1.24)	0.733	0.733[6]	-	-	-	-	97	3.90 (1.44)	4.25 (1.53)	0.877	0.877[6]
Home Observation for Measurement of the Environment & Supplement to the HOME for Impoverished Families (HOME and SHIF)														
Childcare	169	4.15 (0.60)	3.96 (0.82)	**0.056***	0.402[1]	-	-	-	-	154	3.81 (0.77)	3.79 (0.79)	0.431	0.431[8]
Toys and Books	170	7.68 (1.78)	7.30 (1.83)	**0.093***	0.480[2]	-	-	-	-	155	9.35 (0.96)	9.32 (1.02)	0.420	0.653[7]
Daily Routines	168	7.26 (1.38)	7.13 (1.23)	0.278	0.811[3]	-	-	-	-	154	8.11 (1.31)	8.01 (1.28)	0.337	0.737[5]
Play	153	7.20 (1.62)	7.06 (1.40)	0.287	0.779[4]	-	-	-	-	142	7.17 (1.61)	7.06 (1.76)	0.342	0.680[6]
Physical Environment	123	6.14 (1.13)	6.07 (1.12)	0.371	0.846[5]	-	-	-	-	91	6.38 (1.36)	5.98 (1.46)	0.107	0.492[2]
Interaction	71	11.41 (2.03)	11.25 (2.74)	0.393	0.800[6]	-	-	-	-	88	12.06 (2.15)	11.32 (2.94)	0.106	0.514[1]
Restrictions/Not Items	128	5.97 (0.18)	5.99 (0.12)	0.744	0.940[7]	-	-	-	-	94	5.58 (0.64)	5.33 (1.23)	0.129	0.516[3]

Outings	168	4.75 (0.46)	4.80 (0.43)	0.764	0.764[8]	-	-	-	-	154	4.78 (0.51)	4.70 (0.59)	0.194	0.620[4]
Non Step Down Measures														
Total SHIF Score	111	16.88 (1.36)	16.60 (1.30)	0.136	-	-	-	-	-	88	17.33 (1.98)	17.05 (2.05)	0.276	-
Framingham Safety Survey	172	7.39 (0.79)	7.45 (0.69)	0.697	-	-	-	-	-	146	8.37 (0.99)	8.32 (0.95)	0.404	-
Interaction With Baby	173	2.78 (0.62)	2.67 (0.54)	0.119	-	-	-	-	-	153	3.21 (0.49)	3.03 (0.47)	**0.018****	-

Notes: '*N*' indicates the sample size. '*M*' indicates the mean. 'SD' indicates the standard deviation. [(i)] one-tailed (right-sided) *p*-value from an individual permutation test with 100,000 replications applying IPW. [(ii)] one-tailed (right-sided) *p*-value from a Step-down permutation test with 100,000 replications applying IPW and the subscripts indicate the ordering in which the variables are dropped in the Step-down analysis from the largest to smallest T statistic. ***, **, * indicate that the test is statistically significant at the 1%, 5%, and 10% level respectively.

5.2.2 Addressing Missing Data

While the degree of item non-response was minimal for the majority of standardised instruments used (less than 2% at each time point)[14], there were more substantial cases of missing data on the Home Observation for Measurement of the Environment (HOME) and Supplement to the HOME Scale for Impoverished Families (SHIF) instrument. Missing data on the HOME/SHIF measure arises for two reasons. First, as some of the HOME/SHIF items are based on observations of parent-child interactions, if the child is not present or is asleep when the interview takes place, these items cannot be measured. 25% of children were not present at the 6 month interview and 39% were not present at the 18 month interview. Second, as some of the items in the HOME are based on observation of materials available in the home, these items cannot be assessed if the interview is not conducted in the home. 16% of interviews were not conducted in the home at the 6 month interview and 21% were conducted outside of the home at 18 months. One concern is that there may be an element of self-selection by parents who either did not want the interview to be conducted in the home or did not want their child to be present for the interview.

To address this issue, an inverse probability weighting method was applied using a similar approach to that described in Section 5.2.1. However, rather than constructing one set of weights for each data collection wave, instead multiple sets were constructed so that the each of the 31 variables were modelled individually at each wave. Thus, rather than using baseline measure to predict the likelihood of participating in each survey, instead we modelled the likelihood of data being available for each individual measure. Observations which are classified as missing include those that were lost to attrition as well as those who participated in the survey but did provide data for the corresponding outcome. The analyses were again carried out separately for the low and high treatment group to allow for differential missing data patterns in each dosage group. As in

[14] One exception is the *MacArthur-Bates Communicative Development Inventories: Words and Gestures, Short Form (CDI-WG:* Fenson et al., 2000*)* scale items. For the items on this scale, the level of missing data was less than 11% at 12 months and less than 6 percent at 18 months. The CDI-WG manual instructs that measure should not be imputed (Fenson et al., 2000). This instrument is completed by parents using a paper form. It contains 104 items and it is likely that mothers could miss some questions. Therefore, the data are likely to be missing at random.

Section 5.2.1, we first conducted bivariate analysis across 133 baseline variables to check for statistical difference between the individuals who provided data and those who did not. This resulted in 136 separate bivariate analyses. The corresponding logit models were then constructed by using only the significant variables from the bivariate analyses that contained no missing data (with the exception of the cognitive resources measure as described before). Also, in situations where binary predictor measures resulted in a loss of observations in the logit model due to a lack of variation, these measures were also excluded from the right-hand side using the same approach as that described previously. The analyses presented in Tables 8 and 9 represent the impact of the program on child development and parenting while correcting for missing data using IPW. The method leads to very similar results to those which were found for the correction of attrition bias. The only difference found in Table 8 as compared with Table 6 is that the stepdown adjusted *p*-value associated with the *ASQ gross motor score* subdomain at 18 months is no longer statistically significant. With respect to parenting measures, Table 9 differs from Table 7 in the sense that the stepdown adjusted *p*-value associated with the HOME *variety* subdomain is no longer statistically significant and the individual *p*-value associated with the *toys and books* subdomain is no longer statistically significant when tested using the individual permutation testing method.

Correcting for attrition leads to largely similar results regarding child development yet weaker results for the parenting measures. The predictors of staying engaged in the program suggested that mothers who dropout are more socially disadvantaged that those who remain. Therefore, the results in Tables 7 and 9 appear to suggest that more advantaged families benefit more from the program with respect to parenting outcomes. In order to understand the contrast, subgroup analyses were conducted to determine whether the treatment effects for the participants with above median IQ scores differed from those with median IQ scores or below. No obvious pattern emerged from these subgroup analyses. At 6 months, more significant treatment effects were found for the higher IQ group, while at 12 and 18 months there are

more significant treatment effects for the lower IQ group. Also it does not appear that effects differ for child development and parenting. Given that parenting measures, and the HOME measure in particular, suffer most from the problem of missing data, it is possible that the weaker estimation in the parenting outcomes is due to the small sample size and a lack of representative observations with which to weight the estimates.

Table 8: **Estimated Treatment Effects for Child Development Outcomes at 6, 12, 18 Months Accounting for Missing Data Using Inverse Probability Weighting.**

	6 Months					*12 Months*					*18 Months*				
Instrument	*N*	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$p^{(i)}$	$p^{(ii)}$	*N*	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$p^{(i)}$	$p^{(ii)}$	*N*	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$p^{(i)}$	$p^{(ii)}$
ASQ Scores															
ASQ Gross Motor Score	173	40.75 (12.02)	38.36 (13.07)	0.115	0.430[1]	165	42.58 (17.75)	40.60 (18.59)	0.249	0.618[3]	154	56.50 (5.33)	53.12 (12.39)	**0.025****	0.103[1]
ASQ Communication	173	53.22 (7.35)	51.71 (8.62)	0.113	0.395[2]	165	50.13 (10.70)	50.02 (10.64)	0.476	0.805[4]	154	45.92 (13.02)	45.14 (14.16)	0.367	0.564[5]
ASQ Personal Social Score	171	46.36 (12.32)	45.89 (13.65)	0.411	0.788[3]	165	49.97 (8.74)	48.67 (10.55)	0.199	0.579[2]	154	51.40 (7.56)	49.46 (9.00)	**0.077***	0.305[2]
(-)ASQ Social-Emotional Score	173	15.00 (11.08)	15.14 (13.95)	0.469	0.793[4]	165	23.12 (20.76)	21.22 (15.93)	0.736	0.736[6]	154	29.56 (20.66)	30.50 (32.58)	0.430	0.430[6]
ASQ Fine Motor Score	172	51.12 (9.35)	51.25 (10.18)	0.531	0.728[5]	165	54.37 (8.60)	51.71 (10.39)	**0.041****	0.182[1]	154	54.31 (8.14)	52.99 (8.49)	0.179	0.401[4]
ASQ Problem Solving Score	173	52.29 (9.02)	52.44 (9.98)	0.540	0.540[6]	164	46.47 (11.71)	46.43 (12.58)	0.491	0.739[5]	153	46.17 (11.27)	44.26 (11.02)	0.163	0.434[3]
Communicative Development Inventory (CDI)															
First Signs of Understanding		-	-	-	-	151	2.96 (0.19)	2.96 (0.19)	0.469	0.683[1]	148	2.99 (0.11)	2.93 (0.41)	0.137	0.244[1]
First Communicative Gestures		-	-	-	-	147	9.11 (2.20)	9.79 (1.98)	0.966	0.966[2]	149	11.23 (1.43)	11.42 (1.25)	0.785	0.785[2]
Communicative Development Inventory (CDI) NORMS															
Vocabulary Words Produced NORM		-	-	-	-	80	57.97 (34.11)	49.59 (32.55)	0.214	0.329[1]	85	58.38 (30.29)	64.68 (25.93)	0.770	0.897[1]
Vocabulary Words Understood NORM		-	-	-	-	80	72.56 (27.29)	77.75 (19.77)	0.779	0.779[2]	85	70.76 (29.51)	77.15 (22.87)	0.842	0.842[2]
Brief Infant-Toddler Social and Emotional Assessment (BITSEA)															
BITSEA Competence Score		-	-	-	-	162	15.33 (3.44)	14.96 (3.59)	0.256	0.429[1]	154	17.79 (2.62)	17.31 (3.53)	0.185	0.306[1]
(-)BITSEA Problem Score		-	-	-	-	165	8.95 (5.96)	9.06 (6.44)	0.460	0.460[2]	154	9.77 (7.24)	9.52 (7.43)	0.575	0.575[2]

Non Step-down Measures															
(-)Difficult Temperament	173	11.70 (5.69)	12.30 (5.47)	0.247	-	164	12.47 (5.57)	13.10 (5.70)	0.240	-		-	-	-	-
DP3: Cognitive development standardised score		-	-	-	-	165	116.37 (13.98)	115.13 (16.01)	0.301	-	153	119.43 (16.12)	113.78 (18.41)	**0.030****	-

Notes: '*N*' indicates the sample size. '*M*' indicates the mean. 'SD' indicates the standard deviation. [(i)] one-tailed (right-sided) *p*-value from an individual permutation test with 100,000 replications applying IPW. [(ii)] one-tailed (right-sided) *p*-value from a Step-down permutation test with 100,000 replications applying IPW and the superscripts indicate the ordering in which the variables are dropped in the Step-down analysis from the largest to smallest T statistic. (-) indicates the variable was reverse coded for the testing procedure. ***, **, * indicate that the test is statistically significant at the 1%, 5%, and 10% level respectively.

