Life Course Centre Working Paper Series

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

Orla Doyle
UCD School of Economics & UCD Geary Institute, University College Dublin

Colm Harmon
School of Economics, The University of Sydney & IZA

James Heckman
Department of Economics, The University of Chicago; IZA & NBER

Caitriona Logue
UCD School of Economics, University College Dublin

Seong Moon
Department of Economics, The University of Chicago

A more recent version of this paper was published as Doyle O, Harmon C, Heckman J, Logue C and Moon, S. (2017) Early Skill Formation and the Efficiency of Parental Investment: A Randomized Controlled Trial of Home Visiting. Labour Economics, 45, 40-58

No. 2015-09
May 2015

NON-TECHNICAL SUMMARY

Investing in early childhood is increasingly recognized as a key policy mechanism for reducing social disadvantage. Early investment is both equitable and can generate large economic returns in adulthood. 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. Therefore, we examine the impact of an experimentally designed, five-year home visiting program in Ireland - Preparing for Life (PFL) – on parenting skills and early child development. We use methods which deal with some of the common issues that arise when using a randomised controlled trial design. In particular, we take account of the small number of parents included in studies of this kind, we note and account for issues which arise when multiple outcome measures are examined, and finally, we adjust for the fact that not all parents will remain in the study for the full duration.

Preparing for Life (PFL) is a community-based home visiting program which works with families from pregnancy until the children start school at age 4/5 years. The program recruited and randomized 233 pregnant women from a socioeconomically disadvantaged community in Dublin, Ireland into an intervention or control group. The intervention included regular home visits delivered by a trained mentor, as well as a separate parenting course which started when the child was 2 years old. We measured a range of parental investment measures and child cognitive, non-cognitive and physical development when the children in the intervention and control group were 6, 12, and 18 months old.

Overall, we find that the program has an impact on some dimensions of parental investment, specifically on the quality of the child's environment and level of appropriate care provided to the child. However, the program had relatively few effects on child development. We found one effect on the child's cognitive development at 18 months, which is driven by higher cognitive development scores among the intervention group.

This suggests that home visiting programs can be effective at offsetting deficits in parenting skills within a relatively short timeframe, yet continued investment may be required to observe direct effects on child development.

ABOUT THE AUTHORS

Orla Doyle is a lecturer in the School of Economics and a Research Fellow in the Geary Institute for Public Policy in University College Dublin. Her research focuses on the economics of human development, and she is principal investigator on several experimental/quasi-experimental evaluations of early childhood interventions. Recent publications include articles on childcare and cognitive outcomes in *Child Development*; home visiting programs and perinatal outcomes in *European Journal of Obstetrics and Gynaecology;* and university access programmes in *Economics of Education Review*. Email: orla.doyle@ucd.ie

Colm Harmon is Professor of Economics at the University of Sydney since 2012, he is also a Chief Investigator on the Life Course Centre. Prior to Sydney he was Professor at University College Dublin (UCD) and Director of the UCD Geary Institute. His main research interests are in the economics of education, labour economics and applied economics policy. Recent publications include work on parental SES and child education in *IZA Journal of Labor Economics* and the role of non-cognitive traits in study behaviours in *Economics of Education Review*. Email: colm.harmon@sydney.edu.au

James Heckman is the Henry Schultz Distinguished Service Professor of Economics at the University of Chicago, a Nobel Memorial Prize winner in economics and an expert in the economics of human development. He has conducted ground-breaking work with a consortium of economists, developmental psychologists, sociologists, statisticians and neuroscientists showing that quality early childhood development heavily influences health, economic and social outcomes for individuals and society at large. Recent publications include articles on early health shocks and child outcomes in the *Economic Journal*; the Generalized Roy model in *Journal of Political Economy*, and early childhood program and adult outcomes in the *American Economic Review*. Email: jjh@uchicago.edu

Caitriona Logue is a teaching fellow and a part-time lecturer in the UCD School of Economics in University College Dublin. She completed her PhD in 2014 and focus of her Ph.D. was on the econometric methods employed to evaluate the early childhood interventions. Her main research interests are applied economics, micro-econometrics, and the economics of education. Recent publications include articles on early childhood care and education setting in *Childcare in Practice*. Email: caitriona.logue@ucd.ie

Seong Moon is Research Fellow at Economics Research Center in the Department of Economics, University of Chicago. His research interests include empirical labor economics and micro-econometrics, with a particular focus on skill formation, intergenerational transmission and early childhood intervention. His work has been published in journals such as *Journal of Public Economics* and *Quantitative Economics*. Email: moon@uchicago.edu

lifecoursecentre

(ARC Centre of Excellence for Children and Families over the Life Course)
Institute for Social Science Research, The University of Queensland
St Lucia, Qld 4072, Telephone: +61 7 334 **67477**
Email: lcc@uq.edu.au, Web: www.lifecoursecentre.org.au

Abstract

We present 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, non-cognitive 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. This suggests that deficits in parenting skills can be offset within a relatively short timeframe, yet continued investment may be required to observe child effects.

Trial Registration: AEA RCT Registry: AEARCTR-0000066

https://www.socialscienceregistry.org/trials/66

Keywords: skill formation; child development; parental investment; randomized control trial; Ireland

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. The Northside Partnership (through the Irish Government Department of Children and Youth Affairs and The Atlantic Philanthropies) fund the evaluation of the Preparing for Life program. We would like to thank all those who supported this research including the PFL intervention staff and the UCD Geary Institute evaluation team. 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, Alberto Hurtado University and Pontifical Catholic University of Chile 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 Labor Economics and the 6th Irish Conference on Economics and Psychology. The usual disclaimer applies. The UCD Human Research Ethics Committee, the Rotunda Hospital Ethics Committee and the National Maternity Hospital Ethics Committee granted ethical approval for this study.

