Early Skill Formation and the Efficiency of Parental Investment: A Randomized Controlled Trial of Home Visiting

Orla Doyle, Colm Harmon, James J. Heckman, Caitriona Logue, Seong Hyeok Moon

PII: S0927-5371(16)30308-6
DOI: http://dx.doi.org/10.1016/j.labeco.2016.11.002
Reference: LABECO1508

To appear in: *Labour Economics*

Received date: 31 December 2015
Revised date: 30 September 2016
Accepted date: 10 November 2016

Cite this article as: Orla Doyle, Colm Harmon, James J. Heckman, Caitriona Logue and Seong Hyeok Moon, Early Skill Formation and the Efficiency of Parental Investment: A Randomized Controlled Trial of Home Visiting, *Labour Economics,* http://dx.doi.org/10.1016/j.labeco.2016.11.002

Early Skill Formation and the Efficiency of Parental Investment: A Randomized Controlled Trial of Home Visiting[*]

Orla Doyle[a], Colm Harmon[b], James J. Heckman[c], Caitriona Logue[d,♦], Seong Hyeok Moon[e]

[a] UCD School of Economics & UCD Geary Institute for Public Policy, University College Dublin, Belfield, Dublin 4, Ireland (orla.doyle@ucd.ie).

[b] School of Economics H04 - Merewether Building, The University of Sydney NSW 2006, Australia, and Institute for the Study of Labor (IZA) (colm.harmon@sydney.edu.au).

[c] Department of Economics, The University of Chicago, 1126 E. 59th Street Chicago, IL 60637, USA, and Institute for the Study of Labor (IZA) (jjh@uchicago.edu).

[d] UCD School of Economics, University College Dublin, Belfield, Dublin 4, Ireland, and The Economic and Social Research Institute (ESRI) (caitriona.logue@esri.ie).

[e] Center for the Economics of Human Development, The University of Chicago, 1126 E. 59th Street Chicago, IL 60637, USA (seonghmoon@gmail.com).

Abstract

This paper presents evidence on early skill formation and parental investment using an experimentally designed, home visiting program targeting disadvantaged Irish families. Program effects from pregnancy to 18 months are estimated using measures of parenting and child cognitive, noncognitive and physical development. Permutation testing, a stepdown procedure, and inverse probability weighting are applied to account for small sample size, multiple hypothesis testing, and attrition. The program's impact is concentrated on parental behaviors and the home environment with small to moderate effect sizes found. Deficits in parenting skills can be offset within a relatively short timeframe, yet continued investment may be required to observe child effects.

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

JEL Classification: D13, I26, J13, C93.

[*] Corresponding Author: Orla Doyle (orla.doyle@ucd.ie).

[♦] Present Address: The Economic and Social Research Institute, Whitaker Square, Sir John Rogerson's Quay, Dublin 2, Ireland (caitriona.logue@esri.ie).

Highlights:

- RCT examining the impact of an early childhood program up to 18 months of age
- Skill formation measured by cognitive, noncognitive and physical development
- Parental investment measured by the home environment, beliefs and attachment
- Permutation tests, stepdown procedure, and inverse probability weighting applied
- Some identified effects on parenting, but no impact on early skill formation

Trial registration: AEA RCT Registry: AEARCTR-0000066,
https://www.socialscienceregistry.org/trials/66

Ethical Standards Statement: This study was conducted according to the guidelines laid down in the Declaration of Helsinki and all procedures involving human subjects were approved by the University College Dublin's Human Research Ethics Committee and the Rotunda Hospital's Ethics Committee and the National Maternity Hospital's Ethics Committee. Written informed consent was obtained from all participants prior to randomization.

Financial Support: Funding for this study was made available through a European Research Council (ERC) Advanced Investigator Award to James J. Heckman, and the Innovation Academy UCD Bursary awarded to Caitriona Logue. The overall trial was funded by the Northside Partnership, through The Atlantic Philanthropies and the Department of Children and Youth Affairs, awarded to Orla Doyle. The Northside Partnership had no role in design, analysis, or writing of this article.

Conflicts of interest: We declare that we have no conflicts of interests.

Acknowledgements: We thank the European Research Council (ERC) for the Advanced Investigator Award to James J. Heckman, and the Innovation Academy UCD for their Bursary awarded to Caitriona Logue. We also thank The Northside Partnership (which received funding from the Irish Government Department of Children and Youth Affairs and The Atlantic Philanthropies) who funded 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 Early Childhood Research Team. Helpful comments from seminar participants at University College London, University of Stirling, NUI Galway, Queens University Belfast, Royal Holloway University of London, ZEW Mannheim, University of Tasmania, University of Sydney, University of Queensland, Alberto Hurtado University, Pontifical Catholic University of Chile, Aarhus University, and Xiamen University are gratefully acknowledged. We would also like to thank discussants at the 22nd European Workshop on Econometrics and Health Economics, the 2013 Annual Health Econometrics Workshop, the 2012 European Doctoral Group in Economics Jamboree, as well as participants at the CEPR/IZA 14th European Summer Symposium in Labour Economics and the 6th Irish Conference on Economics and Psychology. We would also like to thank the anonymous reviewers for their helpful comments. The usual disclaimer applies.

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 including fewer behavioral problems and criminal convictions, lower dependency on welfare, increased employment, and improved health (Olds et al. 1998; Heckman et al. 2010; Campbell et al. 2014; Kautz et al. 2014; Elango et al. 2015). Cunha and Heckman (2007) and Cunha et al. (2010) present a model of skill formation demonstrating that early skills facilitate the accumulation of later age skills, and these higher level skills make further investment throughout the lifecycle more productive through a process of dynamic complementarity (see Heckman and Mosso 2014). Little is yet known about the mechanisms involved in producing these long-term effects (see however, Heckman et al. 2013 and Heckman and Mosso 2014).

This paper presents empirical evidence on the nature of skill formation and parental investment in the early years based on Preparing for Life (PFL), an experimentally designed, home visiting program in Ireland targeting disadvantaged families. The program begins in utero and continues until age 5 and thus has the potential to influence skill formation during a period in which brain development is thought to be at its most malleable (Nelson 2000; Knudsen et al. 2006). Based on a rich and extensive data set including child cognitive, noncognitive and physical developmental outcomes and various dimensions of parental investment, we investigate the early impact of the program on participating families during infancy and toddlerhood. This allows us to identify the specific areas where effects from targeted intervention programs manifest early in the lifecycle, thus investigating the mechanisms involved in the early investment process.

The importance of the period from in utero to age 3 for the development of skills has been highlighted in recent literature. Studies from neuroscience and epigenetics demonstrate that the brain has higher plasticity at earlier ages, and that a child's abilities and behaviors have both a genetic and environmental component, and in particular, the environment can play a role in shaping the developing brain (see Halfon et al. 2001; Wydner 1998; Weaver et al. 2004; Heckman 2007). For this reason, many interventions have been designed to target the first 1,000 days of life. Such interventions have been shown to be successful at improving early cognitive skills (Barham et al. 2013; Heckman 2000), although fade-out has been found in some cases (Heckman 2000). Heckman and Kautz (2012) note that the Perry Preschool

Program resulted in favorable adult outcomes despite fade-out of early improvements in IQ scores. The authors argue that the long-term effects are driven by improvements in noncognitive or character skills. Recent research has expanded the discussion by examining the impact of early intervention on physical health. Barham et al. (2013) find no impact of early intervention on the physical health of 10 year-olds, while Campbell et al. (2014) find significant improvements in adult health outcomes. Thus, there is evidence that cognitive and noncognitive skills, as well as physical health, can be impacted by early intervention. However, the timing and mechanisms which generate such effects are not yet fully understood.

One potential mechanism which may generate these effects is through changes in parenting behavior. Parental decisions about investment in their children are the primary mechanisms through which the fetal and early childhood environment can be enriched. Hertwig et al. (2002) suggest that parental investment can be divided into at least three categories and propose that material resources, cognitive stimulation and parental interpersonal skills (e.g., affection and encouragement) may each serve divergent roles in the transmission process to the child. Heckman and Kautz (2013) highlight the importance of parenting qualities such as stimulation, attachment, and encouragement. Cunha et al. (2013) suggests that mothers may vary with respect to their understanding of child development and find that disadvantaged mothers have lower expectations of the returns to early life investment. They estimate that policies which increase maternal knowledge about the technology of skill formation could potentially augment child development. Currie (2001) also discusses information failures among parents and highlights the role for government intervention in assisting parents to make decisions about early childhood education. Collectively, this evidence motivates the focus on parents as the first mechanism of change in many early childhood interventions (E.g., Olds et al. 1998 with the Nurse Family Partnership program; Sandner 2013 with the ProKind program).

The primary contribution of this study relates to the multiplicity of skills and behaviors analyzed, as well as the frequency of assessment in infancy and toddlerhood. In particular, given that abilities are often parsed into distinct dimensions, we separate early outcomes into three categories: cognitive development, noncognitive development, and physical development. In addition, we examine the tractability of parenting skills by examining six dimensions of parental investment: physical environment, appropriate care,

interactions with infant, maternal attachment, maternal self-efficacy and beliefs about parenting. Thus, we contribute to the early intervention literature by measuring the malleability of distinct child and parent outcomes during one of the most important developmental phases of life.

We estimate the impact of PFL on parental investment and child skill formation by applying statistical methods specifically tailored for the analysis of multiple outcomes at multiple waves when using small samples. We present results from both classic *t* tests and nonparametric permutation tests. Permutation tests do not make any distributional assumptions and therefore produce valid *p*-values when distributions are skewed (Heckman et al. 2010). Using the methodology of Romano and Wolf (2005) and Romano et al. (2010), we utilize a stepdown procedure to adjust for the increased likelihood of false discoveries when examining multiple outcomes using fixed *p* values for each hypothesis. We also attempt to address differential attrition and non-response by applying inverse probability weights (IPW) and we assess the internal validity of the results by testing for the presence of contamination.

Estimating each treatment effect separately, we find significant program effects (at the 10 percent level) for 17 percent of outcomes (6/35) at six months, 4 percent of outcomes (1/23) at 12 months, and 18 percent of outcomes (5/28) at 18 months. While this is suggestive of a positive program effect at 6 and 18 months, when a more rigorous stepdown method is applied and the *p*-values are adjusted to account for the increased likelihood of a Type I error, we find significantly fewer treatment effects. The stepdown results indicate that the treatment effects are concentrated on parental investment decisions relating to the quality of the home environment and the level of care mothers provide for their children, with small to moderate effect sizes found. The weighted analysis produces fewer significant treatment effects, yet the overall pattern of results is similar to the unweighted case. Finally, we find limited evidence that the results are biased due to contamination.

The rest of the paper is structured as follows. Section I reviews findings from studies of home visiting programs examining early child development and parental investment outcomes. Section II describes the PFL intervention, the recruitment and randomization procedure and the estimation sample. The econometric framework and outcomes assessed are described in Section III. The results are presented in Section IV, and Section V concludes.

I. Home Visiting Programs and Early Outcomes

Family-focused approaches to early intervention have become increasingly popular due to the growing awareness of the importance of parental behaviors on child development (Brooks-Gunn et al. 2000). Table A.1 in Web Appendix A summarizes evidence from a range of home visiting programs which examine child development and parenting outcomes up to 18 months of age.[1] All of the programs focus on similar mechanisms that promote child development such as educating parents about developmental milestones and health, encouraging a healthy lifestyle, affirming maternal perceptions of self-efficacy in the parenting role, and encouraging positive parenting practices. Overall, there is limited evidence in the literature of treatment effects on child development up to 18 months. Two previous studies examine child development outcomes at 6, 12 or 18 months and statistically significant treatment effects are found in just one case (Landsverk et al. 2002). A greater number of studies examine parenting outcomes, yet few identify significant treatment effects. Of the seven studies measuring parental investment at 6, 12 and 18 months, only two identify significant favorable effects (Minkovitz et al. 2001; LeCroy and Krysik 2011).

None of the studies reviewed use methods that address sample size limitations. While some have the advantage of larger samples (e.g., Duggan et al. 1999; Minkovitz et al. 2001; Landsverk et al. 2002; Duggan et al. 2004; Johnston et al. 2004; Drotar et al. 2009), others acknowledge the issue of small samples yet do not adapt their statistical approach (e.g., Koniak-Griffen et al. 2000; LeCroy and Krysik 2011). 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 examined. Avellar and Paulsell (2011) note that few of the studies examined as part of HomVEE review make corrections for multiple hypothesis testing and advise caution when interpreting the significance of the findings. Similarly, none of the studies reviewed address the issue of differential attrition and the potential bias that may result. To ensure comparability with the existing home visiting literature, we present unweighted estimates in this paper, which do not account for differential attrition, as our main results in Section IV.