Table 9: **Estimated Treatment Effects for Parenting Outcomes at 6, 12, 18 Months Accounting for Missing Data Using Inverse Probability Weighting.**

	6 Months					*12 Months*					*18 Months*				
Instrument	*N*	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	*N*	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$
Home Observation for Measurement of the Environment (HOME)														
Variety	169	3.50 (1.13)	3.12 (1.01)	**0.017****	0.124[1]	-	-	-	-	154	4.02 (1.05)	3.96 (1.04)	0.366	0.715[4]
Learning Materials	125	6.60 (1.70)	6.01 (1.88)	**0.084***	0.247[2]	-	-	-	-	96	8.13 (1.00)	7.47 (1.29)	0.186	0.226[1]
Acceptance	119	6.46 (0.58)	6.39 (0.60)	0.296	0.701[3]	-	-	-	-	86	6.07 (0.81)	5.46 (1.48)	**0.028****	0.133[2]
Organisation	140	5.56 (0.64)	5.60 (0.66)	0.556	0.863[4]	-	-	-	-	125	5.47 (0.71)	5.40 (0.82)	0.336	0.773[3]
Responsivity	74	8.55 (1.86)	8.62 (2.22)	0.557	0.763[5]	-	-	-	-	88	8.97 (1.87)	8.81 (2.20)	0.437	0.637[5]
Involvement	125	4.25 (1.18)	4.30 (1.29)	0.590	0.590[6]	-	-	-	-	97	3.90 (1.36)	4.27 (1.59)	0.855	0.855[6]
Home Observation for Measurement of the Environment & Supplement to the HOME for Impoverished Families (HOME and SHIF)														
Childcare	169	4.15 (0.60)	3.99 (0.77)	**0.058***	0.491[1]	-	-	-	-	154	3.81 (0.77)	3.77 (0.78)	0.381	0.721[6]
Play	153	7.25 (1.69)	7.00 (1.40)	0.194	0.712[2]	-	-	-	-	142	7.18 (1.59)	7.07 (1.79)	0.359	0.765[5]
Physical Environment	123	6.24 (1.11)	6.04 (1.13)	0206	0.685[3]	-	-	-	-	91	5.97 (1.49)	5.42 (1.91)	0.268	0.505[2]
Daily Routines	168	7.29 (1.38)	7.12 (1.22)	0.194	0.670[4]	-	-	-	-	154	8.11 (1.31)	8.08 (1.23)	0.443	0.651[7]
Toys and Books	170	7.71 (1.79)	7.48 (1.85)	0.264	0.632[5]	-	-	-	-	155	9.35 (0.96)	9.34 (1.00)	0.459	0.459[8]
Interaction	71	11.20 (2.19)	11.29 (2.44)	0.555	0.890[6]	-	-	-	-	88	11.64 (2.30)	11.20 (2.89)	0.292	0.650[4]

Outings	168	4.79 (0.44)	4.82 (0.41)	0.701	0.914[7]	-	-	-	-	154	4.78 (0.51)	4.70 (0.59)	0.189	0.671[3]
Restrictions/Not Items	128	5.96 (0.20)	5.99 (0.11)	0.797	0.797[8]	-	-	-	-	94	5.57 (0.60)	5.12 (1.34)	**0.065***	0.325[1]
Non Step Down Measures														
Total SHIF Score	111	16.77 (1.37)	16.67 (1.33)	0.377	-	-	-	-	-	88	16.94 (2.16)	16.91 (1.85)	0.475	-
Framingham Safety Survey	172	7.38 (0.80)	7.45 (0.69)	0.730	-	-	-	-	-	146	8.40 (0.94)	8.35 (0.93)	0.398	-
Interaction With Baby	173	2.77 (0.63)	2.67 (0.54)	0.137	-	-	-	-	-	153	3.21 (0.49)	3.03 (0.47)	**0.015****	-

Notes: '*N*' indicates the sample size. '*M*' indicates the mean. 'SD' indicates the standard deviation. [(i)] one-tailed (right-sided) *p*-value from an individual permutation test with 100,000 replications. [(ii)] one-tailed (right-sided) *p*-value from a Step-down permutation test with 100,000 replications and the subscripts indicate the ordering in which the variables are dropped in the Step-down analysis from the largest to smallest T statistic. ***, **, * indicate that the test is statistically significant at the 1%, 5%, and 10% level respectively.

5.3 Reliability of Instruments

One potential limitation of our study is that the majority of instruments used rely on maternal self-reporting. These subjective measures may be less reliable than objective indicators as parents may misreport their children's level of development and their own parenting skills. . There is evidence of low-to-moderate cross-informant correlations in terms of child behavioural/emotional problems (Achenbach, McConaughy, and Howell 1987); and a study conducted within the current catchment area found that parents in the community systematically reported higher child skill levels compared to teacher reports (Doyle, Finnegan, & McNamara, 2012). If parents in both the high and low treatment groups systematically under or over-report the outcomes under analysis, this will not bias the results regarding programme effectiveness as the magnitude of the difference will be the same, however if one group systematically misreports and the other does not, this will bias our estimates of programme effectiveness.

To test the reliability of the self-reported measures we took advantage of the availability of observational items in the HOME and SHIF instruments. Restricting our analysis to solely observational items led to results at 18 months which echo those presented in Table 5 (available upon request). Specifically, at 18 months, the significant differences between the high and low treatment groups with respect to the *acceptance* and *restriction* subdomains remain when the analysis is restricted to observational items. In addition, the significant result for the *total SHIF score* remains. Two additional significant effects were also found with respect to the *organisiation* and *interaction* subdomains which could suggest that the element of noise associated with self-reported measures is masking some true effects.

In addition, our main measure of child development ,the ASQ, is a well-established child development screening tool and a number of studies have found that it is highly correlated with other previously validated measures that are completed by professionals (see

Squires, Bricker, and Potter, 1997; Squires et al., 1999). In particular, the overall level of agreement between the ASQ and standardised assessments such as the Bayley Scales of Infant Development (Bayley, 1969) was 85%, ranging from 76% for the 4 month ASQ to 91% for the 36 month ASQ. In addition, the measure's sensitivity, its ability to detect delayed development, and its specificity, its ability to correctly identify typically developing children, was also in keeping with standards in the literature which identify acceptable levels of sensitivity and specificity for developmental screening tests at 70% and 80% respectively (Barnes, 1982, as cited in Duby et al., 2006). Other studies have also found evidence to suggest that the ASQ is a valid screening tool (Gollenberg et al., 2009; Skellern, Rogers, and O'Callaghan, 2007). Overall the literature suggests that there is considerable agreement between the ASQ and standardised measures that are conducted by professionals.

6. Summary and Conclusion

This paper investigates the effectiveness of investment in an early childhood program from *in utero* to 18 months of age on key indicators of early skill formation, and on parenting skills. We find significant effects of the program on parenting (specifically the *Home Observation for Measurement of the Environment*, the *Supplement to the HOME for Impoverished Families* and *Interactions with Baby*). Although, these results were weakened to some extent when we corrected for attrition bias, a similar pattern of effects emerged which reinforces the original findings. Overall, with respect to child development, we find little evidence of a statistically significant effect, a result which is consistent with previous evaluations of home visiting programs which have examined early child outcomes (U.S. Department of Health and Human Services, 2009; Gomby, Curloss, & Behrman, 1999). The estimated impact of the program on child development was more precise when we corrected for attrition bias, suggesting that the few identified effects in the original analyses represent a true, tangible

impact on child development at 18 months. This is a promising result given the Smitsman and Corbetta (2010) finding that developmental advances and delays are extremely difficult to detect in very young children.

Our analysis suggests that home visiting programs can be an effective means of improving deficits in early parenting skills and the home environment within a relatively short time frame. In home visiting programs such as PFL, parents are conceived as 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 materials, 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. The theory on human skill formation points to a skill multiplier effect (Cunha and Heckman, 2007), and there may be periods of latency.

In addition, the lack of sizable effects on key dimensions of child development may be attributable to dosage and timing. Recall that the average PFL participant began engaging with the program half way into their pregnancy (21 weeks) and had received, on average, 27 home visits between program entry and 18-months. It is possible that this small window of intervention 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).

From a methodological perspective, a naive evaluation strategy (which examines each outcome individually, and calculates the proportion of outcomes for which a significant

difference is found) would indicate a significant effect for 25 percent of outcomes (6/24) at six months, 7 percent of outcomes (1/14) at 12 months, and 16 percent of outcomes (5/30) at 18 months. This could be intrepreted 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"[15]. 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 the literature. 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 small differences we have identified between the low and high treatment groups could potentially result in large returns over time.

This study informs the policy discussion on the effectiveness of early investment in the world context. 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[16]. Further analysis of later waves of outcome data will be examined to understand the true effectiveness of home visiting programs in non-US settings.

[15] 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".

[16] 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

Achenbach, T.M., McConaughy, S.H., & Howell, C.T. 1987. Child/adolescent behavioural and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101,* 213-232.

Alpern, G. D. (2007). *Developmental profile – 3.* Los Angeles, CA: Western Psychological Services.

American Academy of Pediatrics. (1991). *The Framingham Safety Survey*. Elk Grove Village, IL: American Academy of Pediatrics.

Ammerman, R. T., Putnam, F. W., Bosse, N. R., Teeters, A. R., & Van Ginkel, J. B. (2010). Maternal depression in home visitation: A systematic review. *Aggression and Violent Behaviour, 15* (3), 191-200.

Anisfeld, E., Sandy, J., & Guterman, N. B. (2004). *Best Beginnings: A randomized controlled trial of paraprofessional home visiting program: Narrative Summary.* Report to the Smith Richardson Foundation and New York State Office of Children and Family Services. New York: Columbia University School of Social Work.

Avellar, S., & Paulsell, D. (2011). *Lessons learned from the home visiting evidence of effectiveness review*. Princeton, NJ: Mathematica Policy Research.

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

Barlow, J., Davis, H., McIntosh, E., Jarrett, P., Mockford, C., & Stewart-Brown, S. (2007). Role of home visiting in improving parenting and health in families at risk of abuse and neglect: Results of a multicentre randomized controlled trial and economic evaluation. *Archives of Disease in Childhood, 92* (3), 229-233.

Bates, J.E., Freeland, C.B., & Lounsbury, M.L. (1979). Measurement of infant difficulties. *Child Development*, 50, 794-803.

Bavolek, S. J., & Keene, R. G. (1999). *Adult-adolescent parenting inventory - AAPI-2: Administration and development handbook.* Park City, UT: Family Development Resources, Inc.

Bayley, N. (1969). *The Bayley Scales of Infant Development.* New York: Psychological Corporation.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57* (1), 289-300.

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

Black, M.M., Nair, P., Kight, C., Watchel, R., Roby, P., & Schuler, M. (1994). Parenting and early development among children of drug abusing women: Effects of home intervention. *Pediatrics, 94* (4), 440–448.

Bloom, H.S. (2005). Randomizing groups to evaluate place-based programs. In Bloom HS (ed.) *Learning More From Social Experiments: Evolving Analytic Approaches.* New York, NY: Russell Sage Foundation, 115–172.

Briggs-Gowan, M. J., & Carter, A. S. (2006). *BITSEA brief infant-toddler social and emotional assessment. Examiner's manual*. San Antonio, TX: Harcourt Assessment, Inc.

Brooks-Gunn, J., Berlin, L. & Fuligni, A. (2000). 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.