I. Introduction

Investment in early childhood is increasingly recognized as a key policy mechanism for ameliorating social disadvantage. Evidence from the few experimentally designed programs, implemented in childhood but with long term follow-up, suggests positive effects into adulthood including fewer behavioral problems and criminal convictions, lower dependency on welfare, increased employment, and improved health (Olds et al. 1998; Heckman et al. 2010a; Campbell et al. 2014). Cunha and Heckman (2007) present a model of skill formation demonstrating that early skills facilitate the accumulation of more advanced skills, and these higher level skills make further investment throughout the lifecycle more productive through a process of dynamic complementarity. These processes form the theoretical basis explaining why early investment generates high returns in adulthood, yet little is known about the mechanisms involved in producing these long-term effects.

In this paper we present empirical evidence on the nature of skill formation and parental investment in the early years based on an experimentally designed, home visiting program in Ireland targeting disadvantaged families, known as Preparing for Life (PFL). The program begins in utero and continues until age 5 and thus has the potential to influence skill formation during a period in which brain development is 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 aspects of child development and parenting where the effects from a targeted intervention program manifest early in the lifecycle.

An innovation of this study is the application of statistical methods which are specifically tailored for the analysis of multiple outcomes at multiple waves when using small samples. Heckman et al. (2010a) discuss the difficulty of hypothesis testing in small samples where traditional tests, which rely on the central limit theorem, may not be valid. To avoid the potential for drawing incorrect inferences from a small sample, we present results from both classic *t* tests and nonparametric permutation tests. Permutation tests do not incur any distributional assumptions and therefore produce valid *p*-values when distributions are skewed (Heckman et al. 2010a). Using the methodology of Romano and Wolf (2005) and Heckman et al. (2010a), we apply a stepdown procedure to account for the increased likelihood of false discoveries when examining multiple outcomes. For comparison, we also

present results using Bonferroni adjustment, which a popular multi-hypothesis testing method. The stepdown procedure differs from Bonferroni adjustment as it relaxes the independence assumption, allowing interdependence between measures within outcome families.

By estimating each treatment effect individually, we find a significant program effect (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 the 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 outcomes relating to the quality of the home environment and the level of care mothers provide for their children. As a robustness check, we re-estimate the results by applying inverse probability weights (IPW) to account for differential attrition and non-response. We find that the weighted analysis results in fewer significant treatment effects, yet the overall pattern of results is similar.

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 randomisation procedure and the estimation sample. The econometric framework and outcomes assessed are described in Section III. The results are provided in Section IV, and Section V concludes.

II. Home Visiting Programs and Early Outcomes

Family-focused approaches to early intervention have become increasingly popular due to a belief that parental behaviors serve a mediating role in child development (Brooks-Gunn, Berlin, and Fuligni 2000). Table 1 summarizes evidence from a range of home visiting

programs which examine child development and parenting outcomes.[1] All of the programs focus on similar mechanisms that promote favorable child outcomes such as educating parents about child development 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 estimated in just one case (Landsverk et al. 2002). A greater number of studies focus on 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 outcomes 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.

[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 to conduct a thorough and transparent review of the home visiting research literature and provide an assessment of the evidence of effectiveness for home visiting program models that target families with pregnant women and children from birth to age five. Trained reviewers evaluated randomized controlled trials and quasi-experimental designs for each model and authors were given the opportunity to respond to missing information. 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 i.e. those that met the following criteria: random assignment studies with low attrition of sample members and no reassignment of sample members after the original random assignment; and single case and regression discontinuity designs that meet the What Works Clearinghouse (WWC) design standards (Paulsell et al. 2010). Throughout this paper, when discussing the evidence of home visiting effectiveness, we apply this restriction.

Table 1: Summary of Highly Rated Home Visiting Studies Examining Outcomes before and up to 18 months of age

	Author	Sample Size	Program Evaluated	Outcomes Examined	Significant Favorable Effect Found? *	Infant Age
Child Development	Landsverk et al. (2002)	422	Healthy Families America	Cognitive Development	Yes	12 months
				Physical Development	No	12 months
	Drotar et al. (2009)	364	Parents as Teachers	Noncognitive Development	No	18 months
Parenting	Duggan et al. (1999)	564	Healthy Families America	Environment	No	12 months
				Parental Interactions	No	12 months
				Parental Self-efficacy	No	12 months
	Landsverk et al. (2002)	435	Healthy Families America	Environment	No	12 months
				Appropriate Care	No	12 months
				Parental Interactions	No	12 months
	Duggan et al. (2004)	558	Healthy Families America	Appropriate Care	No	12 months
				Parental Interactions	No	12 months
	LeCroy and Krysik (2011)	180	Healthy Families America	Environment	Yes	6 months
				Environment	No	12 months
				Appropriate Care	Yes	12 months
				Parental Interactions	No	6 months
				Parental Interactions	No	12 months
				Parental Beliefs	No	6 months
				Parental Beliefs	No	12 months
	Minkovitz et al. (2001)	1987	Healthy Steps	Environment	No	2-4 months
				Appropriate Care	Yes	2-4 months
				Parental Interactions	No	2-4 months
	Johnston et al. (2004)	396	Healthy Steps	Environment	No	3 months
				Parental Interactions	No	3 months
				Parental Self-efficacy	No	3 months
				Parental Beliefs	No	3 months
	Koniak-Griffin et al. (2000)	97	Early Intervention Program for Adolescent Mothers	Parental Interactions	No	6 weeks

Note: The information in this table was sourced from the Home Visiting Evidence of Effectiveness website (HomVEE; U.S. Department of Health and Human Services, 2009). Studies which were rated 'High' according to the HomVEE criteria (Paulsell et al., 2010), and studies which have examined outcomes before and up to 18 months are included in this summary. * The cutoff for statistical significant in the HomVEE review is 0.05.