[1] The source for this review was the Home Visiting Evidence of Effectiveness website (HomVEE; U.S. Department of Health and Human Services 2009). As described in Paulsell et al. (2010), this site was launched by the U.S. Department of Health and Human Services. In this paper we only consider studies examining outcomes before and up to 18 months of age. Furthermore, we focus only on studies that were rated 'high' quality according to the HomVEE criteria.

As a robustness check, we also present weighted estimates which account for observed attrition patterns.

II. Preparing for Life

A. Treatment

PFL is a five-year program developed to improve social mobility in a multi-generation, suburban community classified by welfare authorities as disadvantaged and consisting mainly of welfare (or social) housing in Dublin, Ireland. The program was initiated and developed by community representatives and local health and education service providers to improve children's early skill formation.[2] The intervention begins during pregnancy and continues until the child starts formal schooling at age 4/5. The program is evaluated using a randomized control trial (RCT) design in which all families who consented to take part were randomly assigned to either a high or a low level of treatment.

High Treatment - Home Visiting Program

Participants in the high treatment group avail of a home visiting mentoring program. The aim of this program is to support and provide education to parents on key child rearing issues through developing a strong parent-mentor relationship (Preparing for Life and The Northside Partnership 2008). The home visits start in the prenatal period, as soon as the participant joins the program (at ~21 weeks), and continues until school entry. Home visiting is a widely used form of early intervention which provides parents with information, emotional support, access to other community services, and direct instruction on parenting practices (Howard and Brooks-Gunn 2009). The program shares some similarities with the Nurse Family Partnership program (Olds et al. 1997), but it is longer in duration and the visits are provided by mentors rather than nurses.

Mentors

[2] Doyle and McNamara (2011) find that prior to the introduction of PFL, children from the catchment area were rated by teachers to be below the norm at school entry across all five domains on the Short Early Development Instrument (EDI) including children's *physical health and wellbeing*, *social competence*, *emotional maturity*, *language and cognitive development*, *communication and general knowledge*. Note that the S-EDI norm is based on a representative sample of Canadian children (Janus and Duku 2005). There is no Irish norm for the short-form of the EDI.

The professional qualifications of the mentors vary and include education, social care, youth studies, psychology, and early childcare and education. Each mentor completed extensive training prior to program implementation and weekly supervision thereafter. The role of the mentor is to build a good relationship with parents, provide them with high quality information and to be responsive to issues that arise. 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 same mentor is assigned to each family over the course of the intervention when possible.

Content

The home visits are tailored based on the age of the child and the needs of the family and are guided by a set of 178 Tip Sheets which present best-practice information on pregnancy, parenting, and child health and development. The Tip Sheets are colorful representations of information presented in a clear, concise manner and were developed by PFL staff based on available information from local organizations such as the Health Service Executive, the Department of Health and Children, and Barnardos Children's Charity. The Tip Sheets are designed at a reading level of a 12 year-old to make them as accessible as possible. 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. It is intended that all participants must have received the full set of Tip Sheets by the end of the program. While some of the Tip Sheets promote multiple aspects of school readiness, the majority of the Tip Sheets focus on physical health and well-being (n = 105), followed by social competence and emotional maturity (n = 60), approaches to learning (n = 30), language (n = 25), and cognitive skills (n =22).[3]

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.

Dosage

[3] Note that these figures do not sum to 178 as some Tip Sheets are classified in more than one domain. An example of a Tip Sheet can be found in Web Appendix B.

The mentors visit the family home twice monthly for between 30 minutes and 2 hours. 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 feasible for all families. Therefore the program 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. Participants were prescribed 49 home visits between program entry during pregnancy and when the children were 18 months of age. On average, participants received 29 home visits during this period which represents 60% of prescribed visits and is consistent with other home visiting programs (Gomby et al. 1999).[4] Table 1 documents prescribed and realized engagement in the program at six monthly intervals. It shows that the number of home visits realized was largely consistent in each period, yet less than prescribed. [5] A previous study on fidelity within the PFL program using qualitative data (Lovett et al. 2016) identified a number of challenges to early engagement in the program generated by cultural and familial barriers to implementation and misconceptions about program aims. Yet as the program progressed, there was a strengthening of the parent-mentor relationship which was facilitated by building mutual rapport and tailored program delivery.

[4] To investigate the predictors of dosage, we examined the relationship between the number of home visits received and 22 socio-demographic and maternal psychosocial characteristics collected at baseline using an OLS regression. We found that 6 of the 22 characteristics had a significant impact on the number of visits. Specifically, mothers who joined the program earlier in pregnancy, mothers with higher cognitive skills, and mothers with greater knowledge of infant development had more home visits since joining the program. Whereas mothers who were married, saved regularly, and smoked during their pregnancy had fewer home visits by 18 months. Thus the results are mixed, in some cases better characteristics are predictive of more home visits, but the converse is also true.

[5] Note that 18 of the 115 participants randomized to the high treatment group did not receive any home visits at all. None of these participants took part in the assessments at baseline, 6, 12 or 18 months. Despite this initial dropout, the high and low treatment groups remained balanced as the baseline assessment (see Section II.B). Our analysis is an intention-to-treat analysis as the actual dosage received by each participant may be less than prescribed. Despite this initial dropout, the high and low treatment groups remained balanced as the baseline assessment (see Section II.B).

Table 1 – Prescribed and Realized Engagement in PFL Home Visits

	Prenatal – birth	Birth – 6 months	6 Months – 12 Months	12 Months – 18 Months	Total
Prescribed number of home visits (bi-monthly)	10	13	13	13	49
Prescribed frequency of home visits (bi-monthly)	Bi-monthly	Bi-monthly	Bi-monthly	Bi-monthly	Bi-monthly
Prescribed length of home visits (bi-monthly)	30mins-2hrs	30mins-2hrs	30mins-2hrs	30mins-2hrs	30mins-2 hrs
Realized number of home visits	6.6 (4.4) 0-32	8.2 (3.8) 0-19	7.6 (3.7) 0-17	7.1 (3.7) 0-21	29.5 (13.1) 4-66
% of prescribed home visits realized	72.4 (46.6) 0-350	62.5 (29.4) 0-146	58.6 (28.7) 0-130	54.8 (28.1) 0-162	60.3 (25.4) 8-137
Realized length of home visits (mins)	55.8 (18.7) 0-111	59.1 (13.4) 0-90	57.9 (13.4) 0-90	58.5 (15.6) 0-105	58.9 (8.9) 40.5-82.3
Realized duration of home visits (hours)	6.3 (4.1) 0-18	8.3 (4.2) 0-19	7.8 (3.9) 0-18	7.2 (3.8) 0-19	29.3 (13.5) 3-57

Note: The table presents the mean, standard deviation in parentheses, and the minimum and maximum values. These statistics are calculated for high treatment participants included in the 18 month estimation sample (n=80).

Common Intervention Components

Both the high and low treatment groups are offered the following supports:

Developmental Materials

Families in both groups receive developmental packs annually to the value of approximately €100pa. By the time the study child had reached 18 months of age, participating families had received the first and second development packs. 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.

Public Health Workshops

Both groups are encouraged to attend public health workshops which are already operating in the community. The Stress Control Program, which is run by external facilitators, involves six one-hour weekly sessions which focus 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 program also teaches techniques and strategies to manage stress. Participants receive a set of booklets and a relaxation CD.

Both groups are also invited to participate in the Healthy Food Made Easy program, which is facilitated by one of the PFL mentors and involves six two-hour sessions. The aim of the program is to improve nutritional knowledge, attitudes and behavior by learning about basic nutritional theories and participating in practical cookery sessions. It is a peer led program which emphasises group learning through discussion, worksheets and hand-outs, quizzes, problem solving games, food preparation and cookery.

Access to a Support Worker and Other Supports

All participants in the high and low treatment groups receive a directory of local services and have access to a PFL support worker who can help them connect to additional community "services as usual" if needed. The service provided to the low treatment group is operated from the PFL centre by a support worker who is not trained to provide advice on child development or parenting. For the high treatment group, the mentor's role subsumes the support worker's responsibilities. Details about PFL coffee mornings and other community events are sent via group text or Facebook messenger. Finally, both treatment groups receive a framed professional photograph of their child, as well as program newsletters and special occasion (e.g., birthday) cards.

By comparing the high and low treatment groups, it is possible to extract the differential impact of the home visiting component, layered on top of the low treatment supports. There could be an element of complementarity between the basic set of provisions delivered to both groups and the additional supports delivered to the high treatment group only. Therefore, we cannot infer that the estimated treatment effects would be replicated in the absence of the common set of provisions. Figure B.1 in Web Appendix B summarizes the components of the low and high treatments, and Doyle (2013) discusses the PFL program and evaluation design in greater detail.

B. Recruitment and Randomization

Recruitment took place between 2008 and 2010. The inclusion criteria included all pregnant women living in the PFL catchment area, regardless of parity or family background. There were no exclusion criteria. Participation was voluntary and eligible candidates were identified using maternity hospital records and self-referral in the community. A total of 233 pregnant women consented to participate.[6] A computerized unconditional probability randomization procedure assigned 115 participants to the high treatment group and 118 to the low treatment group.[7] No stratification or block techniques were used.

To test the validity of the randomization procedure, a baseline survey was administered to 205 (high =104; low = 101) participants post-randomization, yet before treatment began.[8] Seventy-four baseline variables were analyzed using permutation testing (the method is described in detail in Section III.B) for the 3 estimation samples included in the 6, 12, and 18 month analyses. No significant differences between the high and low treatment groups were found on between 92-97% of measures (depending on the estimation sample examined), using the 10% cut-off for significance. Furthermore, when we group the baseline measures into 5 categories for joint hypothesis testing, we fail to reject the null hypothesis of joint significance for any of these categories. This indicates that the

[6] This represents a recruitment rate of 52 percent based on public health records on the number of live births in the community during the recruitment window. 22 percent of potential participants were not identified for recruitment and 26 percent were identified but could not be contacted for a final acceptance, or were contacted and refused to join the program. A socio-demographic profile survey was conducted with a sample of eligible non-participants (n=102) when their children were 4 years old. The survey asked participants about their current socio-demographics and also their socio-demographics when they were pregnant with the eligible study child during the recruitment window. There is some evidence to suggest that the eligible non-participants are of a higher socioeconomic status than the participants who joined the program. This suggests that the program was effective in targeting the families most in need of the intervention.

[7] *PFL* participants were randomized after informed consent was obtained. To ensure randomization was not compromised, a computerized randomization procedure was used whereby the participant pressed a key on a computer which randomly allocated her treatment assignment. Once assignment was complete, an email was generated which included the participant's unique ID number and assignment condition. This email was automatically sent to the *PFL* program manager and the evaluation manager. If there were any attempts to reassign participants from one group to another, by either directly changing the database or repeating the randomization procedure, a second email would automatically highlight this intentional subversion. This preventative measure was important given the evidence of compromised randomization in some of the most influential early childhood interventions such as the Perry Preschool Program (Heckman et al. 2010).

[8] A total of 28 randomized participants (low = 17; high = 11) were not assessed at baseline. Of these, 19 participants (low=13; high=6) elected to withdraw from the program before the baseline interview, 2 participants (low= 1; high=1) miscarried before completing the baseline interview, 5 participants (low = 2; high = 3) missed the baseline interview and did not participate in any subsequent assessments, and 2 participants (low=1; high=1) missed the baseline interview but participated in later assessments. An analysis of a subset (N=12) of these early program exits who agreed to provide limited data 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.

randomization process was generally successful. Full descriptive tables, including all the measures in the baseline analysis, are available in Web Appendix C.

In order to investigate the impact of the program during the early stages of infancy, which has received limited attention in the economic literature, this paper uses data from the baseline, six, 12, and 18 month assessments. Trained interviewers, who were blinded to the treatment condition, collected data through face-to-face interviews conducted primarily in the participant's home using computer-assisted personal interviewing. The structure and design of the questionnaires were varied from wave to wave to assuage respondent fatigue, and, where possible, the repetition of identical questions was avoided in two consecutive interviews. For example, for the parenting measures, the same instruments were never used consecutively.

C. Participant Profile

Table 2 provides baseline descriptive statistics for the estimation sample available at each wave.[9] 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. Over one-quarter indicated that they had a mental health condition, and with respect to substance use during pregnancy, one half of participants smoked and just over a quarter drank alcohol. The participants have a low level of formal education compared to the national average.[10] Using a more refined measure of cognitive capacities, the average level of maternal IQ was approximately 82 using the *Wechsler Abbreviated Scale of Intelligence* (Wechsler 1999) which is below the lower bound on the expected population average range of between 85 and 115. Table 2 also demonstrates that the estimation samples are balanced at each time point.

To place the PFL sample in context, we compare our sample with the nationally representative *Growing up in Ireland (GUI) - Nine Month Cohort Study*, which was

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

[10] 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 high school) or lower, which is effectively minimum compulsory schooling. This compares with an age-cohort completion rate of high school of 74 percent. Thus, the dropout rates from high school are almost three times the national average.