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

Caughy, M.O., Huang, K., Miller, T., & Genevro, J.L. (2004). The effects of Healthy Steps for Young Children Program: Results from observations of parenting and child development. *Early Childhood Research Quarterly, 19* (4), 611–630.

Centers for Prevention of Child Neglect. (2000). *Community support inventory*. South Bend, IN: Centers for the Prevention of Child Neglect, University of Notre Dame.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Culp, A.M., Culp, R.E., Blankemeyer, M., & Passmark, L. (1996). Parent education home visitation program: Adolescent and nonadolescent mother comparison after six months of intervention. *Infant Mental Health Journal, 19* (2), 111-123.

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

Doyle, O. (2013). Breaking the cycle of deprivation: An experimental evaluation of an early childhood intervention. *Journal of the Statistical and Social Inquiry Society of Ireland*. Vol. XLI, 92-111.

—, Finnegan, S., McNamara, K.A. (2012). Differential Teacher and Parent Ratings of School Readiness in a Disadvantaged Community. *European Early Childhood Education Research Journal, 20*(3), 371-389.

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

— , Hickey C. The Challenges of Contamination in Evaluations of Childhood Interventions. *Evaluation*. 19(2), 180-191.

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

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

— , Saias, T., Tubach, F., & Brand, T. (2011). Engagement in experimental home visiting interventions: A cross-national study on the attributes of parental engagement. Under submission at *American Journal of Community Psychology*.

Duby, J. C., Lipkin, P. H., Macias, M. M., Wegner, L. M., Duncan, P., Hagan, J. F. . . Capers, M. (2006). Identifying infants and young children with developmental disorders in the medical home: An algorithm for developmental surveillance and screening. *Pediatrics*, *118*, 405–420.

Duggan, A. K., McFarlane, E. C., Windham, A. M., Rohde, C. A., Salkever, D. S., Fuddy, L., Rosenberg, L. A., Buchbinder, S. B., & Sia, C. C. J. (1999). Evaluation of Hawaii's Healthy Start Program. *The Future of Children, 9*(1), 66-90.

Ertem, I.O., Avni-Singer, A.J., & Forsyth, B.W.C. (1996). *Supplement to the HOME scale for impoverished families*. New Haven, CT: Yale University.

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000). Short-form versions of the MacArthur Communicative Development Inventories. *Applied Psycholinguistics, 21* (1), 95-115.

Gollenberg, A. L., Lynch, C. D., Jackson, L. W., McGuinness, M., & Msall, M. E. (2009). Concurrent validity of the parent-completed Ages and Stages Questionnaires, 2nd Ed. with the Bayley Scales of Infant Development II in a low-risk sample. *Child Care Health and Development*, *36*(4), 485-490.

Gomby, D.S., Curloss, P.L., & Beherman, R.E. (1999). Home visiting: Recent program evaluations: Analysis and recommendations. *The Future of Children, 9* (1), 4-26.

Good, P. (2005). *Permutation, parametric and bootstrap tests of hypotheses* (3rd ed.), New York: Springer.

Gravetter, F.J ., & Wallnau, L. B. (2008). *Statistics for the behavioral sciences* (8th ed.). Belmont, CA: Thomson/Wadsworth Learning.

Guyer, B., Barth, M., Bishai, D., Caughy, M., Clark, B., Burkom, C., & Tang, C. (2003). *The Healthy Steps for Young Children Program National Evaluation*. Baltimore: Johns Hopkins Bloomberg School of Public Health.

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

Heckman, J.J., Holland, M., Oey, T., Olds, D., Pinto, R., & Rosales, M. (2012). A reanalysis of the Nurse Family Partnership Program: the Memphis Randomized Control Trial. *Unpublished manuscript*. University of Chicago.

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

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

Howe A., Keogh-Brown M., Miles S., & Bachmann M. (2007). Expert consensus on contamination in educational trials elicited by a Delphi exercise. *Medical Education* 41: 196–204.

Janus, M., Duku, E.K., & Stat, P. (2005). *Development of the Short Early Development Instrument (S-EDI).* Report for the World Bank.

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

Jungman, T., Ziert, Y., Kurtz, V., & Brand, T. (2009). 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*, *3* (3), 292-298.

Kahn, J., & Moore, K.A. (2010). *What Works for Home Visiting Programmes: Lessons from Experimental Evaluations of Programmes and Interventions?* Washington DC: Child Trends. Retrieved from: www.childtrends.org/Files/Child_Trends-2010_7_1_FS_WWHomeVisitpdf.pdf.

Kitzman, H., Olds, D. L., Henderson, C. R., Hanks, C., Cole, R., Tatelbaum, R., ... Barnard, K. (1997). Effect of prenatal and infancy home visitation by nurses on pregnancy outcomes, child injuries and repeated childbearing. A randomised controlled trial. *Journal of the American Medical Association, 278* (8), 644–652.

Knudsen, E.I., Heckman, J.J., Cameron, J., & Shonkoff, J.P. (2006). 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., Anderson, N. L., Brecht, M. L., Verzemnieks, I., Lesser, J., & Kim, S. (2002). Public health nursing care for adolescent mothers: Impact on infant health and selected maternal outcomes at 1 year postbirth. *Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine*, *30*(1), 44–54.

Lang, D.L., DiClemente, R.J., Hardin, J.W., Crosby, R.A., Salazar, L.F., & Hertzberg, V.S. (2009). Threats of cross-contamination on effects of a sexual risk reduction intervention: Fact or fiction. *Prevention Science* 10:270-275.

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

Lee, E., Mitchell-Herzfeld, S.D., Lowenfels, A.A., Greene, R., Dorabawila, V., & DuMont, K.A. (2009). Reducing low birth weight through home visitation: A randomized controlled trial. *American Journal of Preventive Medicine, 36* (2), 154-160.

Little, R.J.A., & Rubin, D.B. (2002). *Statistical Analysis with Missing Data, 2nd Edition*. New Jersey: Wiley-interscience.

MacPhee, D. (1981). *Knowledge of infant development inventory.* Unpublished questionnaire and manual, University of North Carolina, Department of Psychology, Chapel Hill, NC.

Nelson, C.A. (2000). The neurobiological bases of early intervention. In J. P. Shonkoff & S.J. Meisels (Eds.), *Handbook of Early Childhood Intervention*, Second Edition (pp. 204-227). Cambridge, MA: Cambridge University Press.

Olds, D.L. (2002). Prenatal and infancy home visiting by nurses: from randomized trials to community replication. *Preventive Science 3*(3), pp.153-172.

Olds, D.L., Henderson, C.R., Cole, R., Eckenrode, J., Kitzman, H. & Luckey, D. (1998). 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*, *280* (14), 1238-1244.

— , Kitzman, H., Cole, R., & Robinson, J. (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.

— , Eckenrode, J., Henderson C.R., Kitzman, H., Powers, H., Cole, R., …Luckey, D. (1997). Long-term effects of home visitation on maternal life course and child abuse and neglect. *Journal of the American Medical Association*, *278* (8), 637-643.

— , Robinson, J., O'Brien, R., Luckey, D.W., Pettitt, L.M., Henderson, C.R., …Talmi, A. home visiting by paraprofessionals and by nurses: A randomized, controlled trial. *Pediatrics*, *110* (3), 486-496.

Pearlin, L. I., & Schooler, C. (1978). The structure of coping. *Journal of Health and Social Behaviour*, *19* (1), 2-21.

Pesarin, F., & Salmaso, L. (2010). *Permutation Test for Complex Data: Theory, Applications and Softwares*. Wiley.

Pocock, S.J., Hughes, M.D., & Lee, R.J. (1987). Statistical problems in the reporting of clinical trials. *New England Journal of Medicine*, *317* (7), 426-432.

Preparing for Life, & The Northside Partnership. (2008). *Preparing for Life Programme Manual*. Dublin: Preparing for Life and the Northside Partnership. Retrieved from http://www.preparingforlife.ie/images/Reports/pfl_manual_may_2008.pdf.

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

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

Sanders, M., Markie-Dadds, C., & Turner, K. (2003). Theoretical, scientific and clinical foundations of the Triple P-Positive Parenting Program: A population approach to the promotion of parenting competence. *Parenting Research and Practice Monograph*, *1*, 1-21.

Sanders, M.R., Markie-Dadds, C., Tully, L., & Bor, W. (2000). The Triple P-Positive Parenting Program: A comparison of enhanced, standard, and self directed behavioral family intervention for parents of children with early onset conduct problems. *Journal of Consulting and Clinical Psychology, 68,* 624-640.

Siegel, E., Bauman, K. E., Schaefer, E. S., Saunders, M. M., & Ingram, D. (1980). Hospital and home support during infancy: impact on maternal attachment, child abuse and neglect and health care utilisation. *Pediatrics, 66,* 183–90.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. Boca Raton, FL: Chapman & Hall/CRC.

Schuler, M. E., Nair, P., Black, M. M. (2002). Ongoing Maternal Drug Use, Parenting Attitudes, and a Home Intervention: Effects on Mother-Child Interaction at 18 Months. *Developmental and Behavioral Pediatrics*, *23*(2), 87-94

Shonkoff, J.P., & Phillips, D.A. (2000). *From neurons to neighborhoods: The science of early childhood development*. Washington, DC: National Academies Press.

Skellern, C. Y., Rogers, Y., & O'Callaghan, M.J. (2001). A parent completed developmental questionnaire: A follow up of ex-premature infants. *Journal of Pediatric Child Health, 37,* 125-129.

Smitsman, A.W., & Corbetta, D. (2010). 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.) (pp. 167-203). West Sussex: Blackwell Publishing.

Squires, J., Bricker, D., & Twombly, E. (2003). *The ASQ:SE user's guide*. Baltimore, MD: Brookes Publishing.

Squires, J.K., Potter, L., & Bricker, D. (1997). Revision of a parent completed developmental screening tool: Ages and Stages Questionnaires. *Journal of Pediatric Psychology, 22*(3), 313-328.

Squires, J.K., Potter, L., & Bricker, D. (1999). *The Ages and Stages Questionnaire users guide*. Baltimore, MD: Brookes Publishing.

Stabile, I., & Graham, M. (2000). *Florida Panhandle Healthy Start: A randomized trial of prenatal home visitation*. Tallahassee, FL: FSU Center for Prevention and Early Intervention Policy.

Stolk, M.N., Mesman, J., van Zeijl, J., Alink, L.R.A., Bakermans-Kranenburg, M.J., van Ijzendoorn, M.H.,…Koot, H.M. (2008). Early parenting intervention: Family risk and first-time parenting related to intervention effectiveness. *Journal of Child and Family Studies, 17,* 55-83.

Sumner, G., & Spietz, A. (1994). *Caregiver/parent-child interaction teaching manual,* Seattle, WA: NCAST Publications.

The Panel Study of Income Dynamics. (2010). The Panel Study of Income Dynamics Child Development Supplement: User Guide Supplement for CDS-I. Ann Arbor, MI: Institute for Social Research. Retrieved from: http://psidonline.isr.umich.edu/CDS/CDS1_UGSupp.pdf

Torgerson, D.J. (2001). Contamination in trials: Is cluster randomization the answer? *British Medical Journal* 322: 355–357.

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

U.S. Department of Health and Human Services. (2009). Home Visiting Evidence of Effectiveness (HomVee). Retrieved 21st December, 2012 from http://homvee.acf.hhs.gov/programs.aspx.