III. Experimental Design

A. Treatment

PFL is a five-year program that was developed to address socioeconomic problems 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.

The high treatment consists of twice monthly home visits delivered by mentors from various professional backgrounds who are trained to support and educate parents about child development using role modelling, demonstration, coaching, discussion, encouragement, and feedback.[3] Each family is assigned the same mentor over the course of the intervention where possible. Visits are tailored based on the age of the child and the needs of the family and are guided by a set of Tip Sheets which present best-practice information on pregnancy, parenting, and child health and development.[4] 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).

While the home visiting program is the intervention under investigation, both the high and low treatment groups receive some common supports including developmental materials and

[2] Doyle and McNamara (2011) find that children from the catchment area were rated by teachers to be below the applied norm (Canadian) at school entry across all five domains on the Short Early Development Instrument including children's *physical health and wellbeing*, *social competence*, *emotional maturity*, *language and cognitive development*, *communication and general knowledge*.

[3] 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. Our analysis is on an intention-to-treat basis as the actual dosage received by each participant may be less than prescribed. An examination of implementation records indicates that participants in the high treatment group received an average of 27 home visits between program entry and 18 months.

[4] The Tip Sheets were designed at a reading level of a 12 year-old to make them as accessible as possible. It is required that all participants must have received the full set of Tip Sheets by the end of the program. An example of a Tip Sheet is presented in Web-Appendix A.

book packs. Both groups are also encouraged to attend public health workshops on stress management and healthy eating which are already taking place in the community. The low treatment group also has access to a support worker if needed (to provide, for example, details about public "services as usual" in the area, such as housing services and childcare services), while this function is provided by the mentors for the high treatment group.

By comparing the high and low treatment groups, it is possible to extract the impact of the home visiting component of the program. A full description of the PFL curriculum is available in Web-Appendix A, 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 hospital records and self-referral in the community. A total of 233 pregnant women consented to participate.[5] A computerized unconditional probability randomization procedure assigned 115 participants to the high treatment group and 118 to the low treatment group.[6] 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

[5] 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. Among those who joined the program, fifty-five percent were recruited via community self-referral and the remaining 45 percent were approached during their first appointment at the maternity hospital. 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. Socio-demographic data for these eligible non-participants are not currently available, however, data collection on this group is on-going.

[6] *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., 2010a).

treatment began.[7] Seventy-six baseline variables were analyzed using permutation testing (the method is described in detail in Section III.B) and 93 percent of the measures show no significant differences between the high and low treatment groups using the 10% cut-off for significance. This indicates that the randomization process was generally successful. Full descriptive tables, including all the measures in the baseline analysis, are available in Web-Appendix B.

The study collects data at eight points during program implementation: baseline, six, 12, 18, 24, 36, 48 months, and school entry (age 4/5). Trained interviewers, who are blinded to the treatment condition, collect data through face-to-face interviews conducted primarily in the participant's home using computer-assisted personal interviewing. This paper uses data from the baseline, six, 12, and 18 month assessments.

C. Participant Profile

Table 2 provides baseline descriptive statistics for the estimation sample available at each wave.[8] The participating mothers were 26 years old on average, and 21 weeks pregnant when they joined the program. Approximately 40 percent were employed, over 80 percent had a partner, and almost half were first time mothers. 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.[9] Using a more refined measure of cognitive capacities, the average level of maternal IQ was approximately 82 using the

[7] 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.

[8] 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.

[9] 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.

Wechsler Abbreviated Scale of Intelligence (Wechsler 1999) which is below the lower bound on the expected population average range of between 85 and 115.

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 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 such as smoking, yet there are some similarities such as presence of husband/partner.[10] A detailed comparison of the GUI and PFL samples is presented in Table 3.

[10] 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

	High Treatment Mean (SD)			**Low Treatment Mean (SD)**		
	6 Month Sample	***12 Month Sample***	***18 Month Sample***	***6 Month Sample***	***12 Month Sample***	***18 Month Sample***
Weeks pregnant at program entry	21.78 (7.83)	21.84 (7.88)	21.93 (7.93)	21.18 (6.87)	21.17 (7.02)	21.32 (6.62)
Age	25.67 (5.76)	25.87 (6.01)	25.93 (5.91)	25.69 (6.04)	25.13 (6.02)	25.56 (6.10)
Married	0.16 (0.37)	0.16 (0.37)	0.16 (0.37)	0.17 (0.38)	0.16 (0.37)	0.15 (0.36)
Has partner (including married)	0.80 (0.40)	0.82 (0.39)	0.79 (0.41)	0.83 (0.38)	0.83 (0.38)	0.82 (0.39)
Living with parent(s)	0.55 (0.50)	0.54 (0.50)	0.54 (0.50)	0.45 (0.50)	0.48 (0.50)	0.47 (0.50)
First time mother	0.52 (0.50)	0.51 (0.50)	0.53 (0.50)	0.46 (0.50)	0.49 (0.50)	0.47 (0.50)
Low education	0.29 (0.46)	0.30 (0.46)	0.30 (0.46)	0.36 (0.48)	0.30 (0.46)	0.34 (0.48)
Employed	0.43 (0.50)	0.43 (0.50)	0.43 (0.50)	0.40 (0.49)	0.43 (0.50)	0.41 (0.50)
IQ[a]	82.52 (12.94)	83.11 (12.60)	83.32 (12.35)	80.60 (13.14)	81.54 (12.75)	82.04 (12.16)
Saves regularly	0.50 (0.50)	0.50 (0.50)	0.49 (0.50)	0.53 (0.50)	0.55 (0.50)	0.53 (0.50)
Public housing	0.54 (0.50	0.54 (0.50)	0.55 (0.50)	0.56 (0.50)	0.55 (0.50)	0.53 (0.50)
Prior physical health condition	0.76 (0.43)	0.76 (0.43)	0.75 (0.44)	0.64 (0.48)	0.65 (0.48)	0.63 (0.49)
Prior mental health condition	0.27 (0.45)	0.28 (0.45)	0.26 (0.44)	0.26 (0.44)	0.26 (0.44)	0.26 (0.44)
Smoked during pregnancy	0.51 (0.50)	0.51 (0.50)	0.51 (0.50)	0.49 (0.50)	0.46 (0.50)	0.47 (0.50)
Alcohol during pregnancy	0.27 (0.45)	0.29 (0.46)	0.29 (0.46)	0.25 (043)	0.26 (0.44)	0.27 (0.45)
Drugs during pregnancy	0.01 (0.11)	0.01 (0.11)	0.01 (0.11)	0.03 (0.18)	0.01 (0.11)	0.01 (0.12)
N	82	82	80	89	82	73