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 parents were five years older on average when pregnant with the study child than PFL parents, with education levels in line with expected national averages. Approximately 11 percent of GUI parents 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 versus 50 percent), 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 versus 16 percent), while the percentage that indicated having *either* a partner or spouse was similar to the PFL sample (88 percent versus 81 percent). Overall, this comparison highlights that the PFL cohort reflects a relatively disadvantaged sample when compared with national averages, with significant differences in self-reported and objective health behaviors.[11] A detailed comparison of the GUI and PFL samples is presented in Table C.7 in Web Appendix C.

[11] The GUI data are collected when children are aged 9/10 months and 36 months. We will conduct an outcome comparison with GUI when the PFL 36 month surveys are completed.

Table 2 – Baseline Comparison of High/Low Treatment Participants

	6 Month Sample			12 Month Sample			18 Month Sample		
	High Treatment Mean (SD)	**Low Treatment Mean (SD)**	**p[(i)]**	**High Treatment Mean (SD)**	**Low Treatment Mean (SD)**	**p[(i)]**	**High Treatment Mean (SD)**	**Low Treatment Mean (SD)**	**p[(i)]**
Weeks pregnant at program entry	21.78 (7.83)	21.18 (6.87)	0.594	21.84 (7.88)	21.17 (7.02)	0.566	21.93 (7.93)	21.32 (6.62)	0.608
Age	25.67 (5.76)	25.69 (6.04)	0.987	25.87 (6.01)	25.13 (6.02)	0.437	25.93 (5.91)	25.56 (6.10)	0.709
Married	0.16 (0.37)	0.17 (0.38)	0.861	0.16 (0.37)	0.16 (0.37)	1.000	0.16 (0.37)	0.15 (0.36)	0.842
Has partner (including married)	0.80 (0.40)	0.83 (0.38)	0.654	0.82 (0.39)	0.83 (0.38)	0.839	0.79 (0.41)	0.82 (0.39)	0.595
Living with parent(s)	0.55 (0.50)	0.45 (0.50)	0.197	0.54 (0.50)	0.48 (0.50)	0.438	0.54 (0.50)	0.47 (0.50)	0.379
First time mother	0.52 (0.50)	0.46 (0.50)	0.408	0.51 (0.50)	0.49 (0.50)	0.757	0.53 (0.50)	0.47 (0.50)	0.467
Low education	0.29 (0.46)	0.36 (0.48)	0.355	0.30 (0.46)	0.30 (0.46)	1.000	0.30 (0.46)	0.34 (0.48)	0.577
Employed	0.43 (0.50)	0.40 (0.49)	0.769	0.43 (0.50)	0.43 (0.50)	1.000	0.43 (0.50)	0.41 (0.50)	0.862
IQ[a]	82.52 (12.94)	80.60 (13.14)	0.335	83.11 (12.60)	81.54 (12.75)	0.429	83.32 (12.35)	82.04 (12.16)	0.521
Saves regularly	0.50 (0.50)	0.53 (0.50)	0.715	0.50 (0.50)	0.55 (0.50)	0.535	0.49 (0.50)	0.53 (0.50)	0.566
Resides in public housing	0.54 (0.50	0.56 (0.50)	0.742	0.54 (0.50)	0.55 (0.50)	0.876	0.55 (0.50)	0.53 (0.50)	0.846
Prior physical health condition	0.76 (0.43)	0.64 (0.48)	0.102	0.76 (0.43)	0.65 (0.48)	0.126	0.75 (0.44)	0.63 (0.49)	0.110

Prior mental health condition	0.27 (0.45)	0.26 (0.44)	0.885	0.28 (0.45)	0.26 (0.44)	0.727	0.26 (0.44)	0.26 (0.44)	0.975
Smoked during pregnancy before enrolment	0.51 (0.50)	0.49 (0.50)	0.817	0.51 (0.50)	0.46 (0.50)	0.535	0.51 (0.50)	0.47 (0.50)	0.566
Alcohol during pregnancy before enrolment	0.27 (0.45)	0.25 (043)	0.754	0.29 (0.46)	0.26 (0.44)	0.602	0.29 (0.46)	0.27 (0.45)	0.854
Drugs during pregnancy before enrolment	0.01 (0.11)	0.03 (0.18)	0.355	0.01 (0.11)	0.01 (0.11)	1.000	0.01 (0.11)	0.01 (0.12)	0.948
N	82	89		82	82		80	73	

Note: [a]The Weschler Abbreviated Scale of Intelligence (WASI) was used to measure maternal IQ at 3 months postpartum rather than baseline. [(1)] two-tailed *p*-value calculated from a *t*-test.

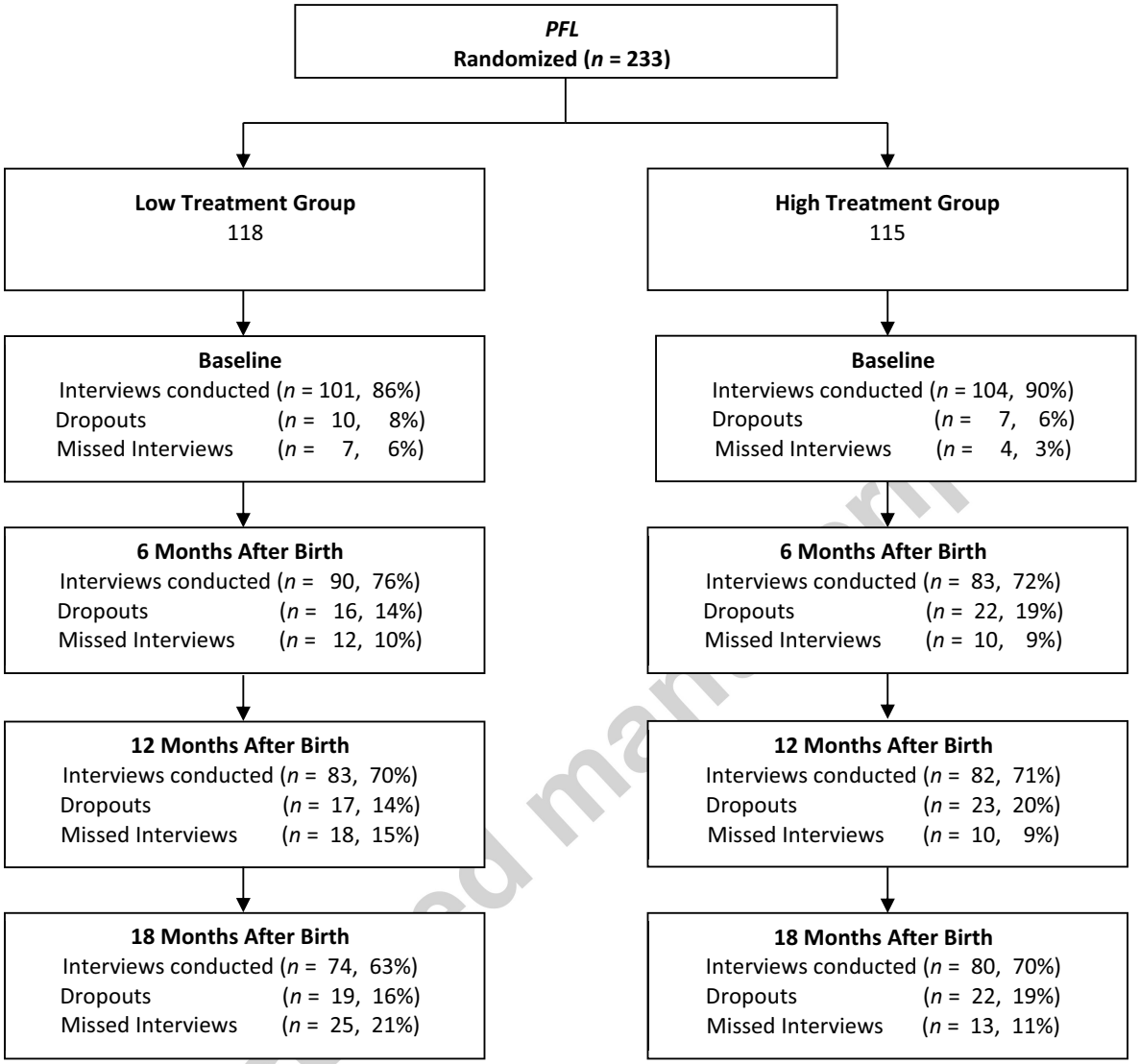
D. Attrition and Non-response.

Figure 1 describes the progression of the PFL participants from program entry until 18 months. After the study began, there were non-negligible dropouts. The 18 month assessment captured 70 percent of the originally randomized high treatment group (80/115) and 63 percent of the originally randomized low treatment group (74/118).

On average, 30 percent of the high treatment group (35/115) and 37 percent of the low treatment group (44/118) had either officially dropped out of the program or did not participate in one of the follow-up assessments between baseline and 18 months. The majority of dropout occurred before 6 months. In comparing the characteristics of individuals who did and did not complete a follow up interview (described in detail in Web Appendix D) some patterns did emerge. Specifically, in the high treatment group, mothers with missing outcome data appear to be a relatively more disadvantaged group. For example, maternal employment, the level of support that mothers receive from friends and family, and mothers' consideration of future consequences were inverse predictors of missing data. The results for the low treatment group were more mixed. For example, at six and 12 months, mothers who had achieved more than minimum schooling were significantly less likely to have missing data. However, mothers who had used child and family services in the community at baseline, and mothers with more children were also less likely to have missing data at 6 months. While at 12 months, teen mothers were also significantly less likely to have missing outcome data. Therefore, the pattern of missingness at 6 and 12 month is mixed. At 18 months, however, the participants in the low treatment group who were most likely to have missing data appear to represent a more advantaged group. For example, mothers who indicated that they did not have difficulty making ends meet, were employed, had taken folic acid supplements during pregnancy and were satisfied with their neighborhood were more likely to have missing 18 month outcome data.

Despite these patterns of attrition and missingness, as shown in Table 2, the groups remain balanced on baseline characteristics across each of the three waves. Specifically, no statistically significant baseline differences emerge for the 6, 12 or 18 month estimation samples. This analysis is further expanded in Tables C.1-C.6 in Web Appendix C. The inverse probability weighting (IPW) technique was used in an attempt to address the potential bias that this attrition or missing data may introduce and is described in detail in Web Appendix D.

Figure 1 Flowchart of Program Participation, Attrition, and Non-response



III. Econometric Framework

A. Estimation Model and Outcome Measures

This study adopts an intention-to-treat analysis and is evaluated using an RCT. The standard model of program evaluation describes the observed outcome Y_i of participant $i \in I$ by

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

where $I = \{1 \ldots N\}$ denotes the sample space, D_i denotes the treatment assignment for participant i ($D_i = 1$ for the intention-to-treat sample , $D_i = 0$ otherwise) and ($Y_i(0)$, $Y_i(1)$) are potential outcomes for participant i. We test the null hypothesis of no treatment effect. H-1: $E[Y_i \mid D = 1] = E[Y_i \mid D = 0]$.

Various standardized psychometric scales were administered at each data collection wave. We examine 53 outcome measures related to child development (15 measures) and parental investment (38 measures). In general, the same instruments were used to measure child outcomes at each wave. However, the parenting instruments were varied from wave to wave. Web Appendix E describes each of the standardized scales in detail. The following child development instruments are used: the *Ages and Stage Questionnaire* (ASQ), the *Ages and Stages Questionnaire: Social-Emotional* (ASQ:SE), an assessment of difficult temperament based on the *Infant Characteristics Questionnaire*, the *MacArthur-Bates Communicative Development Inventories: Words and Gestures, Short Form* (CDI-WG), the *Brief Infant-Toddler Social and Emotional Assessment* (BITSEA), the *Temperament and Atypical Behavior Scale* (TABS), and finally the *Developmental Profile 3, Cognitive Section* (DP-3).

Parental investment is examined using the following standardized scales: the *Parental Cognition and Conduct Towards the Infant Scale* (PACOTIS), the *Adult Adolescent Parenting Inventory 2* (AAPI-2), the *Knowledge of Child Development – Short Form* (KIDI-SF), *Parental Locus of Control* (PLOC), the *Parenting Daily Hassles Scale* (PDH), *Parenting Stress Index* (PSI), *Condon Maternal Attachment Scale* (CMAS), *Maternal Separation Anxiety Scale* (MSAS), a measure of parental interactions based on the activities scales used in the Early Head Start Research and Evaluation Project, the *Framingham Safety Survey* (FSS), the Infant-Toddler version of the *Home Observation for Measurement of the Environment* (HOME), the *Supplement to the HOME Scale for Impoverished Families*

(SHIF), and two indicators of whether the mothers reads to her child and how often she reads to her child. The reliability of these predominantly self-reported instruments is discussed in the results section.

