UCD GEARY INSTITUTE FOR PUBLIC POLICY
DISCUSSION PAPER SERIES

The First 2,000 Days and Child Skills: Evidence from a Randomized Experiment of Home Visiting

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

Geary WP2017/06
July 11, 2017

The First 2,000 Days and Child Skills: Evidence from a Randomized Experiment of Home Visiting

Orla Doyle[1]

[1] The evaluation of the Preparing for Life program was funded by the Northside Partnership through the Department of Children and Youth Affairs and The Atlantic Philanthropies. I would like to thank all those who supported this research for the last ten years, especially the participating families and community organizations, the PFL intervention staff, and the program's Expert Advisory Committee. Thanks also to the Scientific Advisory Committee, including Sylvana Côté, Colm Harmon, James Heckman, Cecily Kelleher, Sharon Ramey, Craig Ramey, and Richard Tremblay, for their guidance and advice throughout the project, and to the Early Childhood Research Team at the UCD Geary Institute for Public Policy who contributed to this project. This paper was written while visiting the University of Chicago, the University of Sydney, the Paris School of Economics, and the University of Bordeaux; I would like to thank all those who helped shaped this paper. The trial was registered with controlled-trials.com (ISRCTN04631728) and the AEA RCT Registry (AEARCTR-0000066). All study procedures were approved by the UCD Human Research Ethics Committee, the Rotunda Hospital's ethics committee, and the National Maternity Hospital's ethics committee. E-mail: orla.doyle@ucd.ie

The First 2,000 Days and Child Skills: Evidence from a Randomized Experiment of Home Visiting

Abstract

Using a randomized experiment, this study investigates the impact of sustained investment in parenting, from pregnancy until age five, in the context of extensive welfare provision. Providing the Preparing for Life program, incorporating home visiting, group parenting, and baby massage, to disadvantaged Irish families raises children's cognitive and socio-emotional/behavioral scores by two-thirds and one-quarter of a standard deviation respectively by school entry. There are few differential effects by gender and stronger gains for firstborns. The results also suggest that socioeconomic gaps in children's skills are narrowed. Analyses account for small sample size, differential attrition, multiple testing, contamination, and performance bias.

Keywords: Early childhood intervention; cognitive skills; socio-emotional and behavioral skills; randomized control trial; multiple hypothesis testing; permutation testing; inverse probability weighting.

JEL Classification: C93, D13, I26, J13

There is a growing evidence base demonstrating that circumstances early in life are critical for the development of the skills and abilities required to lead a successful life. Children exposed to adverse prenatal and postnatal environments typically experience poorer health, education, and labor market outcomes in the long run (Cunha *et al*. 2006; Heckman 2006; Almond and Currie 2011). Intervening early in life to eradicate or compensate for these deficits through early childhood intervention (ECI) programs is becoming an increasingly accepted strategy (see Council of Economic Advisors 2014; OECD 2016). Such investments are considered efficient from both a biological and economic perspective (Doyle *et al*. 2009). Physiologically, there is evidence of greater brain plasticity and neurogenesis in the early years, particularly between pregnancy and age 3 (Thompson and Nelson 2001; Knudsen *et al*. 2006), therefore increased investment during this period of malleability is likely to have a sustained impact on children's skills (Halfon, Shulman, and Hochstein 2001). Such investments are also economically efficient, as by investing early the returns from the improved skill set can be reaped over a longer period (Karoly, Kilburn, and Cannon 2005; Heckman and Kautz 2014). Thus, the 'first 1,000 days' has been predicated as a key period for policy investments (The Lancet 2016).

This paper examines the impact of a prenatally commencing ECI program which targets disadvantaged communities and focuses on parents as the key mechanism of change. By conducting a 5 year intervention, i.e. the first 2,000 days,[2] the impact of early *and* sustained investment during a critical stage of development can be established. This is important as the technology of skill formation, proposed by Cunha and Heckman (2007), establishes that children's early skills facilitate the development of more advanced skills through a process of self-productivity, and this in turn makes investment throughout the lifecycle more productive through a process of dynamic complementarity (Cunha, Heckman, and Schennach 2010; Heckman and Mosse 2014). While there is a genetic basis for the development of skills (Nisbett *et al*. 2012), they can be modified and enhanced by environmental conditions (Weaver *et al*. 2004). The traditional human capital production function shows that skills are determined by inputs of time and market goods/income (Becker 1965; Michael and Becker 1973), and that inequalities in skills arise from differences in the availability of these resources. This contributes to the large and well-documented socioeconomic gap in children's cognitive and non-cognitive skills that can be

[2] Participants joined the ECI program during their 21st week of pregnancy, on average, and left when their children started their first year of primary school when they were 4 years, 9 months old, on average, thus ~1,855 days is the precise figure.