Note: [a]The Weschler Abbreviated Scale of Intelligence (WASI) was used to measure maternal IQ at 3 months postpartum rather than baseline.

Table 3: Comparison of Mothers in the Growing up in Ireland Study and Preparing for Life

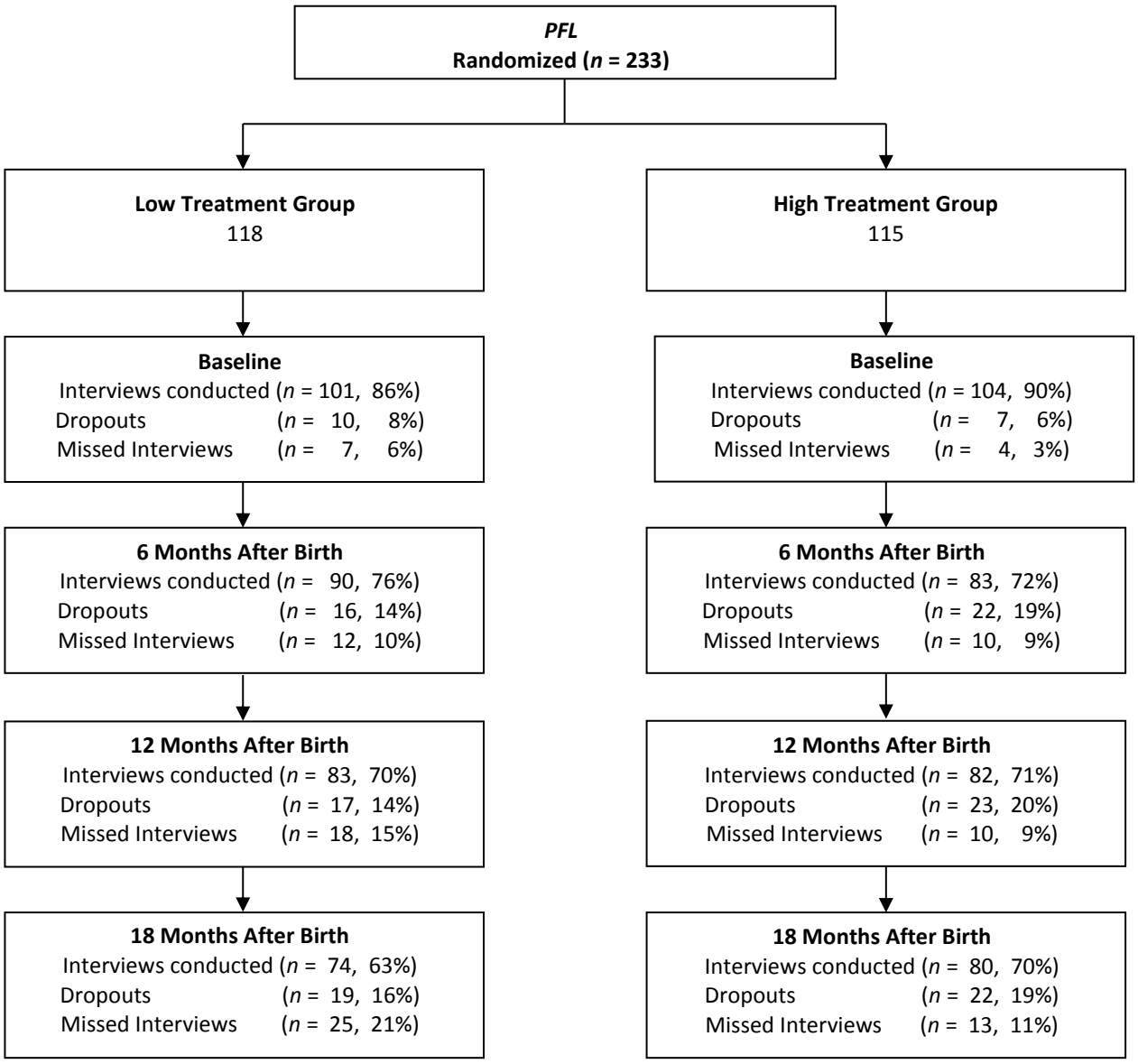
	Growing up in Ireland **Mean (SD)**	***Preparing for Life*** **Mean (SD)**
	9 Month Sample	***Baseline Sample***
Age[a]	31.61 (5.35)	25.38 (5.90)
Married	0.68 (0.47)	0.16 (0.38)
Has partner (including married)	0.88 (0.33)	0.81 (0.39)
Living with parent(s)	0.05 (0.23)	0.52 (0.50)
Low education	0.12 (0.32)	0.37 (0.48)
Employed	0.57 (0.49)	0.38 (0.49)
Social housing	0.08 (0.27)	0.55 (0.50)
Physical or mental health condition	0.11 (0.32)	0.74 (0.44)
Smoked during pregnancy[b]	0.18 --	0.49 (0.50)
Alcohol during pregnancy[b]	0.20 --	0.28 (0.45)
N	11,134	205

Note: [a] Mothers' age in GUI was recorded when the study child was 9 months of age, while in PFL it was recorded during pregnancy (at 21 weeks on average). [b] Data on smoking and alcohol consumption during pregnancy for GUI participants were not available for analysis; thus the means were sourced from Williams et al. (2010).