B. Permutation Testing

Although the RCT design in (1) is a simple specification, the use of traditional *t* tests for hypothesis testing may not be 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 a few recent studies of an early childhood intervention using this approach (Heckman et al. 2010; Campbell et al. 2014), 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). The observed *t*-statistic is recorded and compared to the distribution of *t*-statistics that result from multiple, random permutations of the treatment label.[12] 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:

$$MP(t) = P(t^* > t) + 0.5P(t^* = t) \tag{2}$$

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 favorable effect on child and parenting outcomes compared to the low level treatment. We use one-sided tests as we are testing the hypothesis that the program has a positive impact on outcomes and the use of one-sided tests is consistent with other studies which evaluate early childhood interventions and use permutation tests for hypothesis testing (e.g., Conti et al. 2015; Campbell et al. 2014; Gertler et al. 2014; Heckman et al. 2010). Due to the small sample size, the accepted Type I error rate is set at the 10 percent level.

[12] 100,000 replications are permuted using Monte Carlo resampling in our analyses.

C. *The Stepdown Procedure*

Conducting permutation tests for each of the 53 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 et al. 2010). This procedure adjusts the *p*-values associated with each individual test to account for the effect of testing multiple outcomes.

The stepdown procedure involves placing each measure in a family of related outcomes and calculating a test statistic for each null hypothesis in the family of outcomes - we use the *t*-statistic. The test statistics for each measure are then placed in descending order within each family. 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 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 family 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 fails to be rejected, or only one hypothesis remains. 'Stepping down' through the hypotheses in this manner allow us to isolate the hypotheses that lead to rejection of the null. This method is superior to the well-known Bonferroni adjustment method as it accounts for interdependence across 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. 2010).

The 53 outcome measures are 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.[13] Table F.1 in Web Appendix F shows the stepdown families and the individual measures included in each, and Tables 3-6 in the following

[13] Note that outcomes derived from the same measure should not be included in the same family (e.g., the total score on a standardized instrument may not be included alongside the subdomains of that instrument).

sections are organized accordingly. Note that the composition of the stepdown families vary from wave to wave as, in some cases, different instruments were used at different waves.

Utilizing each of the subdomains of the child development instruments, we derive three stepdown families representing the main skill sets of children at 6, 12, and 18 months – cognitive development, noncognitive development, and physical development. The cognitive development stepdown family captures the children's communication and vocabulary skills, as well as their problem-solving abilities, and general cognitive development. The noncognitive development stepdown family captures the children's socio-emotional skills, temperament, and behavior. The physical development stepdown family captures the children's gross and fine motor skills.

Similarly, utilizing each of the subdomains of the parenting instruments, we derive six stepdown families at 6 months, two at 12 months, and five at 18 months. The stepdown families represent key areas of parental investment including the quality of the home environment provided, appropriate caregiving, parental interactions, parental attachment, parental self-efficacy, and parental beliefs. We selected these areas to best capture aspects of parenting which have been highlighted in Cunha et al (2013), Currie (2001), Heckman and Kautz (2013), and Hertwig et al. (2002) to be important for child development.

D. Inverse Probability Weighting

Due to attrition and non-response, the estimation sample sizes differ at each data collection point. In an attempt to address any bias that attrition, wave non-response or item non-response may introduce[14], we test the robustness of the main analysis using an inverse probability weighting (IPW) technique. Adapting from the description in Campbell et al.

[14] While the degree of item non-response was minimal for the majority of the instruments used (less than 2% at each time point), there were more substantial cases of missing data for some of the home environment, appropriate care and parental interaction measures. First, as some of these 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 environment and appropriate care measures 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 the parents who did not want the interview to be conducted in the home and the parents who did not want their child to be present for the interview. Thus, we apply IPW to deal with this issue. One exception is the *CDI-WG* scale items. This instrument is completed by parents using a paper form. 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 the measure should not be imputed (Fenson et al. 2000). It contains 104 items and it is likely that mothers could miss some questions. Therefore, the data are likely to be missing at random.

(2014), we make the assumption that the outcome is independent of the missing data pattern, conditional on treatment assignment and observable baselines characteristics. This assumption can be written as

(3) $$Y \perp M \mid (D, X)$$

Where Y is the outcome vector $Y = (Y_i: i \in I)$, $I = \{1 \ldots N\}$, M is a missing data indicator $M = (M_i: i \in I)$ where $M_i = 0$ denotes that Y_i is missing, $M_i = 1$ otherwise, D is the treatment indicator as before and X is a set of baseline measures used to predict M. It should be noted that this assumption does not hold if attrition patterns are determined by unobservable characteristics. Thus our analysis is based on the assumption that observable, baseline characteristics X capture the attrition pattern. The availability of rich baseline data, including 74 variables capturing measures of socioeconomic status, health, well-being, IQ, personality traits and parenting, increases our confidence that this assumption holds. However, we cannot formally test this. The probability that an observation has a non-missing outcome can be described as follows

(4) $$P_{i,d} = \Pr(M_i = 1 \mid X_i, D_i = d) \Pr(D_i = d \mid X_i) \quad d \in \{0, 1\}$$

Where Pr(.) is the probability function. $\hat{P}i$ represents the estimate of P_i which we calculate using a logit model. Thus, the weight that is assigned to each observation is defined simply as

(5) $$w_i = {}^{1}/_{\hat{P}_i}$$

In order to test the null hypothesis that the average treatment effect is zero, we calculate the following:

(6) $$\widehat{ATE} = \sum_{i=1}^{N} Y_i \cdot I(D_i = 1, M_i = 1) \cdot {w_i}/{N_1} - \sum_{i=1}^{N} Y_i \cdot I(D_i = 0, M_i = 1) \cdot {w_i}/{N_0}$$

Where

(7) $$N_d = \sum_{i=1}^{N} I\,(D_i = d, M_i = 1) \cdot w_i \quad d \in \{0, 1\},$$ $I(.)$ is the indicator function.

In practice, this method is applied to each outcome separately and involves two main steps: first, baseline data are used to predict each participant's probability of having a non-missing outcome. Observations classified as missing include participants who officially dropped out of the study, those who did not complete the questionnaire at that particular assessment (but may engage at another assessment point), as well as those who participated in the assessment but did not provide data for the corresponding outcome. The predicted probabilities from these logit

models are then applied as weights in the estimation of treatment effects such that a larger weight is applied to individuals that are underrepresented in the sample due to missing observations. The predicted probabilities of having non-missing outcomes were calculated using two separate logit models for the high and low treatment groups to account for differential processes driving the level of missing data associated with treatment assignment. See the Web Appendix D for the technical details of the IPW process.

IV. Results

A. Analysis of Treatment Effects

The impact of the program on child development and parental investment are presented in Tables 3 and 4, respectively. We present the mean outcome scores by treatment group, the *p*-values that result from classic *t* tests ($p^{(i)}$), individual permutation tests ($p^{(ii)}$), the adjusted *p*-values using Bonferroni adjustment ($p^{(iii)}$), and the adjusted *p*-values using the stepdown procedure ($p^{(iv)}$). These results are presented for each wave. 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 presented for the $p^{(iv)}$ values indicate the relative magnitude of the *t*-statistic within each stepdown family, which reflects the order in which the stepdown procedure is executed. Thus superscript 1 indicates the measure corresponding to the largest *t*-statistic. Each adjusted $p^{(iv)}$-value represents the likelihood of rejecting the joint null hypothesis when the variables of higher ordering are excluded. For example, in Table 3, the first adjusted $p^{(iv)}$-value (0.265) in the *Cognitive Development* family is the result of jointly testing the two outcomes in that family. The next adjusted $p^{(iv)}$-value (0.679) is the result of excluding the *Communication* score from the joint hypothesis test. Notice that this is the same as the unadjusted $p^{(ii)}$*-value* (0.679) as only one measure remains in the family (*Problem Solving*). Thus, as we step down through the hypotheses, the most statistically significant variables are excluded until only one measure remains in the subset. In this final step, the adjusted *p*-value is equivalent to the *p*-value that results from individual testing. We order this stepdown reporting in line with the 6 month data in our tables.

Child Development - In order to test the malleability of different skill sets, we separate child development into three categories: *Cognitive Development*, *Noncognitive Development*, and *Physical Development*. The results are presented in Table 3. Focusing on the stepdown

adjusted *p*-values, a rejection of the joint null hypothesis is found only for the *Physical Development* family. Specifically, at 12 and 18 months, when the fine motor score (which is a measure of the child's ability to engage in developmentally appropriate finger and hand movements) is tested in conjunction with the gross motor score (a measure of the child's ability to display developmentally appropriate movement skills such as walking and kicking), the joint null hypothesis can be rejected. The Bonferroni adjusted *p*-value confirms this result. At 12 months, the treatment effect is concentrated on fine motor skills, while at 18 months, the treatment effect is precisely determined for gross motor skills. It should be noted that the magnitude of the estimated effects is small to moderate (Cohen's *d* range = 0.26–0.28) for both fine and gross motor skills).

The lack of significant treatment effects in the *Cognitive Development* and *Noncognitive Development* stepdown families suggests the program's impact is limited to physical development at this stage. While one significant individual treatment effect is identified within the *Cognitive Development* family at 18 months (cognitive development score), the corresponding stepdown adjusted *p*-value indicates that the joint null hypothesis of no effect fails to be rejected.

Given that some of the measures examined are observed at multiple time points, we also investigated whether precision could be improved by taking the average of each outcome over the 3 waves examined. If a respondent had not taken part in all assessments, or if the measure was not collected at all wave, the average score was calculated based on the available non-missing scores. This increased the sample size for some of the repeated measures, as participation in any assessment, at any time point, would lead to a non-missing score. This analysis revealed a similar pattern to that presented in Table 3. Specifically, only one joint null hypothesis was rejected, and that was for the *Physical Development* domain.

In order to test for any potential program impacts in the non-hypothesized direction, we also perform left-sided hypothesis testing.[15] Although three individual significant differences in the non-hypothesized direction are found for words understood at 12 and 18 months and first communicative gestures score at 12 months, none of the left-sided stepdown adjusted *p*-values associated with Child Development outcomes at 6, 12, or 18 months are statistically significant.

[15] Left-sided *p*-values are not reported in the results tables. Left-sided *p*-values for individual hypothesis tests can be calculated directly from the right-sided *p*-values presented in the tables. Stepdown p-values, however, cannot be deduced from the tables. However, any significant results are noted in the text.

Table 3 – Treatment Effects for Child Development Outcomes

Stepdown Family		*6 Months*						*12 Months*						*18 Months*					
	Measure	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$
Cognitive Development	Communication	53.07 (7.84)	51.78 (8.49)	0.149	0.154	0.299	0.265[1]	49.88 (10.74)	50.18 (10.55)	0.572	0.575	1.000	0.769[5]	45.69 (13.16)	45.34 (13.96)	0.437	0.437	0.372	0.697[4]
	Problem Solving	51.87 (9.39)	52.56 (9.92)	0.680	0.679	-	0.679[2]	46.40 (11.71)	46.40 (13.13)	0.500	0.499	-	0.809[4]	45.69 (11.60)	45.07 (10.69)	0.366	0.369	-	0.714[3]
	Words Produced	-	-	-	-	-	-	57.34 (33.90)	55.08 (33.71)	0.383	0.383	-	0.762[3]	53.18 (29.97)	58.61 (26.50)	0.811	0.811	-	0.912[6]
	First Signs of Understanding	-	-	-	-	-	-	2.97 (0.16)	2.96 (0.20)	0.321	0.308	-	0.815[1]	2.99 (0.11)	2.94 (0.37)	0.178	0.178	-	0.554[2]
	First Communicative Gestures	-	-	-	-	-	-	9.04 (2.23)	9.71 (1.97)	0.971	0.972	-	0.998[6]	11.27 (1.37)	11.41 (1.26)	0.740	0.740	-	0.926[5]
	Words Understood	-	-	-	-	-	-	71.71 (26.61)	82.49 (17.01)	0.983	0.984	-	0.984[7]	64.89 (31.20)	73.51 (24.13)	0.922	0.923	-	0.923[7]
	Cognitive Development	-	-	-	-	-	-	116.20 (13.66)	115.13 (16.03)	0.324	0.323	-	0.730[2]	119.01 (15.83)	114.53 (17.94)	**0.053**	**0.053**	-	0.194[1]
Noncognitive Development	(-)Difficult Temperament	11.70 (5.71)	12.21 (5.50)	0.275	0.275	0.824	0.575[1]	12.60 (5.54)	13.30 (5.76)	0.216	0.216	0.928	0.424[4]	-	-	-	-	-	-
	Personal Social Score	46.52 (12.09)	45.94 (13.57)	0.384	0.383	-	0.595[2]	49.88 (8.82)	48.55 (10.46)	0.190	0.190	-	0.475[3]	50.88 (7.91)	49.46 (9.24)	0.155	0.160	0.621	0.396[1]
	(-)Social-Emotional Score	14.76 (10.68)	15.17 (13.75)	0.414	0.403	-	0.403[3]	23.48 (21.51)	21.14 (16.05)	0.784	0.779	-	0.779[6]	29.13 (19.92)	29.05 (31.84)	0.507	0.506	-	0.637[4]
	Competence	-	-	-	-	-	-	15.44 (3.41)	14.88 (3.57)	0.155	0.154	-	0.508[1]	17.85 (2.61)	17.59 (3.45)	0.304	0.305	-	0.530[2]
	(-)Problem Score	-	-	-	-	-	-	8.82 (5.74)	8.90 (6.49)	0.464	0.466	-	0.622[5]	9.44 (6.63)	9.14 (7.18)	0.607	0.606	-	0.606[5]
	(-)Atypical Behavior	-	-	-	-	-	-	0.95 (1.74)	1.23 (2.01)	0.171	0.175	-	0.489[2]	-	-	-	-	-	-
Physical Development	Gross Motor	40.78 (11.93)	38.50 (12.99)	0.115	0.117	0.230	0.207[1]	42.07 (18.34)	40.72 (18.27)	0.318	0.319	**0.098**	0.319[2]	56.31 (5.44)	53.72 (12.02)	**0.046**	**0.047**	**0.092**	**0.088[1]**
	Fine Motor	50.89 (9.47)	51.39 (10.17)	0.630	0.629	-	0.629[1]	54.33 (8.63)	51.87 (10.29)	**0.049**	**0.050**	-	**0.093[1]**	54.13 (8.26)	53.38 (8.28)	0.288	0.291	-	0.291[2]