Wagner, M., & Spiker, D. (1996). *Intervention in support of adolescent parents and their children: A final report on the teen Parents as Teachers demonstration.* Menlo Park, C.A.: SRI International. Retrieved from http://policyweb.sri.com/cehs/publications/TeenParentFinalReport.pdf.

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

Westfall, P.H., & Wolfinger, R.D. (1997). Multiple tests with discrete distributions. *The American Statistician, 51* (1), 3-8.

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

Web Appendix

Appendix A

Literature Review of Home Visiting Program Evaluations

Table A1: Evaluations of Home Visiting Program: Outcomes examined before and up to 18 months of age

Outcome	Author	Sample Size	Programme	Measures used	Sig. Finding	Effect	Timing of Followup
Child Development	Olds et al., (2002)	543-605	Nurse Family Partnership (NFP)	Infant vulnerability (fear stimuli), Infant low vitality (joy stimuli, anger stimuli), Irritable temperament	Infant vulnerability: fear stimuli (6 months)	Favourable (6 months)	6 months
	Anisfeld et al., (2004)	~350	Healthy Families America	ASQ (Ages & Stages Questionnaire), Bayley Scales of Child Development	None (6 months, 12 months)	None (6 months, 12 months)	6 months, 12 months
	Duggan et al. (1999)	564	Healthy Families America	Bayley Scales of Child Development, Mental Development Index	None (12 months)	None (12 months)	12 months
	Jungman et al., (2011)	755 (6 months) 367 (12 months)	ProKind	Cognitive development (MDI), Psychomotor development (PDI)	Cognitive development (6 months) None (12 months)	Favorable (6 months) None (12 months)	6 months, 12 months)
	Wagner et al. (1996)	236	Parents as Teachers (PAT)	Cognitive Development (DPII)	None (12 months)	None (12 months)	12 months
	Barlow et al. (2007)	131	Family Partnership Model	Social and Emotional Adjustment (BITSEA), Development (BSID), Temperament (ITS)	None (12 months)	None (12 months)	12 months
	Black, Nair, Kight, Wachtel, Roby &Schuler, (1994).	43	Home Visiting for drug abusing mothers.	Bayley Scales of Infant Development	None (18 months)	None (18 months)	18 months
	Wasik, Ramey, Bryant &	62	Project CARE	Bayley Scales of Infant development	None (18 months)	None (18 months)	18 months

	Sparling (1990)						
	Caughy, Huang, Miller & Genevro. (2004)	378	Healthy Steps	Attachment Security Q-Sort	None (18 months)	None (18 months)	16-18 months
	Drotar, Robinson, Jeavons & Kirchner. (2009)	364	Parents as Teachers (PAT)	Attachment Security Q-Sort	None (18 months)	None (18 months)	18 months
	Roggman, Boyce & Cook, (2009)	161	Early Head Start- Home Visiting	Attachment Security Q-Sort	Attachment security (18 months)	Favourable (18 months)	18 months
Positive Parenting Practices	Anisfeld et al., (2004)	~350	Healthy Families America	NCAST (Nursing Child Assessment Satellite Training)	None (6 months, 12 months)	None (6 months, 12 months)	6 months, 12 months
	LeCroy & Crysik, (2011)	~180	Healthy Families America	Revised Parent-Child Conflict Tactics Scale, Adolescent Parenting Inventory (AAPI-2), Safety practices, Mother's reading	AAPI-2 (Oppressing child's independence), Safety practices (6 months) Revised Parent-Child Conflict Tactics Scale (Verbal Aggression, Minor Corporal Aggression) (12 months)	Favourable (6 months, 12 months)	6 months, 12 months
	Duggan et al. (1999)	564	Healthy Families America	HOME Inventory (learning environment), NCAST (Nursing Child Assessment Satellite Training)	None (12 months)	None (12 months)	12 months
	Duggan, McFarlane et al. (2004)	558	Healthy Families America	Discipline strategies (CTS-PC), HOME Inventory	None (12 months)	None (12 months)	12 months
	Mitchell-Herzfeld et al. (2005)	1061	Healthy Families America	Adult-Adolescent Parenting Inventory (AAPI), Child Safety Checklist, Knowledge of Child Development (KIDI), Parenting Practices (CTS2)	None (12 months)	None (12 months)	12 months
	Culp et al., (2004)	263	Community Based Family	Massachusetts Safety Checklist, Adult-	Parenting knowledge (developmental	Favourable (6 months, 12	6 months, 12 months

			Resource Service Programmes (CBFRS)	Adolescent Parenting Inventory (AAPI), HOME inventory	expectations, noncorporal punishment), HOME inventory (acceptance and organization subscale) (6 months) Massachusetts Safety Checklist, Adult-Adolescent Parenting Inventory (AAPI: developmental expectations, noncorporal punishment); Parenting knowledge , HOME Inventory (acceptance) (12 months)	months)	
	Koniak-Griffin et al. (2002)	102	Early Intervention Program for Adolescent Mothers (EIP)	Nursing Child Assessment Teaching Scale (NCATS: mother's score, total score)	None (6 weeks, 12 months)	None (6 weeks, 12 months)	6 weeks, 12 months
	Wagner & Spiker (2001)	344	Parents as Teachers (PAT)	Parental Knowledge (AAPI, PSOC), HOME Inventory	None	None	12 months
	Wagner et al. (1996)	236	Parents as Teachers (PAT)	HOME Inventory, Parenting Knowledge (KIDI), Sense of Competence (PSOC)	HOME Inventory (total score, parental responsivity, appropriate play materials) (12 months)	Favourable (12 months)	12 months
	Barlow et al. (2007)	131	Family Partnership Model	Maternal Sensitivity and Child Cooperativeness (CARE index), Maternal Psychopathology (GHQ), Parenting Attitudes and Competence (AAPI, WBPB, Parenting Sense of Competence Scale), HOME Inventory,	Maternal Sensitivity and Child Cooperativeness (CARE index) (12 months)	Favourable (12 months)	12 months
	Johnston et al., (2004)	257	Healhty Steps	Played with baby daily, Parenting sense of competence, Role satisfaction, KIDI score, Knowledge of sleep positions, Endorsed	None (3 months)	None (3 months)	3 months

				appropriate discipline, Home safety index, Safe-sleep practices, Read with infant in past week, Self-efficacy			
	Minkovitz et al., (2001)	1987	Healthy Steps	Car in back seat, Lowered water temperature., Gave baby cereal, Showed picture books daily, Followed at least 2 routines, Played with baby daily	None (2-4 months)	Favourable (2-4 months)	2-4 months
	Shute & Judge, (2005)	359	Starting Well	HOME inventory (total score)	None (6 months)	None (6 months)	6 months
	Armstrong et al., (1999)	181	Family Care	HOME inventory	HOME (multiple subscales & total score) (6 weeks)	Favourable (6 weeks)	6 weeks
	Black et al., (1994)	43	Home Visiting for drug abusing mothers.	Parenting Stress Index (PSI), HOME Inventory	HOME Inventory(Total score, Maternal Responsivity subscale and Variety Subscale (18 months)	Favorable (18 months)	18 months
	Wiggins et al. (2004)	493 (328 control, 165 SHV)	Social Support and Family Health Study	Experiences of looking after baby, (easy/difficult), views on child development	None (14 months)	None (14 months)	14 months
	Caughy et al. (2004)	378	Healthy Steps	Nursing Child Assessment by Satellite Training (NCAST) total score, Parent/Caregiver Involvement Scale (P/CIS), HOME Inventory	None (16 – 18 months)	None (16 – 18 months)	16-18 months
	Caughy, Miller, Genevro, Huang & Nautiyal. (2003)	378	Healthy Steps	Parental Responses to Child Misbehaviour – modified	Use of inductive/authoritative discipline (16-18 months)	Favourable (16 -18 months)	16-18 months
	Larson (1980)	115	Pre/Post natal Home Visiting	Maternal Behaviour Scale, HOME Inventory	Maternal Behaviour Scale, HOME (Total score, also provision of play materials) (18 months)	Favourable (18 months)	18 months
	Mackenzie	294	Starting Well	Home Observation for	Total Home Sore	Favourable	18 months

	et al (2004)			Measurement of the Environment; HOME Inventory			
	Wasik et al., (1990)	62	Project CARE	HOME Inventory	None	None	18 months

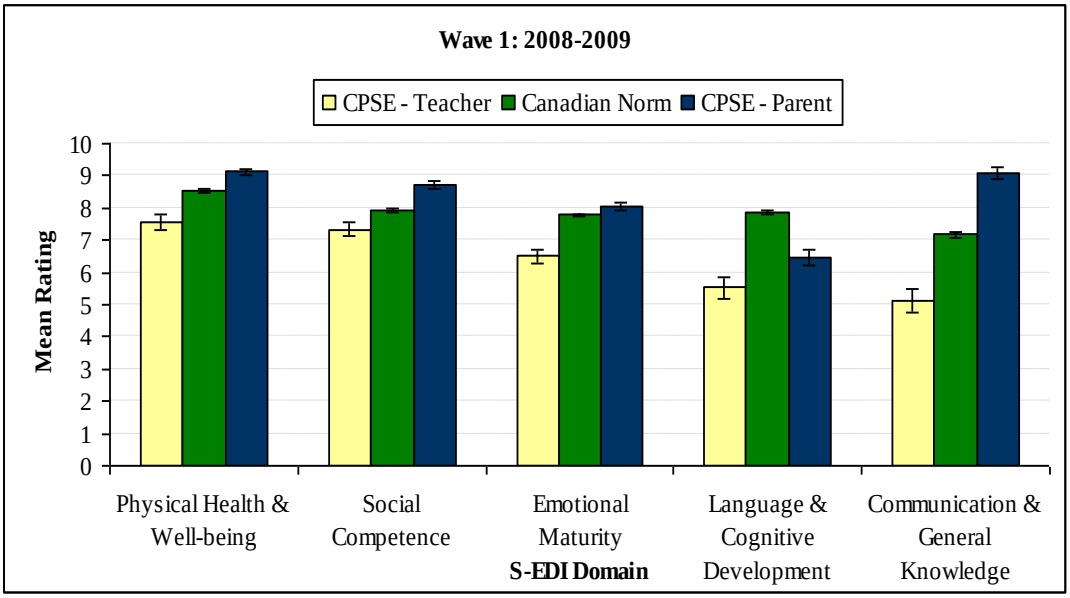
Favourable impact. A statistically significant impact on an outcome measure in a direction that is beneficial for children and parents

Unfavourable or ambiguous impact. A statistically significant impact on an outcome measure in a direction that may indicate potential harm to children and/or parents.

Appendix B

Appendix. B1 PFL Program Background

Table B1: Comparison of S-EDI School Readiness Teacher Ratings in the PFL Study Area and Canadian Norm



Note: CPSE, which represents *Children's Profile at School Entry*, is the assessment of junior infant children's school readiness skills conducted annually in the *PFL* catchment areas.

Appendix B2 PFL Curriculum

Common Supports

Developmental Materials
Families in both the high and low treatment groups receive developmental packs annually to the value of approximately ~€100pa. The first developmental pack includes a number of safety items, such as corner guards, angle latches, and heat sensitive spoons, plus a baby gym/play mat. The second pack includes developmental appropriate toys such as puzzles, activity toys, and bricks. The third pack includes cookery/construction sets, puzzles and memory games. The fourth pack includes a magnetic game, a doctor's case, a lace-up shoe and a tea set. The fifth pack is still under development.