D. *Attrition and Non-response.*

Figure 1 describes the progression of the 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).

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



On average, 19 percent of the high treatment group (22/115) and 16 percent of the low treatment group (19/118) were classified as official 'dropouts' between baseline and 18 months, with the majority of dropout occurring before 6 months. Dropouts are defined as those who actively told the *PFL* program staff or the research team that they wanted to leave the study. Participants who 'missed interviews' are defined as those who have not officially dropped out of the program, but are difficult to engage at the assessment point. The proportion of missed interviews across the high and low treatment groups are 11 percent (13/115) and 21 percent (25/118) respectively between baseline and 18 months. As shown in Table 2, the estimation samples differ at each data collection point, however the groups remain balanced on baseline characteristics across each of the three waves. Specifically, no statistically significant baseline differences emerge for the 6 or 12 month estimation samples, while just one significant difference is found at 18 months (a greater proportion of high treatment participants had a prior physical health condition). We examine attrition and non-response in Section III.D using an inverse probability weighting technique to account for any potential bias due to differential 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. This hypothesis is equivalent to the statement that the counterfactual outcome vectors share the same distribution H-1: $Y(1) \overset{d}{=} Y(0)$ where $\overset{d}{=}$ denotes equality in distribution.

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). Web Appendix C 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 Behaviour 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 with the child based on the *Community Support Inventory*, 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. Web-Appendix D discusses the reliability of these instruments.

B. Permutation Testing

Although the RCT design in (1) is a simple specification, the use of traditional *t* tests for hypothesis testing is not viable given the small sample size and the likely non-normality of the data. Permutation methods do not depend on distributional assumptions and thus facilitate the estimation of treatment effects in small samples. While our analysis replicates a few recent studies of an early childhood intervention using this approach (Heckman et al. 2010a; 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.[11] Upton

[11] 100,000 replications are permuted using Monte Carlo resampling in our analyses.

(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:

$$MP(t) = P(t^* > t) + 0.5P(t^* = t) \quad (2)$$

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 positive effect on child and parenting outcomes compared to the low level treatment. We adopt a 10% *p*-value to assess statistical significance.

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, Shaikh, and Wolf 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 is accepted, 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 the outcomes. The Romano and Wolf (2005) method uses a weaker assumption than other established stepwise methods (Benjamini and Hochberg 1995; Westfall and Wolfinger 1997) – monotonicity with respect to the critical values. This ensures that the largest unadjusted *p*-value corresponds to the largest adjusted *p*-value (Heckman et al. 2010a).

The 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, however outcomes which are 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). Table 4 shows the stepdown families and the individual measures included in each, and Tables 5-8 in the following 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.

Table 4: Stepdown Families

Stepdown Families	6 Months Measures	12 Month Measures	18 Month Measures
CHILD DEVELOPMENT			
Cognitive Development	Communication (0-60; ↑)	Communication (0-60; ↑)	Communication (0-60; ↑)
	Problem Solving (0-60; ↑)	Problem Solving (0-60; ↑)	Problem Solving (0-60; ↑)
		First Communicative Gestures (0-12; ↑)	First Communicative Gestures (0-12; ↑)
		First Signs of Understanding (0-3; ↑)	First Communicative Gestures (0-3; ↑)
		Words Understood (5-99; ↑)	Words Understood (5-99; ↑)
		Words Produced (5-99; ↑)	Words Produced (5-99; ↑)
		Cognitive Development (70-140; ↑)	Cognitive Development (70-140; ↑)
Noncognitive Development	Social-Emotional Score (0-285; ↓)	Social-Emotional Score (0-285;↓)	Social-Emotional Score (0-285;↓)
	Personal Social Score (0-60; ↑)	Personal Social Score (0-60; ↑)	Personal Social Score (0-60; ↑)
	Difficult Temperament (0-42;↓)	Difficult Temperament (0-42;↓)	Competence (0-44; ↑)
		Competence (0-44; ↑)	Problem Score (0-22;↓)
		Problem Score (0-22;↓)	
		Atypical Behavior (0-15;↓)	
Physical Development	Gross Motor (0-60; ↑)	Gross Motor (0-60; ↑)	Gross Motor (0-60; ↑)
	Fine Motor (0-60; ↑)	Fine Motor (0-60; ↑)	Fine Motor (0-60; ↑)
PARENTING			
Environment	Learning Materials (0-9; ↑)		Learning Materials (0-9; ↑)
	Organization (0-6; ↑)		Organization (0-6; ↑)
	Physical Environment (0-9; ↑)		Physical Environment (0-9; ↑)
	Activities in Child Environment (0-5; ↑)		Activities in Child Environment (0-5; ↑)
	Safety (0-10; ↑)		Safety (0-10; ↑)
Appropriate Care	Acceptance (0-8; ↑)		Acceptance (0-8; ↑)
	Variety (0-5; ↑)		Variety (0-5; ↑)
	Suitable Care Provided (0-11; ↑)		Suitable Care Provided (0-11; ↑)
	Hostile-Reactive Behavior (0-11;↓)		
Parental Interactions	Responsivity (0-11; ↑)	Mother Reads to Her Child (0-1; ↑)	Responsivity (0-11; ↑)