Notes: '*M*' indicates the mean. 'SD' indicates the standard deviation. (i) one-tailed (right-sided) *p*-value from a t-test. (ii) one-tailed (right-sided) *p*-value from an individual permutation test with 100,000 replications. (iii) one-tailed (right-sided) *p*-value from Bonferroni adjustment. (iv) one-tailed (right-sided) *p*-value from a stepdown permutation test with 100,000 replications and the superscripts indicate the ordering in which the variables are dropped in the stepdown analysis from the largest to smallest *t*-statistic. (-) indicates the variable was reverse coded for the testing procedure. Statistically significant results are in bold.

Parenting - The impact of the program on the efficiency of parental investment is examined using six dimensions of parenting: physical environment, appropriate care, interactions with infant, maternal attachment, maternal self-efficacy and beliefs about parenting. The results are presented in Table 4. Focusing on the stepdown adjusted *p*-values, we find that the joint null hypothesis is rejected for the *Environment* family at 6 and 18 months, and the *Appropriate Care* family at 6 months. If Bonferroni adjustments were applied, we would fail to reject the joint null hypothesis for the *Environment* family at 6 months.

The *Environment* family includes outcomes reflecting the physical resources available to the child, the safety of the environment, and the quality of the child's day-to-day activities. The rejection of the null is driven by significant differences between the high and low treatment groups with respect to the learning materials available in the home at 6 months and the frequency of activities in the environment at 18 months (for example, trips to a grocery store and visits from relatives). These estimated treatments effects are small to moderate in magnitude (Cohen's *d* range = 0.36–0.37). Individual hypothesis testing also indicates a treatment effect on the frequency of activities at 6 months.

In the *Appropriate Care* family, which includes outcomes that relate to the absence of hostility and the presence of a father in the child's life, the stepdown adjusted *p*-value indicates that the joint null is rejected at 6 months, but not 18 months. The effect at 6 months is driven by differences between the high and low treatment groups in regards the variety of care available in the child's home. The estimated effect is small to moderate in size (Cohen's d = 0.41). An individual treatment effect is also found for a reduction in hostile-reactive behavior at 6 months. In addition, an individual treatment effect is also observed on acceptance at 18 months, which measures how accepting the mother is of the child's behavior, however the joint null on the *Appropriate Care* family fails to be rejected.

Although the joint null hypothesis of no treatment effect on the *Interactions* family also fails to be rejected, it is worth noting that one individual significant difference was found at 18 months for activities to stimulate development, which measures the frequency with which mothers conduct stimulating activities with the child such as playing peek-a-boo games, singing and storytelling.

Similarly, the results for the *Attachment* family at 6 months indicate that the stepdown procedure fails to reject the join null hypothesis of no treatment effect, while individual

hypothesis testing indicates evidence of a statistically significant impact on the baby comparison scale, which indicates the high treatment group are more likely to regard their baby more favorably compared with other babies, and the dysfunctional interactions scale.

For the two remaining parental investment families (*Parental Self-efficacy*, and *Parental Beliefs*), no statistically significant differences between the high and low treatment groups are found in the stepdown tests or the individual tests.

In the same manner in which we carried out a supplementary analysis of the child development results, we compared the high and low treatment groups with respect to the average of each parenting measure across the 3 waves. The results echoed the pattern in Table 4. Specifically, the joint null hypothesis was rejected for the *Environment* and *Appropriate Care* domains.

It should also be noted that we find no statistically significant effects in the non-hypothesized direction when carrying out individual or joint hypothesis testing for the parenting outcomes.

Table 4 – Treatment Effects for Parental Investment Decisions

Stepdown Family		*6 Months*						*12 Months*						*18 Months*					
	Measure	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$
Environment	Learning Materials	6.88 (1.65)	6.26 (1.72)	**0.020**	**0.021**	0.102	**0.096**[1]	-	-	-	-	-	-	8.24 (0.97)	8.04 (1.12)	0.174	0.176	**0.068**	0.503[2]
	Activities in Child Environment	2.33 (0.50)	2.18 (0.50)	**0.027**	**0.026**	-	0.103[2]	-	-	-	-	-	-	2.38 (0.44)	2.22 (0.42)	**0.014**	**0.014**	-	**0.068**[1]
	Physical Environment	7.22 (0.86)	7.15 (0.87)	0.327	0.326	-	0.682[3]	-	-	-	-	-	-	7.34 (1.12)	7.13 (1.03)	0.174	0.175	-	0.428[3]
	Organization	5.57 (0.64)	5.58 (0.66)	0.547	0.543	-	0.791[4]	-	-	-	-	-	-	5.52 (0.69)	5.45 (0.78)	0.293	0.290	-	0.494[4]
	Safety	7.37 (0.77)	7.46 (0.68)	0.781	0.782	-	0.782[5]	-	-	-	-	-	-	8.33 (0.98)	8.33 (0.93)	0.503	0.505	-	0.505[5]
Appropriate Care	Variety	3.54 (1.12)	3.11 (1.01)	**0.005**	**0.005**	**0.020**	**0.020**[1]	-	-	-	-	-	-	4.08 (1.00)	3.99 (1.05)	0.297	0.309	0.106	0.502[2]
	[(-)]Hostile-Reactive Behavior	0.80 (1.13)	1.06 (1.21)	**0.073**	**0.074**	-	0.207[2]	-	-	-	-	-	-	-	-	-	-	-	-
	Suitable Care Provided	9.61 (1.14)	9.52 (1.12)	0.340	0.341	-	0.557[3]	-	-	-	-	-	-	9.92 (1.24)	9.89 (1.55)	0.463	0.463	-	0.463[3]
	Acceptance	6.36 (0.56)	6.36 (0.60)	0.484	0.484	-	0.484[4]	-	-	-	-	-	-	6.12 (0.80)	5.66 (1.45)	**0.035**	**0.035**	-	0.100[1]
Parental Interactions	Activities to Stimulate Development	3.24 (0.91)	3.14 (0.79)	0.229	0.227	0.686	0.493[1]	-	-	-	-	-	-	4.05 (0.76)	3.87 (0.75)	**0.075**	**0.077**	0.449	0.293[1]
	Responsivity	8.83 (1.73)	8.55 (2.32)	0.278	0.276	-	0.439[2]	-	-	-	-	-	-	9.50 (1.59)	9.07 (2.08)	0.142	0.144	-	0.416[2]
	Involvement	4.28 (1.25)	4.40 (1.25)	0.699	0.697	-	0.697[3]	-	-	-	-	-	-	3.88 (1.47)	4.23 (1.56)	0.872	0.872	-	0.872[5]
	Mother Reads to Her Child	-	-	-	-	-	-	0.90 (0.30)	0.90 (0.30)	0.510	0.512	1.000	0.758[1]	0.94 (0.24)	0.95 (0.23)	0.588	0.550	-	0.786[4]
	Mother Reads to Her Child Every Day	-	-	-	-	-	-	0.46 (0.50)	0.53 (0.50)	0.815	0.814		0.814[2]	0.24 (0.43)	0.22 (0.41)	0.377	0.377	-	0.718[3]
Attachment	Baby Comparison	7.52 (1.92)	7.02 (1.91)	**0.044**	**0.045**	0.351	0.266[1]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Dysfunctional Interactions	17.03 (4.90)	18.28 (5.71)	**0.066**	**0.067**	-	0.328[2]	-	-	-	-	-	-	-	-	-	-	-	-

Stepdown Family		6 Months						12 Months						18 Months					
	Measure	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$
	Quality of Attachment	4.70 (0.29)	4.68 (0.38)	0.330	0.331	-	0.869[3]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Difficult Child	19.77 (4.99)	20.13 (5.55)	0.331	0.329	-	0.839[4]	-	-	-	-	-	-	-	-	-	-	-	-
	Pleasure in Interaction	4.34 (0.38)	4.34 (0.43)	0.551	0.552	-	0.964[5]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Parental Overprotection	6.18 (2.19)	6.13 (1.99)	0.557	0.555	-	0.920[6]	-	-	-	-	-	-	-	-	-	-	-	-
	Absence of Hostility	4.39 (0.53)	4.41 (0.53)	0.615	0.617	-	0.828[7]	-	-	-	-	-	-	-	-	-	-	-	-
	Parental Warmth	9.18 (1.17)	9.25 (1.28)	0.653	0.651	-	0.651[8]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Maternal Separation Anxiety	-	-	-	-	-	-	-	-	-	-	-	-	22.02 (5.89)	22.16 (5.19)	0.441	0.440	0.441	0.440
Parental Self-efficacy	[(-)]Control of Child's Behavior	6.92 (2.82)	7.22 (2.64)	0.231	0.234	1.000	0.624[1]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Child Control of Parent's Life	8.46 (3.37)	8.80 (3.09)	0.250	0.251	-	0.605[2]	-	-	-	-	-	-	-	-	-	-	-	-
	Parental Self-efficacy (PACOTIS)	8.80 (1.11)	8.68 (1.25)	0.269	0.270	-	0.544[3]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Parental Efficacy (PLOC)	6.67 (2.44)	6.76 (2.43)	0.410	0.409	-	0.599[4]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Parenting Distress	26.02 (7.98)	25.55 (7.36)	0.657	0.658		0.658[5]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Parenting Daily Hassles	-	-	-	-	-	-	-	-	-	-	-	-	31.43 (11.42)	29.87 (9.16)	0.817	0.818	0.817	0.818
Parental Beliefs	Perceived Parental Impact	7.25 (2.00)	7.04 (2.22)	0.258	0.260	0.773	0.586[1]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Parental Responsibility	12.57 (3.20)	12.86 (3.02)	0.277	0.274		0.508[2]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Parental Belief in Fate	9.79 (3.64)	9.92 (3.31)	0.403	0.404		0.404[3]	-	-	-	-	-	-	-	-	-	-	-	-
	Belief in Appropriate Punishment	-	-	-	-	-	-	6.40 (1.14)	6.29 (1.47)	0.290	0.289	1.000	0.730[1]	-	-	-	-	-	-
	Knowledge of Child	-	-	-	-	-	-	70.30 (7.81)	69.69 (6.81)	0.297	0.299	-	0.700[2]	-	-	-	-	-	-

Stepdown Family	*6 Months*						*12 Months*						*18 Months*					
Measure	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$
Development Realistic Expectations of Children	-	-	-	-	-	-	6.59 (1.90)	6.46 (1.89)	0.325	0.327	-	0.661[3]	-	-	-	-	-	-
Promoting Children's Independence	-	-	-	-	-	-	5.35 (2.23)	5.27 (2.13)	0.397	0.399	-	0.678[4]	-	-	-	-	-	-
Appropriate Parent-Child Roles	-	-	-	-	-	-	6.11 (2.15)	6.09 (2.23)	0.472	0.474	-	0.606[5]	-	-	-	-	-	-
Parental Empathy	-	-	-	-	-	-	4.93 (2.40)	4.94 (2.00)	0.514	0.515	-	0.515[6]	-	-	-	-	-	-

Notes: '*M*' indicates the mean. 'SD' indicates the standard deviation. [(i)] one-tailed (right-sided) *p*-value from a t-test. [(ii)] one-tailed (right-sided) *p*-value from an individual permutation test with 100,000 replications. [(iii)] one-tailed (right-sided) *p*-value from Bonferroni adjustment. [(iv)] one-tailed (right-sided) *p*-value from a stepdown down permutation test with 100,000 replications and the superscripts indicate the ordering in which the variables are dropped in the stepdown analysis from the largest to smallest *t*-statistic. [(-)] indicates the variable was reverse coded for the testing procedure. Statistically significant results are in bold.