observed as early as 18 months of age (Cunha and Heckman 2007; Fernald, Marchman, and Weisleder 2013). While such deficits have been partly explained by poverty, credit constraints (e.g. Carnerio and Heckman 2003), and parental time investments (e.g. Bernal and Keane 2001; Del Boca, Flinn, and Wiswall 2014; Del Bono *et al*. 2016), these factors may also influence and/or serve as proxies for the child's environment. Indeed, empirical research has identified the quality of the home environment (Bakermans-Kranenburg, van IJzendoorn, and Bradley 2005; Todd and Wolpin 2007), parenting skills (Dooley and Stewart 2007; Fiorini and Keane 2014), and parental stimulation (Miller *et al*. 2014) as important predictors of children's ability. As a result, many production function models have been amended to include parenting skills, behaviors, and beliefs, and several economic models of parenting have emerged (e.g. Burton, Phipps, and Curtise 2002; Doepke and Zilibotti 2014; Cunha 2015; Cobb-Clarke, Salamanca, and Zhu 2016). While these models differ in their focus, they all recognize the important role of parenting in the production of children's skills and the inequalities that can result as a consequence.

Families from disadvantaged backgrounds often face financial constraints which limits their ability to sufficiently invest in their children, however they may also be constrained in their capacity to parent. Evidence suggest that parents from low socioeconomic status (SES) backgrounds engage in poorer parenting styles and behaviors (Lareau 2011; Cunha, Elo, and Culhane 2013). For example, lower SES parents tend to engage in more negative parenting styles such as permissive or harsh parenting (Bradley and Corwyn 2002), while providing less stimulating materials and experiences such as going to a library or providing learning materials and books (Bradley *et al*. 1989). This partly may be attributed to a knowledge gap concerning both appropriate parenting practices and techniques for optimizing child development. Specifically, Cunha *et al.* (2013) identify a lack of parenting knowledge and differing beliefs about the importance of parenting among low SES parents. There is also evidence of less pre-academic stimulation, such as reading to children and helping them to recognize letters, in disadvantaged homes (Miller *et al.* 2014). Thus, promoting ECI strategies which increase parenting knowledge and encourage parental stimulation in developmental appropriate activities may counteract the adverse effects of poverty on children's skills.

Much of the policy focus on ECI has been attributed to the long-run findings from preschool programs which target children directly (e.g. Head Start). Interventions which target parents and/or start in pregnancy have a smaller evidence base concerning their

long-term effectiveness. Parent-focused interventions are delivered in a home or group based setting, and home visiting programs in particular have become increasingly popular, especially in the US where the Federal Maternal, Infant, and Early Childhood Home Visiting program has invested over $1.85 billion in home visiting (Maternal and Child Health Bureau 2016). Yet evidence on the effectiveness of these programs on children's early development is mixed, and effects are typically modest in size and not consistent across programs (Sweet and Applebaum 2004; Gomby 2005; Filene *et al.* 2013; Peacock *et al.* 2013; Avellar *et al.* 2016).[3] The best known prenatally commencing home visiting program, that has followed participants into early adulthood, is the Nurse Family Partnership (NFP) program (Eckenrode *et al.* 2010). They find that girls in the treatment group are less involved in crime, have fewer children, and are less likely to receive Medicaid at age 19, however there are no effects for boys, or for any educational outcome.

The evidence base for the effectiveness of ECI programs, and home visiting programs in particular, is mainly based on studies from the US, and more recently from developing countries.[4] One may expect lower SES inequalities in Europe where many countries are characterized by universal health insurance, generous welfare payments, and a social safety net which protects the most vulnerable in society. Yet inequalities in children's skills are a universal phenomenon, and continue to persist in Europe despite these arguably more redistributive policies (Martins and Veiga 2010; Lecerf 2016).[5] The existence of such inequalities suggests that family economic circumstances alone may not be the primary driver of these differences in skills. Thus, the expanded human capital

[3] A small number of home visiting studies identify favorable effects on early cognitive development, including Early Head Start (EHS) at 36 months (Roggman, Boyce, and Cook 2009), Parents as Teachers (PAT) at ages 4 to 5 (Drazen and Haust 1993), and the Nurse Family Partnership (NFP) program at age 6 (Olds *et al.* 2004). However, other studies of NFP and EHS find no significant treatment effects for cognition between the ages 2 and 5 (Olds, Henderson, and Kitzman 1994; Jones Harden *et al.* 2012). There is also evidence that home visiting programs can impact language development, as found in Home Instruction for Parents of Preschool Youngsters (HIPPY) at ages 3 to 5 (Necoechea 2007), NFP at age 6 (Olds *et al.* 2004), and PAT at ages 4 to 5 (Drazen and Haust, 1993). Yet many of these effects are absent when measured at school entry, including the Mother-Child Home Program, HIPPY, and EHS (Madden, O'Hara, and Levenstein 1984; Baker and Piotrkowski 1996; Jones Harden *et al.* 2012). A number of programs have also identified positive treatment effects on children's social and emotional skills between age 3 and school entry including fewer internalizing, externalizing, and social problems (e.g. Olds *et al.* 1994 (NFP); Landsverk *et al.* 2002 (Healthy Families America); Olds *et al.* 2004 (NFP); Fergusson *et al.* 2005 (Early Start); Connell *et al.* 2008 (Family Check-Up); Shaw *et al.* 2009 (Family Check-Up); Jones Harden *et al.* 2012 (EHS)).