Public Health Information
Both groups are also encouraged to attend two public health workshops in the community. The Stress Control Programme, which is run by external facilitators, involves six one-hour weekly sessions which focuses on enabling individuals to identify how they consciously and subconsciously feed their stress, as well as describing what stress is, and the indicators of stress. The programme also teaches techniques and strategies to manage stress. Participants receive a set of booklets and a relaxation CD. For more details on this programme please see www.glasgowsteps.com.

The second health programme offered is the Healthy Food Made Easy programme, which is facilitated by one of the mentors and involves six two hour sessions. The aim of the programme is to improve nutritional knowledge, attitudes and behaviour by learning about basic nutritional theories and participating in practical cookery sessions. It is a peer led programme which emphasises group learning through discussion, worksheets and hand-outs, quizzes, problem solving games, food preparation and cookery.

Facilitated Access to Enhanced Preschool
A preschool place for one-year has also been reserved for all *PFL* children in the local childcare centres. *PFL* covers the cost of this for those families experiencing exceptional financial difficulty. However, it should be noted that all *PFL* children will now be eligible for a new Government scheme which provides every three year old child in Ireland with access to a free preschool place for one year.

Access to a Support Worker
Participants are given a directory of local services and have access to a *PFL* support worker who can help them connect to additional community services if needed. Details about coffee mornings and other community events are sent via group text or online. Finally, both treatment groups receive a framed professional photograph of their child as well as programme newsletters and special occasion (e.g., birthday) cards.

It should be noted that families in the low treatment group have access to an Information Officer who acts as the point of contact for parents; while families in the high treatment group have a home visiting mentor assigned to each household. The role of the information officer is to meet

with the families in the low treatment group before birth and contact the family at various intervals, such as when sending developmental packs, and when the child is due to begin childcare. Families in the low treatment group may contact the information officer at any time with queries regarding public services-as-usual for their child or queries regarding PFL events. However, the Information Officer may not provide the participants in the low treatment group with any information related to parenting or child development.

Additional High Treatment Supports

Home Visits from a Trained Mentor

Participants in the high treatment group avail of a home-visiting mentor support service. Each family has an assigned mentor who visits the home for between 30 minutes and two hours starting during pregnancy and continuing until the child starts school. Originally, it was anticipated that each family would receive a weekly home visit. However, early on in the implementation process it became evident that weekly home visits were not achievable from the families' point of view. Therefore the programme changed this weekly requirement, such that the frequency of the visits depends on the needs of the families, with the majority of families receiving fortnightly visits, and some monthly.

The home visits are facilitated by trained mentors with a cross section of professional backgrounds including education, social care, youth studies, psychology, and early childcare and education. Although the professional qualifications of the mentors are diverse, each mentor completed extensive training on the *PFL* Programme Manual. The role of the mentors is to build a good relationship with parents, provide them with high quality information and to be responsive to issues that arise. Through these efforts the *PFL* Programme aims to enable parents to make informed choices and connect them to other community services (*Preparing for Life* & The Northside Partnership, 2008). The mentors focus on five general areas related to child development: 1) pre-birth; 2) nutrition; 3) rest and routine; 4) cognitive and social development; and 5) mother and her supports. These areas were selected during the development phase as they were highlighted as areas of need in this community.

The aim of the home visits is to support and help the parents with key parenting issues using a set of *PFL* developed Tip Sheets. The Tip Sheets are colourful representations of information related to child development presented in a clear, concise manner and were developed by *PFL* staff based on available information from local organisations such as the Health Service Executive, the Department of Health and Children, and Barnardos Children's Charity. The Tip Sheets were designed at a reading level of a 12 year old and are used to facilitate the home visiting sessions. The Tip Sheets are given to the participant after discussion with the mentor and remain with the participant to serve as an on-going parenting resource. The Tip Sheets are designed to be delivered based on the age of the child and the needs of the family, however, the participants must have received the full set of Tip Sheets by the end of the programme.

Participants in the high treatment group can also avail of baby massage through individual or group sessions with one of the mentors until their baby is approximately 10 months old. There are three individual baby massage sessions and four group-based baby massage sessions,

followed by a refresher session. Finally, the high treatment group are invited to coffee mornings hosted by the mentors.

Triple P Group Parent Training

Secondly, participants in the high treatment supports group also participate in group parent training using the Triple P Positive Parenting Programme (Sanders, Markie-Dadds, & Turner, 2003). Triple P aims to improve positive parenting through the use of videos, vignettes, role play, and tip sheets in a group-based setting for eight consecutive weeks (two hours per week for the first four weeks followed by three weeks of phone support and a final two hour group session on week eight). The group-based component of the Triple P programme has been subject to multiple rigorous evaluations which have demonstrated positive effects for both parents and children (Sanders, Markie-Dadds, Tully, & Bor, 2000). The Triple P programme is delivered to participants in the high treatment group when their children are at least two-years old. As this paper examines child outcomes at six, 12 and 18 months of age, we cannot yet examine the impact of the Triple P component.

Appendix B3 Example of Tipsheets

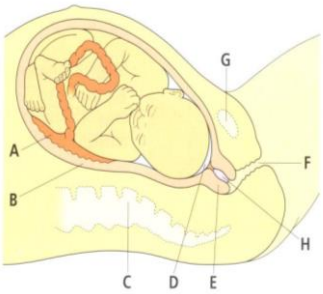
Labour

Labour is the process of giving birth

Every birth follows its own pattern and timetable, and usually takes between four and twelve hours.

The baby in the drawing is ready to be born. It is lying in the correct position with head downwards.

A placenta (afterbirth)
B uterus (womb)
C backbone
D amniotic fluid
E cervix (neck of womb)
F vagina
G pelvis (hip bone)
H plug of mucus



The three stages of labour

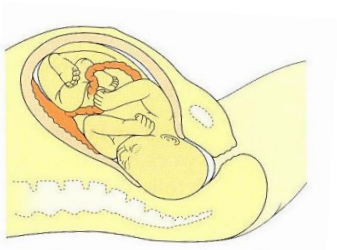
Stage 1: The neck of the womb opens

When you notice one or more of the following signs, you will know that labour has started:

- A show – a bloodstained mucus plug that comes from the opening of the cervix at the bottom of the womb. This may happen 1-3 days before labour begins.
- Your waters break – this is a gush of water from the vagina which shows that the water sac around the baby has burst. This usually means that labour pains will start within the next 24-48 hours. Always inform your midwife or doctor when your waters break.
- The most obvious sign of labour is frequent, painful tummy cramps (labour pains) every 10 minutes or more often. Let your midwife or doctor know when this happens and prepare to go into the hospital.

If you're not sure whether you are in labour or not, ring your maternity hospital, midwife or doctor for advice.

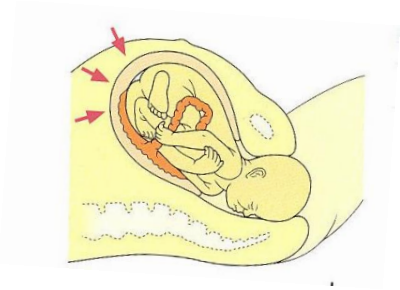
Stage 1 continued: The neck of the womb opens


The first stage of labour

During the first stage of labour, contractions of the muscles in the wall of the womb gradually open the cervix. At some time during this stage, the waters break. The first stage of labour is the longest stage and it comes to an end when the cervix has opened wide enough for the baby's head to pass through. (10 cms)

Stage 2: The baby passes through the birth canal

The womb, cervix and vagina have by now become one birth canal. The contractions are very strong and they push the baby head-first through the birth canal. The mother must help to push.

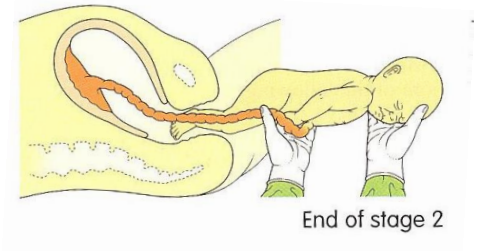


When the baby's head comes out, the midwife may clear mucus from the baby's nose and mouth.

The midwife or doctor now eases the baby's shoulders through the birth canal and the baby slides out into the world.

Stage 3: The baby becomes a separate person

Once the baby is breathing, the umbilical cord is clamped in two places and a cut is made between them. This separates the baby from the mother. Cutting the cord does not hurt the mother or the baby.


End of stage 2

When the baby is first born, the skin is a bluish colour. As soon as the baby starts to breathe the skin quickly turns pink.

The contractions continue until the placenta (afterbirth) is separated from the wall of the womb and has been pushed out through the vagina. The mother may be given an injection to speed up the process. Labour is now completed.

Listening and Talking

Children get better at talking when they are given lots of chances to listen, and also to use words. You can make this fun for yourself and your child.

Things you can do to help your child:

- Listen together and name some of the sounds you hear around you

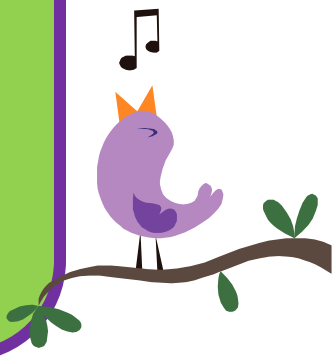
Sounds around us

Indoors:

- ✓ tap running
- ✓ radio and TV
- ✓ baby crying
- ✓ children playing
- ✓ washing machine

Outdoors:

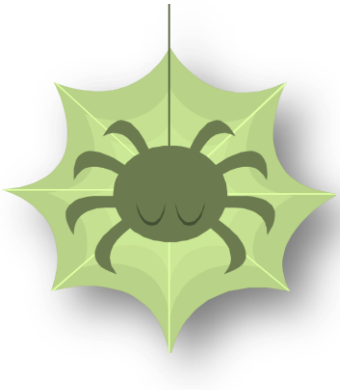
- ✓ plane overhead
- ✓ car, bus, train
- ✓ wind in the trees
- ✓ someone calling
- ✓ birds or insects



- Play 'I hear with my little ear' something that goes *'woof'* (or *'miaow'*.)
- Say an alphabet sound and help your child to find something that starts with that sound, e.g. *b for ball; s for sock; d for doll.*
- Make up rhymes or songs about everyday activities that your child is doing.
- Sing or read nursery rhymes.

Here are two action rhymes you could play with your child:

Show your child how to make a spider with his/her hand, and do actions for the rain falling down and the spider climbing up and falling down.



Incy Wincy spider

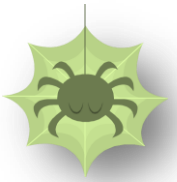
Climbing up the spout

Down comes the rain

And washes poor Wincy out

Out comes the sun

Dries up all the rain



Incy Wincy spider

Climbing up again

Show your child how to bend one arm for the spout, the other for the handle, and do the actions for tipping up and pouring out.

I'm a little teapot

Short and stout

Here's my handle

Here's my spout

When the tea is ready

Hear me shout

'Pick me up and

Pour me out!'