B. Robustness Tests

Tables 5 and 6 present the IPW-adjusted weighted results and can be read in the same manner as Tables 3 and 4.

Child Development - Table 5 shows that correcting for attrition and non-response bias changes some of the child development results. As before, few significant differences between the high and low treatment groups emerge. At 12 and 18 months, the non IPW-results indicated a precisely determined treatment effect on *Physical Development*. In contrast, Table 4 shows that this effect is no longer significant when the IPW method is applied. In addition, the non IPW-results reported that the joint null hypothesis failed to be rejected for the *Cognitive Development* stepdown family. When IPW is applied, the joint null is rejected at 18 months, and this result is driven by a significant difference between the high and low treatment groups on their cognitive development scores. This result is replicated using Bonferroni adjustment. Note that when hypothesis testing is conducted in the non-hypothesized direction, the stepdown adjusted *p*-value associated with words understood at 18 months which is in the *Cognitive Development* family is also statistically significant.[16] Finally, when the IPW method is applied to the *Noncognitive Development* family, the joint null fails to be rejected at any time point, although the individual permutation tests indicates evidence of an effect on the personal social score which was not evident in the non IPW-results.

Overall, the IPW analysis suggests that when we correct for misrepresentation due to attrition and non-response bias, the original results for the *Noncognitive Development* family are echoed, although a more favorable treatment effect emerges for the *Cognitive Development* family. However, this must be balanced with a less favorable result for the measure of words understood at 18 months. Correcting for attrition and non-response also suggests that the effect on *Physical Development*, identified in the non IPW-results, may be spurious.

[16] A significant difference is found in the non-hypothesized direction for first communicative gestures, but the stepdown adjusted p-value is not statistically significant. No other significant differences are found in the non-hypothesized direction for child development outcomes when IPW is applied.

Table 5 – Treatment Effects for Child Development Outcomes – Inverse Probability Weighted

Stepdown Family		*6 Months*						*12 Months*						*18 Months*					
	Measure	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$
Cognitive Development	Communication	53.04 (7.71)	51.95 (8.34)	0.187	0.187	0.373	0.325[1]	49.91 (10.67)	52.00 (10.33)	0.899	0.782	1.000	0.977[5]	45.61 (12.91)	44.14 (13.72)	0.247	0.303	**0.029**	0.730[3]
	Problem Solving	51.69 (9.42)	52.66 (9.74)	0.747	0.738	-	0.738[2]	46.67 (11.40)	47.54 (11.05)	0.691	0.669	-	0.951[3]	45.53 (12.02)	44.47 (10.03)	0.277	0.295	-	0.705[4]
	Words Produced	-	-	-	-	-	-	56.63 (33.00)	61.97 (35.53)	0.756	0.638	-	0.945[4]	51.46 (28.37)	55.41 (27.84)	0.741	0.697	-	0.876[6]
	First Signs of Understanding							2.97 (0.18)	2.97 (0.17)	0.556	0.541		0.944[2]	2.99 (0.12)	2.95 (0.34)	0.228	0.328		0.773[2]
	First Communicative Gestures							8.94 (2.32)	9.84 (1.98)	0.997	0.994		0.994[7]	11.20 (1.51)	11.29 (1.37)	0.649	0.599		0.898[5]
	Words Understood	-	-	-	-	-	-	73.19 (24.54)	82.98 (22.05)	0.968	0.893	-	0.989[6]	65.79 (28.94)	81.21 (20.68)	0.997	0.995	-	0.995[7]
	Cognitive Development Score	-	-	-	-	-	-	116.09 (13.63)	115.70 (14.44)	0.430	0.429		0.907[1]	119.47 (15.89)	111.99 (18.43)	**0.004**	**0.023**	-	**0.084**[1]
Noncognitive Development	[(-)]Difficult Temperament	11.82 (5.85)	12.16 (5.39)	0.343	0.355	1.000	0.637[1]	12.75 (5.68)	12.64 (5.35)	0.550	0.542	1.000	0.887[1]	-	-	-	-	-	-
	[(-)]Social-Emotional Score	15.18 (11.26)	15.33 (13.35)	0.469	0.468	-	0.656[2]	24.17 (20.92)	21.80 (14.39)	0.800	0.802	-	0.922[5]	30.56 (20.42)	30.55 (28.88)	0.501	0.502	0.132	0.675[3]
	Personal Social Score	45.47 (13.03)	46.16 (13.40)	0.634	0.607	-	0.607[3]	49.56 (9.10)	50.85 (10.39)	0.800	0.651	-	0.937[4]	50.94 (7.92)	48.67 (8.96)	**0.049**	**0.072**	-	0.162[2]
	Competence Score	-	-	-	-	-	-	15.50 (3.36)	17.40 (3.00)	1.000	0.618	-	0.618[6]	17.78 (2.56)	16.82 (3.73)	**0.033**	0.116	-	0.195[1]
	[(-)]Problem Score	-	-	-	-	-	-	9.29 (5.93)	8.88 (5.81)	0.672	0.672	-	0.923[2]	9.74 (6.95)	9.02 (6.60)	0.744	0.729	-	0.729[4]
	[(-)]Atypical Behavior	-	-	-	-	-	-	0.97 (1.77)	0.75 (1.66)	0.792	0.617	-	0.938[3]	-	-	-	-	-	-
Physical Development	Gross Motor	40.50 (12.30)	39.10 (13.09)	0.235	0.257	0.469	0.393[1]	54.01 (8.72)	53.40 (9.77)	0.335	0.386	0.670	0.510[1]	56.43 (5.44)	54.93 (10.92)	0.143	0.154	0.286	0.278[1]
	Fine Motor	51.04 (9.36)	51.31 (10.20)	0.574	0.570	-	0.570[1]	41.80 (17.71)	44.04 (18.16)	0.789	0.651	-	0.651[2]	54.26 (7.98)	52.96 (7.87)	0.155	0.167	-	0.167[2]

Notes: '*M*' indicates the mean. 'SD' indicates the standard deviation. [(i)] one-tailed (right-sided) *p*-value from a t-test. [(ii)] one-tailed (right-sided) *p*-value from an individual permutation test with 100,000 replications. [(iii)] one-tailed (right-sided) *p*-value from Bonferroni adjustment. [(iv)] 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. Statistically significant results are in bold.

Parenting - Table 6 shows that when the IPW method is applied to the parenting outcomes, individual treatment effects are identified in the same stepdown families as the non IPW-results, however fewer remain statistically significant after stepdown adjustment. With respect to the *Environment* family, the non IPW-results indicated a rejection of the joint null hypothesis at 6 and 18 months, however the IPW adjustment leads to a failure to reject the joint null at either time points using the stepdown procedure. However, the joint null is rejected using the Bonferroni procedure at 18 months. Using individual hypothesis testing, a precisely determined treatment effect on the frequency of activities in the child's environment is identified at both 6 and 18 months. Regarding the *Appropriate Care* family, the rejection of the joint null hypothesis at 6 months is consistent with the non IPW-result, suggesting that the treatment leads to improvements on the variety of care provided in the home. However consistent with the non IPW-results at 18 months, we fail to reject the joint null hypothesis for the *Appropriate Care* family, and observe only one individual treatment effect on the mothers' acceptance of the child's behavior.

Examining the IPW results using the individual permutation tests indicate that across the *Environment*, *Appropriate Care*, *Interactions* and *Attachment* stepdown families, a similar pattern emerges to the non IPW-results. For the *Parental Self-efficacy* and *Parental Beliefs* families, consistent with the non IPW-results, no significant treatment effects are identified when IPW-adjusted permutation testing and the stepdown procedure is applied. Also, similar to the non IPW-results, when we carry out hypothesis testing for the parenting outcomes in the non-hypothesized direction we find no statistically significant differences resulting from individual or joint hypothesis testing.

Table 6 – Treatment Effects for Parental Investment Decisions – Inverse Probability Weighted

Stepdown Family		*6 Months*						*12 Months*						*18 Months*					
	Measure	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$
Environment	Activities in Child Environment	2.31 (0.51)	2.17 (0.52)	**0.037**	**0.055**	0.185	0.245[1]	-	-	-	-	-	-	2.39 (0.45)	2.24 (0.42)	**0.017**	**0.028**	**0.087**	0.252[1]
	Learning Materials	6.59 (1.62)	6.19 (1.84)	**0.098**	0.121	-	0.425[2]	-	-	-	-	-	-	8.21 (0.97)	8.24 (1.10)	0.566	0.534	-	0.534[5]
	Organization	5.66 (0.58)	5.60 (0.64)	0.272	0.282	-	0.664[3]	-	-	-	-	-	-	5.53 (0.68)	5.47 (0.81)	0.339	0.397	-	0.785[2]
	Physical Environment	7.19 (0.81)	7.13 (0.91)	0.330	0.357	-	0.605[4]	-	-	-	-	-	-	6.99 (1.27)	6.97 (1.06)	0.471	0.476	-	0.890[3]
	Safety	7.32 (0.79)	7.44 (0.70)	0.842	0.804	-	0.804[5]	-	-	-	-	-	-	8.39 (0.93)	8.39 (0.97)	0.493	0.496	-	0.806[4]
Appropriate Care	Variety	3.52 (1.12)	3.11 (1.00)	**0.007**	**0.008**	**0.027**	**0.077**[1]	-	-	-	-	-	-	4.00 (1.07)	3.79 (1.13)	0.129	0.207	0.244	0.304[2]
	Acceptance	6.44 (0.57)	6.30 (0.65)	**0.099**	0.171	-	0.402[2]	-	-	-	-	-	-	6.08 (0.81)	5.77 (1.22)	**0.081**	**0.076**	-	0.278[1]
	[(-)]Hostile-Reactive Behavior	0.79 (1.09)	1.00 (1.19)	0.116	0.116	-	0.288[3]	-	-	-	-	-	-	-	-	-	-	-	-
	Suitable Care Provided	9.38 (1.19)	9.47 (1.08)	0.670	0.622	-	0.622[4]	-	-	-	-	-	-	9.91 (1.23)	9.94 (1.30)	0.532	0.532	-	0.532[3]
Interactions	Activities to Stimulate Development	3.19 (1.02)	3.17 (0.79)	0.429	0.433	1.000	0.781[1]	-	-	-	-	-	-	4.02 (0.80)	3.87 (0.83)	0.127	0.215	0.636	0.549[1]
	Involvement	4.25 (1.15)	4.33 (1.25)	0.640	0.633	-	0.839[2]	-	-	-	-	-	-	3.90 (1.42)	4.39 (1.46)	0.949	0.891	-	0.891[5]
	Responsivity	8.52 (1.90)	8.81 (2.00)	0.738	0.722	-	0.722[3]	-	-	-	-	-	-	9.40 (1.64)	8.97 (1.93)	0.133	0.149	-	0.492[2]
	Mother Reads to Her Child	-	-	-	-	-	-	0.90 (0.31)	0.92 (0.27)	0.703	0.682	1.000	0.896[1]	0.93 (0.26)	0.95 (0.22)	0.706	0.685	-	0.854[4]
	Mother Reads to her Child Every Day	-	-	-	-	-	-	0.47 (0.50)	0.55 (0.50)	0.826	0.814		0.814[2]	0.23 (0.43)	0.21 (0.41)	0.360	0.372	-	0.742[3]
Attachment	[(-)]Dysfunctional Interactions	16.84 (4.90)	18.41 (5.88)	**0.032**	**0.043**	0.258	0.250[1]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Difficult Child	19.33 (5.02)	20.21 (5.71)	0.147	0.171	-	0.659[2]	-	-	-	-	-	-	-	-	-	-	-	-
	Baby Comparison Scale	7.48 (1.91)	7.18 (1.94)	0.149	0.168	-	0.627[3]	-	-	-	-	-	-	-	-	-	-	-	-