Stepdown Families	6 Months Measures	12 Month Measures	18 Month Measures
	Involvement (0-6; ↑) Activities to Stimulate Development (0-5; ↑)	Mother Reads to Her Child Every Day (0-1; ↑)	Involvement (0-6; ↑) Activities to Stimulate Development(0-5; ↑) Mother Reads to Her Child (0-1; ↑) Mother Reads to Her Child Every Day (0-1; ↑)
Parental Attachment	Quality of Attachment (1-5; ↑) Pleasure in Interaction (1-5; ↑) Absence of Hostility (1-5; ↑) Parental Warmth (1-11; ↑) Parental Overprotection (1-11;↓) Baby Comparison (1-11; ↑) Dysfunctional Interactions (12-60;↓) Difficult Child (12-60;↓)		Maternal Separation Anxiety Scale (7-35;↓)
Parental Self-efficacy	Control of Child's Behavior (4-20;↓) Child Control of Parent's Life (4-20;↓) Parental Efficacy (PLOC) (4-20;↓) PSI Parenting Distress (12-60;↓) Parental Self-efficacy (PACOTIS) (1-11; ↑)		Parenting Daily Hassles (0-100;↓)
Parental Beliefs	Parental Responsibility (4-20;↓) Parental Belief in Fate (4-20;↓) Perceived Parental Impact (1-11; ↑)	Belief in Appropriate Punishment (1-10; ↑) Realistic Expectations of Children (1-10; ↑) Promoting Children's Independence (1-10; ↑) Appropriate Parent-Child Roles (1-10; ↑) Parental Empathy (1-10; ↑) Knowledge of Child Development (1-100; ↑)	

Note: The numbers in parentheses represent the minimum and maximum value of each measure and the arrow represents the direction of the effect. Thus, ↑ indicates that higher values on the scale represent favorable outcomes, and ↓ indicates that higher values on the scale represent unfavorable outcomes.

Utilizing each of the subdomains of the child development instruments, we derived 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 family captures the children's communication and vocabulary skills, as well as their problem solving abilities, and general cognitive development. The noncognitive development family captures the children's socio-emotional skills, temperament, and behavior. The physical development family captures the children's gross and fine motor skills.

Similarly, utilizing each of the subdomains of the parenting instruments, we derived 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, Elo and Culhane (2013), Currie (2001), Heckman and Kautz (2013), and Hertwig, Davis and Sulloway (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.[12] To account for any bias that attrition and item non-response may introduce, we test the robustness of the main analysis using an inverse probability weighting (IPW) technique. Adapted from the description in Campbell et al. (2014), we make the assumption that the outcome is independent of the missing data pattern, conditional on treatment assignment and observable baselines characteristics. This can be written as

[12] 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.

(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. The probability that an observation has a non-missing outcome can be described as follows

(4) $$P_i = D_i\,(\Pr(M = 1 \mid X_i, D = 1)) + (1 - D_i)(\Pr(M = 1 \mid X_i, D = 0))$$

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) $$ATE = \sum_{i=1}^{N} Y_i \cdot 1(D_i = 1, M_i = 1) \cdot {}^{w_i}/_{N_1} - \sum_{i=1}^{N} Y_i \cdot 1(D_i = 0, M_i = 1) \cdot {}^{w_i}/_{N_2}$$

Where

(7) $$N_d = \sum_{i=1}^{N} 1\,(D_i = d, M_i = 1) \cdot w_i \;\; d \in \{0, 1\}$$

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 calculate 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 E 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 5 and 6, 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 5, 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*, N*oncognitive Development*, and *Physical Development*. The results are presented in Table 5. 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.

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 fails to be rejected. Three individual significant differences in the non-hypothesised direction are found for words understood at 12 and 18 months and first communicative gestures score at 12 months, however the stepdown adjusted *p*-values (not displayed in Table 5) associated with these mean differences are not significant at 12 or 18 months.[13]

[13] *p*-values in the non-hypothesised direction are not reported in the tables, however any significant results are noted in the text.

- *Table 5: 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	8.90	0.464	0.466	-	0.622[5]	9.44	9.14	0.607	0.606	-	0.606[5]

								(5.74)	(6.49)					(6.63)	(7.18)				
	[(-)]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. Statistical significance at the 1, 5, and 10 percent levels is indicated by ***, **, *, respectively.

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 6. 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). Individual hypothesis testing also indicates a treatment group 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. 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 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 favourably 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.

Table 6: Treatment Effects for Parental Investment 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*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$
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	-	-	-	-	-	-	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]

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*)	$P^{(i)}$	$P^{(ii)}$	$P^{(ii i)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$
	Every Day																		
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]	-	-	-	-	-	-	-	-	-	-	-	-
	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

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*)	$P^{(i)}$	$P^{(ii)}$	$P^{(ii\ i)}$	$P^{(iv)}$	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$
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 Development	-	-	-	-	-	-	70.30 (7.81)	69.69 (6.81)	0.297	0.299	-	0.700[2]	-	-	-	-	-	-
	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. Statistical significance at the 1, 5, and 10 percent levels is indicated by ***, **, *, respectively.

B. Robustness Tests

Tables 7 and 8 present the IPW-adjusted weighted results and can be read in the same manner as Tables 5 and 6.

Child Development - Table 7 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 7 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-hypothesised direction, the stepdown adjusted *p*-value associated with words understood at 18 months is also statistically significant. As in the non IPW-results, when the IPW method is applied to the *Noncognitive Development* family, the result of hypothesis testing indicates that 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 *Noncognitive Development* family are echoed. Although a more favourable treatment effect emerges for the *Cognitive Development* family as a result of a new, precisely identified impact on the cognitive development score at 18 months. However, this must be balanced with a less favourable 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.