Appendix C

Baseline Analysis

Table C1: Baseline comparison of the high treatment and low treatment groups: Parental demographics & SES indicators (continuous variables)

Variable	N (n_{LOW}/ n_{HIGH})	M_{LOW} (SD)	M_{HIGH} (SD)	$M_{LOW} - M_{HIGH}$	p
Mother's Age	205 (101/104)	25.30 (5.99)	25.46 (5.85)	-0.16	ns
Teenage Mothers	205 (101/104)	0.20 (0.40)	0.16 (0.37)	0.04	ns
Primiparous Mothers	205 (101/104)	0.50 (0.50)	0.54 (0.50)	-0.04	ns
Number of Biological Children	99 (51/48)	1.80 (1.00)	2.04 (1.20)	-0.24	ns
Mother in a Relationship	205 (101/104)	0.84 (0.37)	0.78 (0.42)	0.06	ns
Mother Married	205 (101/104)	0.18 (0.38)	0.14 (0.35)	0.04	ns
Biological Father's Age	203 (100/103)	27.58 (7.33)	27.47 (6.52)	0.11	ns
Teenage Fathers	203 (100/103)	0.12 (0.33)	0.10 (0.30)	0.02	ns
Mothers with Junior Certificate Qualification or Lower	205 (101/104)	0.40 (0.49)	0.34 (0.47)	0.06	ns
Mothers with Primary Degree	205 (101/104)	0.03 (0.17)	0.03 (0.17)	0.00	ns
Age Mother Left Full-time Education	187 (91/96)	17.43 (3.08)	17.41 (2.53)	0.02	ns
Mothers with Literacy/Numeracy Problems	205 (101/104)	0.29 (0.45)	0.35 (0.48)	-0.06	ns
Fathers with Junior Certificate Qualification or Lower	183 (94/89)	0.48 (0.50)	0.46 (0.50)	0.02	ns
Fathers with Primary Degree	183 (94/89)	0.03 (0.18)	0.03 (0.18)	0.00	ns
Age Father Left Full-time Education	149 (73/76)	16.75 (2.14)	16.61 (2.36)	0.14	ns
Mothers in Paid Work	205 (101/104)	0.40 (0.49)	0.37 (0.48)	0.03	ns
Mothers in Full-time Work	78 (41/37)	0.63 (0.49)	0.51 (0.51)	0.12	ns
Annual Income of Working Mothers (in Euros)	75 (38/37)	19,602 (8,093)	19,224 (9,851)	378	ns

Mothers Unemployed	205 (101/104)	0.41 (0.49)	0.43 (0.50)	-0.02	ns
Fathers in Paid Work	198 (97/101)	0.57 (0.50)	0.49 (0.50)	0.08	ns
Fathers in Full-time Work	102 (56/46)	0.84 (0.37)	0.87 (0.34)	-0.03	ns
Annual Income of Working Fathers (in Euros)	76 (40/36)	25,698 (12,059)	27,208 (13,179)	-1510	ns
Fathers Unemployed	198 (97/101)	0.31 (0.46)	0.43 (0.50)	-0.12	ns
Residing in Social Housing	204 (101/103)	0.55 (0.50)	0.55 (0.50)	0.00	ns
In Possession of a Medical Card	205 (101/104)	0.66 (0.47)	0.60 (0.49)	0.06	ns
In Possession of Private Health Insurance	202 (100/102)	0.07 (0.26)	0.09 (0.29)	-0.02	ns
In Receipt of Social Welfare Payments	203 (99/104)	0.65 (0.48)	0.64 (0.48)	0.01	ns
Saves Regularly	205 (101/104)	0.51 (0.50)	0.47 (0.50)	0.05	ns
Materially Deprived (on at least one item)	202 (100/102)	0.32 (0.47)	0.44 (0.50)	-0.12	ns
Material Deprivation Index	202 (100/102)	0.64 (1.16)	0.91 (1.47)	-0.27	ns

Note. Permutation tests were conducted using regression tests for normally distributed data unless otherwise noted. [a]Permutation tests were conducted using the Wilcoxon Rank-Sum tests for non-normally distributed data.

Table C2: Baseline comparison of the high treatment and low treatment groups: Parental demographics & SES indicators (categorical variables)

Variable	Low % (*n*)	High % (*n*)	*p*
Maternal Ethnicity			ns [a]
Irish	92.08 (93)	95.19 (99)	
Irish Traveller	7.92 (8)	2.88 (3)	
Other	0.00 (0)	1.92 (2)	
Household Income			ns
< €250 per week	4.95 (5)	4.81 (5)	
€250-€500 per week	25.74 (26)	24.04 (25)	
€500-€750 per week	14.85 (15)	21.15 (22)	
€750-€1000 per week	19.80 (20)	16.35 (17)	
> €1000 per week	34.65 (35)	33.65 (35)	
Ability to Make Ends Meet			ns
With Difficulty	29.70 (30)	28.16 (29)	
Get by	33.66 (34)	35.92 (37)	
Easily	36.63 (37)	35.92 (37)	

Note: Chi-square test used unless otherwise noted. [a] Fisher's exact test used.

Table C3: Baseline comparison of the high treatment and low treatment groups: Maternal well-being & personality (continuous variables)

Variable	*N* (n_{LOW}/ n_{HIGH})	M_{LOW} (*SD*)	M_{HIGH} (*SD*)	$M_{LOW} - M_{HIGH}$	*p*
Well-being					
WHO-5 Percentage Score	205 (101/104)	57.94 (22.96)	54.27 (20.24)	3.67	ns
Low WHO-5 Percentage Score	205 (101/104)	0.37 (0.48)	0.42 (0.50)	-0.05	ns
Incidence of Postnatal Depression in Previous Pregnancies	97 (51/46)	0.22 (0.42)	0.17 (0.38)	0.05	ns
Vulnerable Attachment Style Questionnaire (VASQ)					
Insecurity Score	205 (101/104)	7.58 (2.79)	7.82 (2.48)	-0.24	ns
High Insecurity	205 (101/104)	0.40 (0.49)	0.53 (0.50)	-0.13	ns
Proximity Seeking Score	205 (101/104)	10.24 (2.16)	10.42 (2.16)	-0.18	ns
High Proximity Seeking	205 (101/104)	0.93 (0.26)	0.95 (0.21)	-0.02	ns
Total Vulnerable Attachment Score	205 (101/104)	17.82 (3.98)	18.24 (3.77)	-0.42	ns
High Vulnerable Attachment	205 (101/104)	0.70 (0.46)	0.77 (0.42)	-0.07	ns
Pearlin Self Efficacy Scale					
Mastery	205 (101/104)	2.88 (0.60)	2.77 (0.63)	0.11	ns
Lowest 10% Mastery	205 (101/104)	0.08 (0.27)	0.13 (0.33)	-0.05	ns
Parenting Self Efficacy	205 (101/104)	3.17 (0.53)	3.04 (0.55)	0.13	ns
Lowest 10% Parenting Self Efficacy	205 (101/104)	0.07 (0.26)	0.15 (0.36)	-0.08	ns
Total Self Efficacy Score	205 (101/104)	3.02 (0.52)	2.90 (0.52)	0.12	ns
Lowest 10% Total Self Efficacy Score	205 (101/104)	0.08 (0.27)	0.14 (0.35)	-0.06	ns
Rosenberg Self Esteem Scale					
Total Self Esteem Score	205 (101/104)	12.78 (2.86)	12.82 (2.69)	-0.04	ns

Lowest 10% Self Esteem Score	205 (101/104)	0.18 (0.38)	0.13 (0.34)	0.05	ns
Ten Item Personality Inventory (TIPI)					
Extraversion	205 (101/104)	5.17 (1.35)	5.11 (1.34)	0.06	ns
Agreeableness	205 (101/104)	5.79 (1.17)	5.65 (1.16)	0.14	ns
Conscientiousness	205 (101/104)	5.53 (1.23)	5.38 (1.29)	0.15	ns
Emotional Stability	205 (101/104)	4.04 (1.54)	3.83 (1.60)	0.21	ns
Openness to Experience	205 (101/104)	5.09 (1.25)	4.96 (1.26)	0.13	ns
Consideration of Future Consequences (CFC) Scale	205 (101/104)	10.33 (3.18)	9.50 (3.23)	0.83	ns
Indicators of Household Social and Emotional Risk	203 (99/104)	0.70 (1.18)	0.79 (1.08)	-0.09	ns

Note. Permutation tests were conducted using regression tests for normally distributed data unless otherwise noted. [a]Permutation tests were conducted using the Wilcoxon Rank-Sum tests for non-normally distributed data.

Table C4: Baseline comparison of the high treatment and low treatment groups: Maternal health & pregnancy (continuous variables)

Variable	N (n_{LOW}/ $n_{HIGH)}$	M_{LOW} (SD)	M_{HIGH} (SD)	$M_{LOW} - M_{HIGH}$	p
Regular Exercise	205 (101/104)	0.45 (0.50)	0.38 (0.49)	0.06	ns
Health Service Use					
# Health Services Used in Previous Year	205 (101/104)	2.39 (1.25)	2.44 (1.53)	-0.05	ns
# of Non-pregnancy Related GP Visits in Previous Year	200 (100/100)	2.95 (3.56)	3.37 (6.41)	-0.42	ns
The Pregnancy					
Age at First Pregnancy	98 (51/47)	21.39 (4.51)	21.11 (3.75)	0.28	ns
Birth Control Practices	203 (99/104)	0.33 (0.47)	0.33 (0.47)	0.00	ns
Planned Pregnancy	203 (100/103)	0.30 (0.46)	0.29 (0.46)	0.01	ns
Week Pregnancy Confirmed	204 (100/104)	6.56 (3.86)	6.32 (3.56)	0.24	ns
Week of First Antenatal Visit	158 (78/80)	16.77 (5.43)	15.83 (5.77)	0.94	ns
Participation in Antenatal Classes	190 (93/97)	0.33 (0.47)	0.40 (0.49)	-0.07	ns
Health Supplement Use					
Regular Exercise	205 (101/104)	0.45 (0.50)	0.38 (0.49)	0.06	ns
Health Service Use					
# Health Services Used in Previous Year	205 (101/104)	2.39 (1.25)	2.44 (1.53)	-0.05	ns
# of Non-pregnancy Related GP Visits in Previous Year	200 (100/100)	2.95 (3.56)	3.37 (6.41)	-0.42	ns
The Pregnancy					
Age at First Pregnancy	98 (51/47)	21.39 (4.51)	21.11 (3.75)	0.28	ns
Birth Control Practices	203 (99/104)	0.33 (0.47)	0.33 (0.47)	0.00	ns
Planned Pregnancy	203 (100/103)	0.30 (0.46)	0.29 (0.46)	0.01	ns

Week Pregnancy Confirmed	204 (100/104)	6.56 (3.86)	6.32 (3.56)	0.24	ns
Week of First Antenatal Visit	158 (78/80)	16.77 (5.43)	15.83 (5.77)	0.94	ns
Participation in Antenatal Classes	190 (93/97)	0.33 (0.47)	0.40 (0.49)	-0.07	ns
Health Supplement Use					
# Drinks per Week (during pregnancy)	53 (27/26)	2.93 (1.38)	3.15 (1.71)	-0.22	ns
Drug Behaviour					
Ever Used Drugs Before Pregnancy	205 (101/104)	0.15 (0.36)	0.13 (0.34)	0.02	ns
Ever Used Drugs During Pregnancy	205 (101/104)	0.03 (0.17)	0.01 (0.10)	0.02	ns

Note. Permutation tests were conducted using regression tests for normally distributed data unless otherwise noted.
[a]Permutation tests were conducted using the Wilcoxon Rank-Sum tests for non-normally distributed data.