Stepdown Family		*6 Months*						*12 Months*						*18 Months*					
	Measure	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$
	Quality of Attachment	4.71 (0.29)	4.70 (0.36)	0.422	0.425	-	0.935[4]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Parental Overprotection	6.27 (2.19)	6.28 (2.03)	0.481	0.481	-	0.946[5]	-	-	-	-	-	-	-	-	-	-	-	-
	Pleasure in Interaction	4.36 (9.16)	4.37 (0.42)	0.581	0.579	-	0.881[6]	-	-	-	-	-	-	-	-	-	-	-	-
	Parental Warmth	9.16 (1.16)	9.25 (1.30)	0.677	0.665	-	0.857[7]	-	-	-	-	-	-	-	-	-	-	-	-
	Absence of Hostility	4.40 (0.52)	4.44 (0.54)	0.706	0.691	-	0.691[8]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Maternal Separation Anxiety Scale	-	-	-	-	-	-	-	-	-	-	-	-	22.08 (6.42)	21.75 (5.85)	0.627	0.567	0.627	0.567
Parental Self-efficacy	[(-)]Child Control of Parent's Life	8.42 (3.37)	9.06 (3.14)	0.102	0.125	0.508	0.405[1]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Control of Child's Behavior	6.76 (2.79)	7.21 (2.57)	0.137	0.141	-	0.419[2]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Parental Efficacy (PLOC)	6.70 (2.46)	6.88 (2.45)	0.323	0.33	-	0.639[3]	-	-	-	-	-	-	-	-	-	-	-	-
	Parental Self-efficacy (PACOTIS)	8.82 (1.11)	8.81 (1.24)	0.497	0.496	-	0.712[4]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Parenting Distress	26.38 (8.09)	25.69 (7.29)	0.721	0.712	-	0.712[5]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Parenting Daily Hassles	-	-	-	-	-	-	-	-	-	-	-	-	31.24 (11.30)	30.04 (9.96)	0.750	0.710	0.750	0.710
Parental Beliefs	[(-)]Parental Responsibility	12.28 (3.31)	12.81 (3.03)	0.137	0.152	0.411	0.378[1]	-	-	-	-	-	-	-	-	-	-	-	-
	Perceived Parental Impact	7.19 (1.98)	7.15 (2.20)	0.446	0.449	-	0.647[2]	-	-	-	-	-	-	-	-	-	-	-	-
	[(-)]Parental Belief in Fate	10.00 (3.60)	10.00 (3.34)	0.495	0.494	-	0.494[3]	-	-	-	-	-	-	-	-	-	-	-	-
	Belief in the Use of Appropriate Punishment	-	-	-	-	-	-	6.42 (1.11)	6.23 (1.33)	0.162	0.164	0.410	0.531[3]	-	-	-	-	-	-
	Knowledge of Child Development	-	-	-	-	-	-	70.48 (7.95)	68.83 (7.16)	**0.083**	0.159	-	0.503[2]	-	-	-	-	-	-
	Realistic Expectations of Children	-	-	-	-	-	-	6.64 (1.93)	6.89 (2.01)	0.786	0.611	-	0.611[6]	-	-	-	-	-	-

Stepdown Family	*6 Months*						*12 Months*						*18 Months*					
Measure	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$
Promoting Children's Independence	-	-	-	-	-	-	5.33 (2.23)	4.82 (2.11)	**0.068**	0.200	-	0.572[1]	-	-	-	-	-	-
Appropriate Parent-Child Roles	-	-	-	-	-	-	6.09 (2.13)	6.23 (2.24)	0.664	0.649	-	0.747[5]	-	-	-	-	-	-
Parental Empathy	-	-	-	-	-	-	4.98 (2.45)	5.07 (2.01)	0.595	0.584	-	0.771[4]	-	-	-	-	-	-

Notes: '*M*' indicates the mean. 'SD' indicates the standard deviation. [(i)] one-tailed (right-sided) *p*-value from a t-test. [(ii)] one-tailed (right-sided) *p*-value from an individual permutation test with 100,000 replications. [(iii)] one-tailed (right-sided) *p*-value from Bonferroni adjustment. [(iv)] 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. Statistically significant results are in bold.

C. Reliability of Instruments

One potential limitation of our study is that the majority of child and parent instruments 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. For example, there is evidence of low-to-moderate cross-informant correlations in terms of child behavioral/emotional problems (Achenbach et al. 1987); and a study conducted within the PFL catchment area found that parents in the community systematically reported higher child skill levels compared to teacher reports (Doyle et al. 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 program impact 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 program effectiveness.

To test whether the participants in the high and low treatment groups differ with respect to the level of social desirability exhibited, we examine a measure of defensive responding which is a defined subdomain of the Parenting Stress Index administered to the participants at the six month assessment. The measure is derived from the widely used Crowne-Marlowe Social Desirability Scale and the questions pertain to routine parenting experiences, thus a denial of these experiences can be interpreted as defensive, rather than accurate, responding. A score of 10 or above on this measure may suggest defensive responding. Although the high (M = 15.24, SD = 4.82) and low (M = 14.99, SD = 4.50) treatment groups both scored above 10 on average, they did not differ significantly in their scores (p-value > 0.1). Thus, it appears that the group do not differ systematically in their level of misreporting.

As an additional test of the reliability of the self-reported measures we took advantage of the availability of interviewer reported observational items in the *Home Observation for Measurement of the Environment (HOME)* parenting scale. Consistent with the main results, restricting our analysis to solely observational items at 18 months led to a statistically significant difference between the high and low treatment groups with respect to the 'acceptance' measure within the Appropriate Care stepdown family. One additional significant effect was also found with respect to the 'involvement' measure in the Parental

Interactions stepdown family which could suggest that an element of noise associated with the self-reported measures is masking some true effects.[17]

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 et al. 1999). In particular, the overall level of agreement between the ASQ and standardized assessments such as the Bayley Scales of Infant Development (Bayley 1969) is 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 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 et al. 2001). Overall, the literature suggests that there is considerable agreement between the ASQ and standardized measures that are conducted by professionals.

Finally, we also examined the level of correlation between each of the child development and parenting outcomes (at 6, 12, and 18 months) with the participating mothers' IQ scores in order to test for measurement error in the outcome variables. If measurement error is low, we would expect to find a positive correlation between child and parent outcomes and mothers' IQ scores, in keeping with the gradient typically found in these associations (i.e. mothers with higher IQ tend to have better parenting skills, as well as children with more advanced skills; Plomin et al. 2001; Sattler 2008). Overall, we found that the level of correlation was low to moderate: the Pearson's correlation coefficient (*r)* was less than 0.3 for 76% of the outcomes examined, and ranged from 0.3-0.5 for the remaining 26% of outcomes. As these data do not indicate a clear gradient in the outcomes examined, it is possible that an element of measurement error is obscuring some of the program impacts.

[17] None of the measures where program effects were found at 6 months contain solely observational items. Thus, it was not possible to check if the treatment effects identified at 6 months remained after restricting the analysis to observational items. No new effects emerged for any of the observational measures at 6 months.

D. Testing for Contamination

As the PFL program is taking place in one small community, the potential for contamination is high given the geographical proximity of the participants. Contamination, also known as spillover effects (Bloom 2005), may occur if participants in the high treatment group engage in cross-talk or share their parenting materials, information, strategies, or advice which they receive from their mentors, with participants in the low treatment group. A number of strategies were devised to measure cross-talk and information flows between the two groups (information on these strategies can be found in Doyle and Hickey 2013).

At 6 months, we tested for the presence of contamination using 'blue-dye' questions, whereby participants from both groups were asked if they have heard of two particular parenting phrases - *'mutual gaze*' and *'circle of security*' and if they knew what these phrases meant. In theory, a greater proportion of high treatment group participants should be aware of these terms as the mentors spend time discussing these parenting techniques when delivering the program prior to 6 months. In particular, there are Tip Sheets on both '*mutual gaze*' and *'circle of security*'.[18] These questions may be used as proxies for contamination as, if participants in the low treatment group report knowledge of these phrases and they accurately describe how to promote these behaviors, it suggests that they may have accessed material intended for the high treatment group only.

The first two rows in Table 7 show that significantly greater proportions of the high treatment group report knowledge of '*mutual gaze*' and *'circle of security*' (59% and 49%) compared to the low treatment group (8% and 12%). While this suggests a lack of contamination, it is possible that parents are simply reporting knowledge of the term without knowing how to engage in these behaviors. Therefore, in order to provide a more accurate measure of contamination, participants who stated that they had heard of these parenting phrases, yet provided incorrect responses regarding how best to engage in these behaviors, were removed from the analysis. Tests comparing the two groups were re-estimated and are presented in the third and fourth rows of Table 7. They show that significantly greater proportions of the high treatment group (54% and 33%) than the low treatment group (7% and 10%) report knowledge of the phrases and accurately know how to engage in them. Again, this suggests that contamination may not be an issue in this sample.

[18] These terms are commonly used in parenting books and online parenting resources. As they are not *PFL* specific terms it is possible that some members of the low treatment group may have heard of the terms also.

This analysis is limited as only two areas of parenting are examined. It is possible that the high treatment group may have shared material about other parenting or child development resources. However, in the absence of alternative measures, this proxy suggests that contamination may be low in the *PFL* trial at 6 months. Indeed, as *PFL* is a complex intervention which aims to change parental behavior by building relationships of trust between mentors and participants, minimal contamination may be expected as it is often difficult to achieve behavioral change. Thus, even if contamination between the two groups exists, it may not be enough to meaningfully affect the results (Howe et al. 2007).

Table 7 –Testing for Contamination Across Groups

	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	Permutation Test *p*
Heard the phrase 'Mutual Gaze'	0.59 (0.49)	0.08 (0.27)	**0.000**
Heard the phrase 'Circle of Security'	0.49 (0.50)	0.12 (0.33)	**0.000**
Heard the phrase 'Mutual Gaze' & accurately reports how to engage in this behavior	0.54 (0.50)	0.07 (0.27)	**0.000**
Heard the phrase 'Circle of Security' & accurately reports how to engage in this behavior	0.33 (0.47)	0.10 (0.30)	**0.000**

Note: 'M' indicates the mean. 'SD' indicates the standard deviation. [1]two-tailed p value from an individual permutation test with 100,000 replications.

V. Conclusion

This study investigates the effectiveness of investment in an Irish early childhood, home visiting intervention from *in utero* to 18 months of age on key indicators of early skill formation and parenting skills. Rigorous evaluation of early intervention programs has received relatively little attention in Europe, yet given the social, economic, and cultural differences, especially with respect to welfare systems, it cannot be assumed that the findings from the seminal American studies can be replicated. In this study, permutation testing, a stepdown procedure, and inverse probability weighting are applied to account for small sample size, multiple hypothesis testing, and attrition. Overall, we find evidence of significant, robust treatment effects on some dimensions of parental investment, specifically on the quality of the child's environment and level of appropriate care provided to the child,

but the evidence of effects on child development is inconclusive and the effect sizes for the parenting and home environment outcomes are small to moderate.

With respect to the parental investment measures, all significant treatment effects are in the hypothesized direction. In home visiting programs such as PFL, parents are conceived as the primary mechanism for change. Thus the main avenue by which a child's skills can be developed and enhanced is through changes in parenting skills and abilities. Hertwig et al. (2002) suggest that various dimensions of parenting practices (such as material resources, cognitive stimulation and parental interpersonal skills) may impact on disparate areas of child development. However, it is unclear how malleable these dimensions of parenting are to intervention given the lack of impact demonstrated in the majority of the home visiting literature. In addition, according to Shonkoff and Phillips (2000) parenting behavior is extremely difficult to change.

By examining multiple dimensions of parenting encompassing 38 measures, our analysis suggests that home visiting programs can be an effective means of improving some deficits in parenting skills within a relatively short time frame, particularly regarding the level of appropriate care provided to children and the quality of the home environment. These dimensions relate to material resources that can be observed in the household and activities that the mother carries out with her child. The questions used to measure these skills refer to what the mother *does* with her infant rather than questions about parenting style. Thus, it is possible that the intervention has an impact on tangible aspects of parenting rather than the more subjective maternal perceptions and beliefs. This is consistent with the analysis of Brooks-Gunn and Markman (2005) who state that parenting interventions may be more effective at changing parental behavior rather than parental emotional states. These results are also consistent with the PFL curriculum up to 18 months. For example, a number of the Tip Sheets delivered during this period focus on promoting activities which encourage 'learning through play' and 'establishing a daily routine', as well as encouraging parents to create a safe environment for their child and acknowledge their development.

In terms of the home visiting literature, few studies to date have identified significant effects on parenting outcomes within the first 18 months. Exceptions include LeCroy and Krysik (2011), who found treatment effects on measures related to the quality of the environment and appropriate care, and Minkovitz et al. (2001), who identified a program impact on appropriate care between 2-4 months. Thus our results concerning improvements

in the quality of the environment and appropriate care at 6 and 18 months are consistent with these studies regarding the areas of parenting that can be impacted by home visiting. Yet it is important to note that comparing the results from different home visiting programs is complicated by significant variation across programs with respect to the types of families targeted by the intervention, the timing of program delivery, the program's content, intensity and duration, the background of the home visitors, as well as the goals and outcomes of the program (Gomby et al. 1999).

While the PFL program has some impact on certain dimensions of parenting, which may have been developed through interactions with mentors and materials, it may take time for these new parenting strategies and skills to have a direct impact on infant behavior and development. Indeed, we find little evidence of improvements on key dimensions of child development by 18 months. The only significant impact in the main results relates to the child's physical development, which may be attributed to the curriculum's concentration on promoting physical health and well-being in the Tip Sheets. Indeed, almost 60% of all Tip Sheets delivered focus on improving the child's physical development. However, as this result is not replicated in the IPW analysis, caution should be applied in placing too much weight on this finding.