Table 7: 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. Statistical significance at the 1, 5, and 10 percent levels is indicated by ***, **, *, respectively.

Parenting - Table 8 shows that when the IPW method is applied to parenting outcomes, treatment effects are identified in the same stepdown families as the non IPW-results, but fewer are statistically significant. 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, at 18 months the joint null is rejected using the Bonferroni procedure. 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* 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 IWP-adjusted permutation testing and the stepdown procedure is applied.

Table 8: Treatment Effects for Parental Investment 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*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$
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)	0670	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]

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*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$
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]	-	-	-	-	-	-	-	-	-	-	-	-
	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	[(-)]Parental	12.28	12.81	0.137	0.152	0.411	0.378[1]	-	-	-	-	-	-	-	-	-	-	-	-

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*)	$P^{(i)}$	$P^{(ii)}$	$P^{(iii)}$	$P^{(iv)}$
Beliefs	Responsibility	(3.31)	(3.03)																
	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]	-	-	-	-	-	-
	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. Statistical significance at the 1, 5, and 10 percent levels is indicated by ***, **, *, respectively.

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. Overall, we find 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, and few effects on early skill formation.

With respect to child development, there is little evidence of a statistically significant program effect, a result which is consistent with previous evaluations of US home visiting programs which have examined early child outcomes (U.S. Department of Health and Human Services 2009). When we correct for attrition bias, a treatment effect is identified on the cognitive development family at 18 months, which is driven by higher cognitive development scores among the high treatment group. However, multi-hypothesis testing in the non-hypothesised direction indicates that the low treatment group appear to understand more words at 18 months compared to the high treatment group. Thus, the evidence of a treatment impact on child development is inconclusive. The lack of sizable effects on key dimensions of child development may be attributable to dosage and timing of the intervention. The average high treatment participant began engaging with the program half way into her pregnancy and had received, on average, 27 home visits between program entry and 18 months. It is possible that this small window of intervention did not allow enough time for the participants to adopt the strategies advised by their mentors as the bond between mentor and participant was still forming (Ammerman et al. 2010).

With respect to the parental investment measures, all significant treatment effects were in the hypothesised direction. Our analysis suggests that home visiting programs can be an effective means of improving deficits in the level of appropriate care provided to children and the quality of the home environment within a relatively short time frame. This pattern remains when the IPW method is applied, although fewer treatment effects are statistically significant. The literature has examined multiple parenting outcomes and for the majority of measures, no significant treatment effects are identified (see Table 1). Consistent with our results, the areas

where effects have been identified relate to the child's environment and the appropriate level of care.

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 develop and grow is via changes in parenting skills and abilities. However, according to Shonkoff and Phillips (2000) changing parenting behavior is extremely difficult. Hertwig et al. (2002) propose that material resources, cognitive stimulation and parental interpersonal skills may each serve divergent roles in the transmission process to the child. By examining multiple dimensions of parental investment, our analysis sheds light on the aspects of parenting which are most malleable. The measures of the environment and appropriate care where significant treatment effects are found, relate to material resources that can be observed in the household and activities that the mother carries out with her child. The items in these categories are essentially yes/no questions relating to what the mother *does* with her infant. 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 Brooks-Gunn and Markman (2005) who state that parenting interventions may be more effective at changing parental behavior rather than parental emotional states.

These new parenting strategies and skills, which have been developed through interactions with PFL mentors and materials, may take time to have an impact on infant behavior and development. Indeed, the majority of studies that calculate high returns to early childhood investment are based on analyses conducted when the participating children have reached the teenage years or adulthood (Olds et al. 1997; Heckman et al. 2010b). It is important to understand the environmental factors that are most malleable in early childhood as the theory on human skill formation points to a skill multiplier effect (Cunha and Heckman 2007), and thus, a small change early in life may have large effects later in life. This paper suggests that improvements in early parenting skills may be one such mechanism that accounts for these later effects.

As the potential for contamination in PFL is high given the geographical proximity of the participants, a number of strategies were devised to measure cross-talk and information flows between the two treatment groups (information on these strategies can be found in Doyle and Hickey, 2013). Analysis of these data finds that while the conditions for contamination or spillover effects is quite high as participants are regularly in contact with each and share materials, the blue-dye analysis suggests that these practices do not translate into improved

parenting knowledge among the low treatment group suggesting that contamination from the high to low treatment group is minimal (see Doyle and PFL Evaluation Team, 2013).

This paper contributes to the literature in a number of ways. First, we acknowledge the distinction between cognitive, noncognitive and physical development and various aspects of parental investment. Second, we adopt appropriate statistical tests to overcome issues inherent in RCTs including the analysis of small samples and accounting for the increased likelihood of Type I error when examining multiple hypotheses. Studies of home visiting programs typically report many outcomes, yet few account for multiple testing which risks overstating the true impact of the program. Finally, we apply an inverse probability weighting method to account for the potential bias that attrition and non-response may cause.

From a methodological perspective, when individual hypothesis testing methods are applied, this study identifies a significant effect for 17 percent of outcomes (6/35) at 6 months, 4 percent of outcomes (1/23) at 12 months, and 18 percent of outcomes (5/28) at 18 months. The results in the non IPW-analysis are the same regardless of whether classic *t*-tests or permutation tests are applied. This addresses the concern that the small sample size may lead to skewness in the distribution of outcomes and, thus, the concern that incorrect inferences may be drawn if only asymptotically valid tests are applied. However, the results indicate that *t* tests produce quite different *p*-values to the permutation tests when inverse probability weighting (IPW) is applied, indicating that caution is needed when analysing weighted data using traditional methods. When the *p*-values are adjusted to account for the increased likelihood of a Type I error in a multiple hypotheses setting, fewer treatment effects emerge overall. We use both Bonferroni adjustment and a stepdown procedure to conduct multiple hypothesis testing and find relatively similar results. Yet, in contrary to the stepdown method, the Bonferroni method fails to reject the joint null hypothesis associated with the *Environment* category at 6 months in the non IPW-analysis.