Table C5: Baseline comparison of the high treatment and low treatment groups: *Maternal health & pregnancy (categorical variables)*

Variable	Low *n* (%)	High *n* (%)	*p*
Maternal Health Across the Lifespan			
BMI Scale			ns
Underweight	6.17 (5)	4.55 (4)	
Normal weight	60.49 (49)	56.82 (50)	
Overweight	23.46 (19)	27.27 (24)	
Obese	9.88 (8)	11.36 (10)	
Self Rated Healthy Eating Habits			ns
Healthy	41.58 (42)	33.98 (35)	
Average	45.54 (46)	56.31 (58)	
Unhealthy	12.87 (13)	9.71 (10)	
Maternal Substance Use			
Smoking Status During Pregnancy			ns
Reduced smoking	36.63 (37)	36.54 (38)	
Increased smoking	2.97 (3)	2.88 (3)	
Stopped smoking	12.87 (13)	10.58 (11)	
No change in smoking habits	7.92 (8)	11.54 (12)	
Never smoked	39.60 (40)	38.46 (40)	
Maternal Health Across the Lifespan			
BMI Scale			ns
Underweight	6.17 (5)	4.55 (4)	

Normal weight	60.49 (49)	56.82 (50)	
Overweight	23.46 (19)	27.27 (24)	
Obese	9.88 (8)	11.36 (10)	
Self Rated Healthy Eating Habits			ns
Healthy	41.58 (42)	33.98 (35)	
Average	45.54 (46)	56.31 (58)	
Unhealthy	12.87 (13)	9.71 (10)	
Maternal Substance Use			
Smoking Status During Pregnancy			ns
Reduced smoking	36.63 (37)	36.54 (38)	
Increased smoking	2.97 (3)	2.88 (3)	
Stopped smoking	12.87 (13)	10.58 (11)	
No change in smoking habits	7.92 (8)	11.54 (12)	
Never smoked	39.60 (40)	38.46 (40)	

Note: Chi-square test used unless otherwise noted.

Table C6: Baseline comparison of the high treatment and low treatment groups: *Thoughts about parenting (continuous variables)*

Variable	*N* (n_{LOW}/ n_{HIGH})	M_{LOW} (*SD*)	M_{HIGH} (*SD*)	$M_{LOW} - M_{HIGH}$	*P*
Knowledge of Infant Development Short Form (KIDI-SF) Score	205 (101/104)	69.82 (8.18)	72.25 (7.60)	-2.43	<.05
Lowest 10% KIDI-SF Score	205 (101/104)	0.20 (0.40)	0.11 (0.31)	0.09	Ns
Adult Adolescent Parenting Inventory (AAPI-2)					
Parental Expectations of Children	205 (101/104)	5.48 (1.68)	5.91 (1.61)	-0.43	Ns
Parental Empathy Towards Children's Needs	205 (101/104)	4.04 (2.23)	4.21 (1.98)	-0.17	Ns
Use of Corporal Punishment	205 (101/104)	6.19 (1.68)	6.02 (1.58)	0.17	ns
Parent-child Family Roles	205 (101/104)	4.73 (2.14)	5.17 (1.91)	-0.44	ns
Children's Power and Independence	205 (101/104)	5.16 (1.84)	4.93 (2.07)	0.23	ns
Total AAPI-2 Score	205 (101/104)	5.12 (1.42)	5.25 (1.38)	-0.13	ns
Total Number of Scales At Risk	205 (101/104)	1.31 (1.43)	1.08 (1.38)	0.23	ns
Breastfeeding Intentions					
Breastfed Previous Child	97 (51/46)	0.20 (0.40)	0.15 (0.36)	0.05	ns
Intention to Breastfeed Current Child	186 (92/94)	0.30 (0.46)	0.33 (0.47)	-0.03	ns
Childcare Questions					
Intention to Use Childcare	194 (98/96)	0.60 (0.49)	0.45 (0.50)	0.15	<.05
Age Intend to Start Childcare (in months)	96 (55/41)	6.31 (5.46)	8.66 (6.33)	-2.35	<.05 [a]
Community Services	204 (101/103)	0.57 (0.97)	0.90 (1.12)	-0.33	<.05
Residents Associations	204 (101/103)	0.03 (0.22)	0.07 (0.29)	-0.04	ns
Adult Education Services	204 (101/103)	0.08 (0.34)	0.11 (0.34)	-0.03	ns
Other Useful Services	204 (101/103)	0.32 (0.47)	0.38 (0.49)	-0.06	ns

Total Service Use	204 (101/103)	4.13 (3.13)	4.83 (3.24)	-0.70	ns

Note. Permutation tests were conducted using regression tests for normally distributed data unless otherwise noted.
[a]Permutation tests were conducted using the Wilcoxon Rank-Sum tests for non-normally distributed data.

Table C7: Baseline comparison of the high treatment and low treatment groups: Social support (continuous variables)

Variable	N (n_{LOW}/ n_{HIGH})	M_{LOW} (SD)	M_{HIGH} (SD)	$M_{LOW} - M_{HIGH}$	p
Social Support					
From Partner	166 (85/81)	4.87 (0.53)	4.96 (0.25)	-0.09	ns
From Parents	197 (98/99)	4.57 (0.86)	4.68 (0.75)	-0.11	ns
From Relations	204 (100/104)	4.46 (0.76)	4.43 (0.89)	0.03	ns
From Friends	205 (101/104)	4.39 (0.77)	4.36 (0.81)	0.03	ns
From Neighbours	191 (94/97)	3.15 (1.10)	3.04 (1.12)	0.11	ns
From People in Workplace	78 (41/37)	4.27 (0.87)	4.19 (0.94)	0.08	ns
Service Use					
Emergency Services	204 (101/103)	0.63 (0.70)	0.61 (0.85)	0.02	ns
Health Services	204 (101/103)	1.14 (1.04)	1.20 (1.03)	-0.06	ns
Child/Family Services	204 (101/103)	0.97 (0.92)	1.15 (1.02)	-0.18	ns
Employment Services	204 (101/103)	0.39 (0.77)	0.41 (0.77)	-0.02	ns
Community Services	204 (101/103)	0.57 (0.97)	0.90 (1.12)	-0.33	<.05
Residents Associations	204 (101/103)	0.03 (0.22)	0.07 (0.29)	-0.04	ns
Adult Education Services	204 (101/103)	0.08 (0.34)	0.11 (0.34)	-0.03	ns
Other Useful Services	204 (101/103)	0.32 (0.47)	0.38 (0.49)	-0.06	ns
Total Service Use	204 (101/103)	4.13 (3.13)	4.83 (3.24)	-0.70	ns

Note. Service use only was assessed in the *PFL* communities. Permutation tests were conducted using regression tests for normally distributed data unless otherwise noted

Table C8: Baseline comparison of the high treatment and low treatment groups: *social support (categorical variables)*

Variable	Low % (*n*)	High % (%)	*p*
Frequency of Meeting Friends/Relatives			ns[a]
Regularly	60.40 (61)	67.31 (70)	
Sometimes	36.63 (37)	26.92 (28)	
Rarely/Never	2.97 (3)	5.77 (6)	
Number of Neighbours Known Personally			ns
None	11.88 (12)	8.65 (9)	
Few: 1-6	35.64 (36)	37.50 (39)	
Many: 7+	52.48 (53)	53.85 (56)	
Satisfaction with Neighbourhood			ns
Dissatisfied	9.90 (10)	12.50 (13)	
Neither Satisfied or Dissatisfied	24.75 (25)	15.38 (16)	
Satisfied	65.35 (66)	72.12 (75)	

Note: Chi-square test used unless otherwise noted. [a] Fisher's exact test used.

Appendix D. Standardized Scales Uses to Measure Child and Parent Outcomes

Child Development

We focus on six child development instruments: the *Ages and Stage Questionnaire* (ASQ; Squires et al., 1999); the *Ages and Stages Questionnaire: Social-Emotional* (ASQ:SE; Squires et al., 2003); an assessment of temperament based on the *Infant Characteristics Questionnaire* (Bates et al., 1979); the *MacArthur-Bates Communicative Development Inventories: Words and Gestures, Short Form (CDI-WG:* Fenson et al., 2000*),* the *Brief Infant-Toddler Social and Emotional Assessment* (*BITSEA*; Briggs-Gowan and Carter, 2006); and finally the *Developmental Profile 3, Cognitive Section* (*DP-3*; Alpern, 2007).

Ages and Stage Questionnaire

Using the ASQ questionnaire that specifically relates to child development at six months of age, mothers were asked to indicate whether their children could perform thirty different tasks. There are three possible answers to each question: *yes* (10 points), *sometimes* (5 points), and *no* (0 points). These items are divided into five sub domains entitled *communication*, *gross motor*, *fine motor*, *problem solving*, *personal-social* each with possible scores ranging from zero to 60. Higher scores are indicative of more advanced child development.

Ages and Stage Questionnaire: Social-Emotional

The infant's social-emotional behaviour was assessed using the ASQ:SE questionnaire, a supplement to the traditional ASQ. The 19 items in this questionnaire relate to the child's tendency towards the following behaviours: to calm and settle down, to accept direction, to communicate feelings, to cope with physiological needs (e.g. sleeping, eating), to respond without guidance (move to independence), to demonstrate feelings for others, to initiate social responses to parents and others. The mother indicates whether her child exhibits each behaviour *most of the time* (10 points), *sometimes* (5 points), or *never* (0 points). Additionally, the mother was asked whether the particular behaviour was a concern for her. If the mother answered *yes* to any item, an additional 5 points was added resulting in a total possible score ranging from zero to 285. Higher scores indicate a more negative outcome.

Difficult Temperament

The mothers were asked seven questions taken from the Quebec Longitudinal Study of Child Development (QLSCD) which are based on the Infant Characteristics Questionnaire (Bates, Freeland, & Lounsbury, 1979). The questions relate to whether the child gets upset easily or demonstrates fussy behaviour. Each item is scored on a six point scale resulting in a total possible range of one to 42, where higher scores are indicative of a more difficult temperament.

MacArthur-Bates Communicative Development Inventories: Words and Gestures (CDI-WG)

CDI-WG is a parent report instrument for assessing language and communication skills in children. It provides norms for children aged 8 to 16 months of age. The CDI inventories measure a range of early communicative and representational skills that are related to language development in typically developing and language-delayed children. The CDI-WG consists of

three sections; first signs of understanding, First communicative gestures, and a vocabulary checklist.

Participants were asked to complete the CDI-WG with pen and paper before beginning the main part of the interview. The first section, 'first signs of understanding', contained 3 questions with the response options *yes* or *no*. In the second section, 'first communicative gestures', there were 12 questions with the response options *not yet, sometimes* and *often.* The final section contained an 89-word 'vocabulary checklist' with 3 separate columns; *understands, understands and says* and *does not understand or say*. *First signs of understanding,* and *first communicative gestures* each produced a summed raw score. The vocabulary checklist columns 'understands*'* and 'understands and says' provided 2 scores: *words understood* and *words produced.* These were then normed by age and gender, according to Fenson et al. (2000*).* In total, the CDI produces four scores.