There are a number of potential explanations which may account for the absence of short term effects on the majority of child outcomes. Within the first 18 months of life there is considerable variability in the rate at which children develop. For example, some 'normally' developing children meet their developmental milestones later than others, without any observable long-term negative effects. Such variation makes it difficult to detect differences in problematic developmental delays within the first two years. This may explain why this study, similar to the majority of the home visiting literature, fails to identify early effects on child development.

The lack of effects also may be attributed to the timing of assessment and dosage. The PFL program is a five year intervention and this study was based on data collected when parents were exposed to the program for approximately two years (pregnancy - 18 months). The high treatment participants had received, on average, 29 home visits during this time. The program's theory of change suggests that the intervention's impact on child outcomes will be mediated by changes in parenting behavior. 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 forming (Ammerman et al. 2010). Indeed in a qualitative analysis of the PFL program (Lovett et al., 2016), describing the barriers to early engagement as elicited from focus groups and semi-structured interviews when the participating children were, on average, 5 months old, the authors note that the parent-mentor relationship was perceived as strengthening over time. As the intervention progresses and the relationship becomes stronger, changes in parenting behavior may lead to changes in child development. For example, improved parenting behaviors and an enhanced home environment at age 2 were found to mediate treatment effects on cognitive scores at age 6 in the Nurse Family Partnership program (Heckman et al. 2014). The PFL programs costs approximately $US 2,250 (€2,000) per family per year to be delivered. A cost–benefit analysis of multiple, primarily US-based, home visiting programs found returns ranging from $US 0·21 to $US 30·46 per $US invested, with a median return of $US 1·62 (Washington State Institute for Public Policy, 2014). If the effects identified in the present paper regarding early parenting skills are a mechanism for promoting optimal child development at later ages, it is possible that the intervention may generate similar positive returns in the long run; thus future work should include a cost–benefit analysis.

In conclusion, this study demonstrates that home visiting programs can be effective at raising the efficiency of parental investment in children during infancy, yet continued investment and assessment may be required to observe direct effects on skill formation.

References

Achenbach, T.M., McConaughy, S.H., and Howell, C.T. 1987. "Child/adolescent Behavioral and Emotional Problems: Implications of Cross-informant Correlations for Situational Specificity." *Psychological Bulletin* 101: 213-232.

Ammerman, R.T., Putnam, F.T., Bosse, N.R., Teeters, A.R., and Van Ginkel, J.B. 2010. "Maternal Depression in Home Visitation: A Systematic Review." *Aggression and Violent Behavior* 15 (3): 191-200.

Avellar, S., and Paulsell, D. 2011. *Lessons Learned from the Home Visiting Evidence of Effectiveness Review*. Princeton, NJ: Mathematica Policy Research.

Barham, T., Macours, K. and Maluccio, J. A. 2013. "Boys' Cognitive Skill Formation and Physical Growth: Long-Term Experimental Evidence on Critical Ages for Early Childhood Interventions." *American Economic Review* 103 (3): 467-471.

Barnes, K.E. 1982. *Preschool Screening: The Measurement and Prediction of Children At-Risk.* Springfield, IL: Charles C. Thomas.

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

Benjamini, Y., and 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.

Bloom, H.S. 2005. "Randomizing Groups to Evaluate Place-based Programs". In: *Learning More From Social Experiments: Evolving Analytic Approaches* by Bloom, H.S. (ed.) New York: Russell Sage Foundation.

Brooks-Gunn, J., and Markman, L.B. 2005. "The Contribution of Parenting to Ethnic and Racial Gaps in School Readiness." *The Future of Children* 15 (1): 139-168.

Brooks-Gunn, J., Berlin, L.J., and Fuligni, A.S. 2000. "Early Childhood Intervention Programs: What About the Family?" In *Handbook of Early Childhood Intervention* edited by Shonkoff, J.P. and Meisels, S.J. Third Edition. New York: Cambridge University Press.

Campbell, F., Conti, G., Heckman, J.J, Moon, S.H., Pinto, R., Pungello, E., and Pan, Y. 2014. "Early Childhood Investments Substantially Boost Adult Health." *Science* 343 (6178): 1478-1485.

Conti, G., Heckman, J.J, and Pinto, R. 2015. "The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behaviors." *NBER Working Paper* No. 21454.

Cunha, F., Elo, I., and Culhane, J. 2013. "Eliciting Maternal Expectations about the Technology of Cognitive Skill Formation." *NBER Working Paper* No. 19144.

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

—, —, and Schennach, S.M. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78: 883–931.

Currie, J. 2001. "Early Childhood Education Programs." *The Journal of Economic Perspectives* 15 (2): 213–238.

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*. XLI, 92-111.

— , Finnegan, S., and 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.

— , and Hickey, C. 2013. "The Challenges of Contamination in Evaluations of Childhood Interventions." *Evaluation* 19(2): 180-191.

— , and 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.

Drotar, D., Robinson, J., Jeavons, L., and Lester Kirchner, H. 2009. "A Randomized, Controlled Evaluation of Early Intervention: The Born to Learn Curriculum." *Child: Care, Health & Development* 35 (5): 643–649.

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., Fuddy, L., Burrell, L., Higman, S.M., Windman, A.M, and Sia, C. 2004. "Randomized Trial of a State-wide Home Visiting Program: Impact in Preventing Child Abuse and Neglect." *Child Abuse and Neglect* 28 (6): 597-622.

— , — , Windham, A.M., Rohde, C.A., Salkever, D.S., Fuddy, L., Rosenberg, L.A., Buchbinder, S.B., and Sia, C.C. 1999. "Evaluation of Hawaii's Healthy Start Program." *The Future of Children* 9 (1): 66-90.

Elango, S., García, J.L., Heckman, J.J., and Hojman, A.P. 2015. "Early Childhood Education". NBER Working Paper No. 21766. Forthcoming in Economics of Means-Tested Transfer Programs in the United States, Volume 2, Moffitt.

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S. and Reznick, J. S. 2000. "Short-form Versions of the MacArthur Communicative Development Inventories." *Applied Psycholinguistics* 21(01): 95-116.

Gertler, P.J, Heckman, J.J., Zanolini, A., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., Chang, S.M., and Grantham-McGregor, S. 2014. "Labor market returns to an early childhood stimulation intervention in Jamaica." *Science* 344(6187): 998-1001.

Gollenberg, Audra L., C. D. Lynch, L. W. Jackson, , M. McGuinness and M. E. Msall. 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., Culross, P.L., and Behrman, R.E. 1999. "Home visiting: recent program evaluations-analysis and recommendations." *Future Child* 9: 4–26.

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

Halfon, N., Shulman, E., Hochstein, M. 2001. "Brain development in early childhood." In *Building Community Systems for Young Children*, edited by Halfon, N., Shulman, E., Hochstein, M. Los Angeles: UCLA Center for Healthier Children, Families and Communities.

Heckman, J.J. 2000. "Policies to Foster Human Capital." *Research in Economics* 54(1): 3-56.

—. 2007. "The Economics, Technology and Neuroscience of Human Capaility Formation." *Proceedings of the National Academy of Sciences* 104(33):13250-13255.

—, Holland, M., Makino, K., Olds, D., Pinto, R., and Rosales, M. 2014 "The Nurse Family Partnership Program: a Reanalysis of the Memphis Randomized Controlled Trial". University of Chicago, mimeo.

—, and Kautz, T. 2012. "Hard Evidence on Soft Skills." *Labour Economics* 19(4): 451-464.

—, and Kautz, T. 2013. "Fostering and Measuring Skills: Interventions that Improve Character and Cognition." *NBER Working paper,* No. 19656.

—, Moon, S.H., Pinto, R., Savelyev, P.A., and Yavitz, A. 2010. "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the HighScope Perry Preschool Program." *Quantitative Economics* 1 (2): 1-46.

—, and Mosso, S. 2014. "The Economics of Human Development and Social Mobility." *Annual Review of Economics* 6 (1): 689–733.

—, Pinto, R., and Savelyev, P.A. 2013. "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103: 2052–86.

Hertwig, R., Davis, J.N., and Sulloway, F.J. 2002. "Parental Investment: How an Equity Motive can Produce Inequality." *Psychological Bulletin* 128 (5): 728–745.

Howard, K.S., and Brooks-Gunn, J. 2009. "The Role of Home-visiting Programs in Preventing Child Abuse and Neglect". *The Future of Children* 19 (2): 119-46.

Howe, A., Keogh-Brown, M., Miles, S., and Bachmann, M. 2007. "Expert Consensus on Contamination in Educational Trials Elicited by a Delphi Exercise." *Medical Education* 41: 196–204.

Janus, M., and Duku, E.K. 2005. "Development of the Short Early Development Instrument (S-EDI)". *Report for the World Bank*, Available from http://www.offordcentre.com/readiness/files/REPORT.short_edi_june2005.pdf. Accessed Sept, 2014.

Johnston, B.D., Huebner, C.E., Tyll, L.T., Barlow, W.E., and 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.

Kautz, T., Heckman, J.J., Diris, R., ter Weel, B., and Borghans, L. 2014. Fostering and Measuring Skills: Improving Cognitive and Non-cognitive Skills to Promote Lifetime Success. OECD Education Working Papers No. 110. Paris Organisation for Economic Co-operation and Development (OECD).

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

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

Landsverk, J., Carrilio, T., Connelly, C. D., Ganger, W., Slymen, D., Newton, R., et al. 2002. *Healthy Families San Diego clinical trial: Technical report.* San Diego, CA:

The Stuart Foundation, California Wellness Foundation, State of California Department of Social Services: Office of Child Abuse Prevention.

LeCroy, C.W., and Krysik, J. 2011. "Randomized Trial of the Healthy Families Arizona Home Visiting Program." *Children and Youth Services Review* 33 (10): 1761-1766.

Lovett J., Palamaro Munsell, E., McNamara, K., and Doyle, O. 2016. "Friend, Foe or Facilitator? The Role of the Parent-service provider Relationship in the Early Implementation of a Family-based Community Intervention." *Community Psychology in Global Perspective* 2(1): 52-72.

Minkovitz, C., Strobino, D., Hughart, N., Scharfstein, D., and Guyer, B., and Healthy Steps Evaluation Team. 2001. "Early Effects of the Healthy Steps for Young Children Program." *Archives of Pediatrics & Adolescent Medicine* 155: 470-479.

Nelson, C.A. 2000. *"The Neurobiological Bases of Early Intervention"*.In *Handbook of Early Childhood Intervention*, edited by Shonkoff, J.P. and Meisels, S.J. Second Edition (pp. 204-227). Cambridge, MA: Cambridge University Press.

— , Eckenrode, J., Henderson, C.R., Kitzman, H., Powers, J.H., Cole, R., Sidora, K., Morris, P., Pettitt, L.M., and 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.

Olds, D.L., Henderson, C.R., Cole, R., Eckenrode, J., Kitzman, H., Luckey, D., Pettitt, L., Sidora, K., Morris, P., and Powers, J. 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.

Paulsell, D., Avellar, S., Martin, E.S., and Del Grosso, P. 2010. *Home Visiting Evidence of Effectiveness Review: Executive Summary.* Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Washington, DC.

Plomin, R., and DeFries, J.C., McClearn, G.E., and McGuffin, P. 2001. *Behavioral genetics* (4th ed.), New York: Worth Publishers.

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

Preparing for Life and The Northside Partnership. 2008. "Preparing for Life Programme Manual." Available from:

www.preparingforlife.ie/sites/default/files/pfl_manual_may_2008.pdf. Accessed Aug, 2014.

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

—, Shaikh, A., and Wolf, M. 2010. "Hypothesis Testing in Econometrics." *Annual Review of Economics* 1 (2): 75-104.

Sandner, M. 2013. "Effects of Early Childhood Intervention on Child Development and Early Skill Formation. Evidence from a Randomized Controlled Trial." *Hannover Economic Papers (HEP)* No. dp-518.

Sattler, J.M. 2008. *Assessment of Children: Cognitive Foundations* (5th ed.), San Diego, CA: Jerome M. Sattler Publisher.

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

Skellern, C.Y., Rogers, Y., and 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.

Squires, J., Potter, L. and Bricker, D. D. 1999. *The ASQ User's Guide,* Baltimore, MD: Brookes Publishing Co.

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.Washington State Institute for Public Policy. 2014. "Benefit–Cost Results – Public Health and Prevention". Retrieved 23rd April 2015 from http://www.wsipp.wa.gov/BenefitCost.

Weaver, I.C., Cervoni, N., Champagne, F.A., D'Alessio, A.C., Sharma, S., Seckl, J.R., Dymov, S., Szyf, M., and Meaney, M.J. 2004. "Epigenetic programming by maternal behavior." *Nature Neuroscience* 7 (8): 847-54.

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

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

Williams, J., Greene, S., McNally, S., Murray, A., and Quail, A. 2010. *Growing Up in Ireland: The Infants and Their Families: Report 1*. Dublin: Government Publications.

Wynder, E.L. 1998. “Introduction to the report on the conference on the "critical" period of brain development.” *Preventative Medicine* 27 (2): 166-7.