To the best of our knowledge, this is the first study examining early childhood outcomes that employs a robust procedure to deal with multiple outcomes, small samples and differential attrition. The HomVee Review (U.S. Department of Health and Human Services 2009) served as our source of information on other high quality home visiting studies. Avellar and Paulsell (2011) note that few of the studies examined as part of this review make corrections for multiple outcomes. In addition, this is the only study of early life outcomes which has approached attrition using an Inverse Probability Weighting technique to ensure that a larger weight is given to participants who are under-represented in the study. This approach has

been employed to examine adult health outcomes in Campbell et al. (2014), yet this rigorous approach has not yet been applied to studies of early life outcomes. The widespread use of the methods applied here would help harmonise the early intervention literature and improve within study internal validity, the absence of which has limited insight from previous research.

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

References

Ammerman, Robert. T., Frank W. Putnam, Nicole R. Bosse, Angelique R. Teeters, and Judith B. Van Ginkel. 2010. "Maternal Depression in Home Visitation: A Systematic Review." *Aggression and Violent Behaviour* 15 (3): 191-200.

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

Benjamini, Yoav, and Yosef Hochberg. 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.

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

Brooks-Gunn, Jeanne, Lisa J. Berlin, and Allison Sidle Fuligni. 2000. "Early Childhood Intervention Programs: What About the Family?" In *Handbook of Early Childhood Intervention* edited by Jack Shonkoff and Steven Meisels, (3rd ed.). New York: Cambridge University Press.

Campbell, Frances, Gabriella Conti, James J. Heckman, Seong Hyeok Moon, Rodrigo Pinto, Elizabeth Pungello, and Yi Pan. 2014. "Early Childhood Investments Substantially Boost Adult Health." *Science* 343 (6178): 1478-1485.

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

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

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

Doyle, Orla. 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.

— , & McNamara, Kelly 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.

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

— , & PFL Evaluation Team. 2013. *Assessing the Impact of Preparing for Life at Twenty-four Months.* Report to Preparing for Life Programme (Atlantic Philanthropies & Department of Children and Youth Affairs). http://geary.ucd.ie/preparingforlife/?page_id=146

Duggan, Anne K., Elizabeth C. McFarlane, Loretta Fuddy, Lori Burrell, Susan M. Higman, Amy M. Windman, and Calvin Sia. 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.

Duggan, Anne K., Elizabeth C. McFarlane, Amy M. Windham, Charles A. Rohde, David S. Salkever, Loretta Fuddy, Leon A. Rosenberg, Sharon B. Buchbinder, and Calvin C.J.

Sia. 1999. “Evaluation of Hawaii’s Healthy Start Program.” *The Future of Children* 9 (1): 66-90.

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

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

Heckman, James J., and Tim Kautz. 2013. “Fostering and Measuring Skills: Interventions that Improve Character and Cognition.” *NBER Working paper,* No. 19656.

—,, Seong Hyeok Moon, Rodrigo Pinto, Peter A Savelyev, and Adam Yavitz. 2010a. “Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the HighScope Perry Preschool Program.” *Quantitative Economics* 1 (2): 1-46.

—, —, —, —, &, —. 2010b. “The Rate of Return to the HighScope Perry Preschool Program.” *Journal of Public Economics* 94 (1-2): 114-128.

Hertwig, Ralph, Jennifer N. Davis, and Frank J. Sulloway. 2002. “Parental Investment: How an Equity Motive can Produce Inequality.” *Psychological Bulletin* 128 (5): 728–745.

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

Johnston, Brian D., Colleen E. Huebner, Lynda T. Tyll, Willaim E. Barlow, and Robert S. Thompson. 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.

Knudsen, Eric I., James J. Heckman, Judy L. Cameron, and Jack P. Shonkoff. 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, Deborah, Nancy L. Anderson, Inese Verzemnieks, and M.L. Brecht. 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, Craig W., and Judy Krysik. 2011. “Randomized Trial of the Healthy Families Arizona Home Visiting Program.” *Children and Youth Services Review* 33 (10): 1761-1766.

Minkovitz, Cynthia, Donna Strobino, Nancy Hughart, Daniel Scharfstein, Bernard Guyer, 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, Charles A. 2000. *“The Neurobiological Bases of Early Intervention”.*In *Handbook of Early Childhood Intervention*, edited by Jack P. Shonkoff & Stephen .J. Meisels, Second Edition (pp. 204-227). Cambridge, MA: Cambridge University Press.

Olds, David L., Charles R. Henderson, Robert Cole, John Eckenrode, Harriet Kitzman, and Dennis Luckey, Lisa Pettitt, Kimberly Sidora, Pamela Morris, and Jane Powers. 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.

— , John Eckenrode, Charles R. Henderson, Harriet, Kitzman, Jane H. Powers, Robert Cole, Kimberly Sidora, Pamela Morris, Lisa M. Pettitt, and Dennis Luckey. 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.

Paulsell, Diane, Sarah Avellar, Emily Sama Martin, and Patricia Del Grosso. 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.

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

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

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

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

Upton, Graham 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.

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

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

Williams, James, Sheila Greene, Sinead McNally, Aisling Murray, and Amanda Quail. 2010. *Growing Up in Ireland: The Infants and Their Families: Report 1*. Dublin: Government Publications.