Brief Infant-Toddler Social and Emotional Assessment (BITSEA)

The BITSEA is a 42-item screening tool for social-emotional/behavioural problems and delays in competence in children aged twelve months to thirty-six months. This version is a shortened version of the *Infant-Toddler Social and Emotional Assessment* (ITSEA). The BITSEA yields a *problem score* and a *competence score*. *Problem score* items include externalising (6 items), internalising (8 items) and dysregulation problems (8 items). *Competence score* items include areas of attention, compliancy, mastery, motivation, pro-social peer relations, empathy, play skills and social relatedness (11 items). The interviewer asked participants to verbally rate each item on a 3 point scale (0 = *not true/rarely*, 1 = *somewhat true/sometimes*, 2 = *very true/often*). Items were summed to obtain a total score, with higher *Problem scores* indicating greater levels of social-emotional or behavioural problems and lower *Competence scores* indicating possible delays/deficits in competence. These scores were normed by child gender.

Developmental Profile 3- Cognitive Section

The DP-3 is a parent report measure of child development from birth to age 12 years and 11 months. The *PFL* evaluation included the DP-3 cognitive section which measures cognitive abilities in an indirect manner. This is a 38 item scale, starting at number 1 and continuing until the stop rule is satisfied (i.e. when five consecutive no responses are recorded). Each of the items refer to tasks which require cognitive skill and are arranged in order of difficulty, for example, 'When an adult points to something, does the child usually look where the adult has pointed?'. For each item, participants were asked whether their child had carried out the task and responded *yes* or *no* accordingly. The *Yes* responses were tabulated to create a continuous score whereby higher values indicated greater cognitive development.

Parenting

Parenting behavior is examined using three standardized scales: a measure of parental interactions with child based on the *Community Support Inventory* (Centres for the Prevention of Child Neglect, 2000); the *Framingham Safety Survey* (FSS; American Academy of Pediatrics, 1991); and the Infant-Toddler version of the *Home Observation for Measurement of the Environment* (HOME; Caldwell and Bradley, 1984) combined with the *Supplement to the HOME Scale for Impoverished Families* (SHIF; Ertem et al., 1996). Further information on each of these scales and the subdomains listed in Table 4 can be found in Appendix B.

Parental Interactions with Baby

Mothers were asked 16 questions ($\alpha = .74$) relating to how often they did certain activities (e.g., singing songs, dancing, telling stories) with their baby. These items were taken from the *My Baby and Me* program and *Parenting for the First Time* program (Centres for the Prevention of Child Neglect, 2000). Answers were given on a 6 point scale ranging from 0 representing *not at all* to 5 signifying *more than once a day*. A scale representing the frequency of the mother's interaction with her baby was created by taking an average of all responses, with higher scores indicating more interaction.

Framingham Safety Survey

The *PFL* evaluation combined multiple measures to assess the safety of the physical environment. Specifically, 20 of the 23 items on the birth to 12 month version of the *Framingham Safety Survey* were used in the six month PFL interview and 15 of the 30 items on the 1 to 4 year version were used in the 18 month PFL survey. These items were combined with questions assessing the presence of five common safety items (e.g., safety gate) in the house. Items are rated on a scale from zero to 10, with higher scores representing a safer environment. An overall safety score is obtained by taking the average of all items.

Home Observation for Measurement of the Environment and Supplement to the HOME Scale for Children Living in Impoverished Urban Environments

The Infant-Toddler version of the Home *Observation for Measurement of the Environment* is a 45-item instrument completed by a trained interviewer. It measures the stimulation potential of the child's home environment, and may be used as a substitute for reliance on social class as an indicator of quality of the child's home environment. The HOME Inventory comprises six domains. *Responsivity* (11 items) illustrates the degree to which a parent is responsive to the child's behaviour. *Acceptance* (8 items) represents parental acceptance of negative behaviour from the child and avoidance of unnecessary punishment. *Organisation* (6 items) pertains to the degree of routine in a family's schedule, safety of the environment, and community supports utilised. The *learning materials* domain (9 items) assesses the appropriateness of play materials for the child. *Involvement* (6 items) illustrates the degree to which the parent is involved in the child's learning and promotes child development. Finally, the *variety* domain (5 items) assesses visitation of people and attendance of activities that introduce variety into the child's life. Each item was scored by a trained interviewer as true or not. Items were scored based on observations while in the home. For items where this was not possible, the mother is directly asked the question in an interview format. If the item was *true* it is scored as a 1, if it is *not true* it is scored as 0. Scores for each domain on the HOME Inventory were obtained by averaging the responses to each question in that domain resulting in a score ranging from 0 to 1 with higher scores indicating a more nurturing home environment.

The *Supplement to the HOME Scale for Children Living in Impoverished Urban Environments* (SHIF; Ertem, Avni-Singer, & Forsyth, 1996) consists of 20 items that were combined with the HOME Infant/Toddler Inventory and administered by a trained interviewer along with the HOME Inventory. The SHIF was developed to be used in conjunction with the HOME Infant/Toddler Inventory to provide a more suitable and accurate assessment of the home environment of young children living in low socioeconomic urban areas. Additionally, four items assessing child interaction with adult figures (not father figures) and the level of noise generated

inside and outside the house were added to this measure as they were thought to be particularly relevant to low income populations. SHIF items, as well as these additional four items, are scored in the same way as items on the HOME Inventory, with a score of one representing that the statement is *true* and a zero representing that it is *not true.* In addition to the individual HOME Inventory domains described above, a combined score using information from all questions related to the home environment (e.g., HOME + SHIF) was used to form a composite measure of stimulation in the home environment. The combined measure consists of 69 items, 45 from the HOME Inventory, 20 from the SHIF, and the four additional questions described above which form eight subscales. *Daily routines* (10 items) contains items pertaining to the child's eating and sleeping patterns and the availability of food and safe sleeping facilities. *Child care* (5 items) provides details about the range, adequacy and appropriateness of childcare used by parents. *Outings* (5 items) measures the variety of stimulation the child receives in the form of trips made outside the home environment. *Toys and books* (10 items) measures the variety of appropriate play and learning materials available to the child in the home environment. *Play* (10 items) contains items relating to stimulating interactions between the parent and the child, and the parent's conscious encouragement of the child's development. *Physical environment* (10 items) is an observational subscale which contains items relating to cleanliness and safety in the home, as well as the presence of literacy materials. *Interaction* (13 items) measures the parent's warmth and responsiveness in interacting with the child. Finally, *restriction* (6 items) measures the level of restraint the parent places on the child during the visit, in the form of physical punishment and scolding, as well as inappropriate handling by older children. Scores for each domain were obtained by averaging the responses to each question in that domain. All calculated scores range from 0 to 1 with higher scores indicating a more stimulating home environment. In addition, a total SHIF score was obtained by summing the responses to 20 SHIF items, resulting in a score ranging from 0 -20.

Appendix E. Addressing Attrition and Non-response

Table E1: Estimated Regression Coefficients for Logit Models Employed to Calculate the Probability of Participating in Each Round of Data Collection

VARIABLES	(1) 6m High	(2) 6m Low	(3) 12m High	(4) 12m Low	(5) 18m High	(6) 18m Low
Socioeconomic						
Mother's age	-	0.746	-	-	-	-
		(0.684)				
Teen Mother	-	-	-	1.012	-	-
				(2.126)		
# Biological children (including foetus)	-	11.93	-	-	-	-
		(3,709)				
First time mother	-	13.09	-	-	-	-
		(3,709)				
IQ (WASI)	-	0.0846	0.0294	0.0545	0.0395	0.0389
		(0.140)	(0.0349)	(0.0729)	(0.0359)	(0.0325)
Low education	-0.798	-0.762	-	-3.462*	-	-0.0765
	(0.579)	(2.182)		(1.908)		(0.705)
Literacy Difficulties	-	-	-1.014	-	-0.657	-0.869
			(0.834)		(0.884)	(0.762)
Mother in Paid Employment	0.555	-	1.061	-	1.493	-
	(1.151)		(1.904)		(2.042)	
Irish Ethnicity (not including Irish Travellers)	-	-	-	-0.242	-	0.537
				(2.447)		(1.298)
Mother's Health						
Number of cigarettes smoked during pregnancy	-	0.209	-	-	-	-
		(0.301)				
Consumed alcohol during pregnancy	-	-	0.672	-	-	-
			(0.934)			
Ever took drugs	1.672	-	-	-	-	-
	(1.132)					
Exercise more than 3 times per week	-	-	-	1.606	-	-
				(1.604)		
Using birth control at time of pregnancy	-	-	-	-	-2.255***	-
					(0.859)	
Personality						
Consideration for future consequences (CFC)	0.153*	-	0.212	-	0.213	-
	(0.0917)		(0.135)		(0.129)	
Conscientiousness Personality Score (TIPI)	-	-	0.336	-	-	-
			(0.285)			
AAPI						
Low risk of child abuse/neglect (AAPI total raw score)	-	-	-	-	0.0411	-
					(0.0337)	
Low risk of child abuse/neglect (AAPI sten score)	-	-	-	-1.864	-	-
				(1.837)		
At risk of child abuse/neglect (AAPI cutoff)	-0.543	-	-	-	0.104	-2.667
	(0.858)				(1.462)	(1.673)
Risk of child abuse/neglect (AAPI total number of subdomains for which mother is at risk)	-	-	-	-1.297	-	0.968*
				(1.422)		(0.503)
Low risk of oppressing children's power (AAPI sten score)	-	0.378	-	0.258	-	-0.191
		(1.221)		(0.822)		(0.249)
At risk of oppressing children's power (AAPI cutoff)	-	-2.864	-	-2.386	-	-2.191*
		(3.038)		(2.592)		(1.159)
At risk of parent-child role reversal (AAPI cutoff)	-	-	-	-2.642	-	-
				(2.833)		
At risk of inappropriate parental expectations (AAPI cutoff)	-	-	-	-	-	-2.859**
						(1.414)
At risk of parent-child role reversal (AAPI cutoff)	-	-	-	-2.642	-	-
				(2.833)		
Social Support						
Level of support from work colleagues	0.466	-	-0.261	-	-0.513	-
	(0.535)		(0.758)		(0.774)	

Mother receives support from relations	0.461	-	-		-	-
	(0.290)					
Meets with friend regularly	-	-	-	2.562	-	-
				(1.719)		
Level of satisfaction with neighbourhood	-	-	-	-2.109	-	-0.0674
				(1.685)		(0.727)
Service Use						
# Child/family services used	-	15.22	-	-	-	-
		(3,435)				
# Total services used	-	0.234	-	-	-	-
		(0.620)				
Uses child/family services	-	-13.49	-	-	-	-
		(3,435)				
# Health services used	-	-	-	0.186	-	0.512
				(0.741)		(0.378)
# Employment services used	-	-	-	-	-	0.343
						(0.480)
Constant	-3.139	-44.74	-3.001	12.26	-5.419	-1.439
	(1.946)	(7,418)	(3.439)	(11.03)	(5.158)	(3.149)
Observations	104	91	90	91	90	91
Pseudo R-squared	0.201	0.484	0.215	0.568	0.312	0.226

Notes: Coefficients reported and Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

The Weschler Abbreviated Scale of Intelligence (WASI) was used to measure IQ at 3 months postpartum. The Consideration of Future Consequences (CFC) Scale is a measure of the extent to which people consider distant versus immediate consequences of possible behaviours. Higher scores are indicative of more consideration for future consequences. The Ten Item Personality Inventory (TIPI) was used to measure the Big-Five Personality framework. Higher scores on the conscientiousness subdomain are indicative of more conscientious behaviour. The Adult Adolescent Parenting Inventory (AAPI) measures approaches to parenting with higher scores indicating a lower risk endorsement of abuse/neglect. AAPI cutoff scores are binary measures which indicate whether the participant was at risk of abuse/neglect.