[4] There is evidence that home visiting programs delivered in developing countries have led to short (e.g. the Columbian Conditional Cash Transfer Program, see Attanasio *et al.* 2015), and long (e.g. the Jamaica home visiting program, see Walker *et al.* 2011) term impacts on children's skills.

[5] Martins and Veiga (2010) find that socioeconomic status represents between 14.9 percent and 34.6 percent of the overall inequality in mathematics scores in the EU using PISA (Programme for International Student Assessment) data, with Germany scoring the highest and Sweden the lowest. In Ireland, the figure is 25 percent.

production function, which moves beyond income and time investments as the main determinants of skills, to also consider parenting practices, may provide a more informative model for testing the impact of ECIs in a European setting.

With this in mind, this study explores the role of intensive and continued investment in parenting from pregnancy until entry into formal schooling within a highly disadvantaged community in Dublin, Ireland. Theoretically, if the *in utero* and infancy periods are critical for optimizing brain development, and parenting and the quality of the home environment is strongly implicated in the development of children's skills, then intervening early and focusing on parents may generate larger effects than centre-based pre-school programs on which much of the ECI literature is based. The program, known as *Preparing for Life* (*PFL*), incorporates a home visiting program from pregnancy until age five, baby massage classes in the first year, and group-based parenting classes in the second year. The program aims to reduce SES inequalities in children's school readiness skills by working directly with parents to improve their knowledge of child development and parenting, as well as encouraging greater stimulation and investment in their children. Previous reports of the *PFL* trial have identified some treatment effects at earlier ages, primarily using parent report measures of children's health and skills.[6] This paper examines the impact of the program on children's cognitive, language, socio-emotional, and behavioral development during the program at 24, 36, 48 months of age and at the end of the program at 51 months utilizing both parent report and direct assessment of children's skills.

By exploiting program design, the study makes a number of contributions to the empirical literature. First, unlike many ECI programs, the impact of intervening during pregnancy and sustaining the investment until school entry can be tested. The majority of home visiting programs, including NFP the most frequently cited program, operate from pregnancy until age two, yet building on the technology of skill formation, continued

[6] For example, Doyle *et al.* (2014) focus on birth outcomes utilizing hospital data and identify a significant treatment effect regarding a reduction in the incidence of caesarean section, yet no impact on any neonatal outcomes. Doyle *et al.* (2017a), the only other study to date to examine the program's impact on children's skills, finds no effect on parent reported cognitive or non-cognitive skills at 6, 12, or 18 months, yet there are significant improvements in the quality of the home environment at 6 and 18 months. O'Sullivan, Fitzpatrick, and Doyle (2017) find evidence of improved nutrition at 24 months in terms of increasing protein intake, and Doyle *et al.* (2015) identify a number of significant treatment effects for parent reported child health at 24 months in terms of reducing the incidence of asthma, chest infections, and health problems. Finally, Doyle *et al.* (2017b) find few treatment effects on maternal well-being.

investment may be required to foster appropriate parental investment in response to the child's growing skill set (Heckman and Mosse 2014).

Second, much of the ECI literature focuses on primiparous parents. While first time parents may be more receptive to external support given the increased sense of vulnerability associated with first pregnancies (Olds *et al.* 1999), multiparous parents face additional financial and time constraints (Behrman, Pollak, and Taubman 1982; Becker and Tomes 1986). As the *PFL* program is provided to all women regardless of parity, the program's impact on non-firstborn children can be tested.[7] Thus, tests for differential treatment effects by parity status are conducted. Similarly, as differential treatment effects for girls and boys have been identified for interventions starting later in childhood, (e.g. Anderson 2008; Eckenrode *et al.* 2010; Heckman *et al.* 2010), differential treatment effects by gender are also tested to determine how early such potential differences may emerge.

Third, the *PFL* program operated in Ireland between 2008 and 2015; a period in which, despite national financial difficulties, the social welfare system of payments to disadvantaged families[8] and the 'care as usual'[9] package for mothers and children was largely retained; both of which are more substantial than the countries frequently studied in this field. The most similar European study is an experimental evaluation of *Pro Kind*, a German version of NFP, which included first time mothers only and ended at age two

[7] While not the focus of the current paper, spillover effects to older and younger children in the family can also be explored.

[8] The generous welfare system in Ireland, particularly for disadvantaged families, can be demonstrated by analyzing the tax wedge, a measure of taxes on labor income paid by employees and employers, minus family benefits through cash transfers received, as a percentage of the labor costs of the employer. Ireland has the lowest tax wedge out of 35 OECD countries for a single person with two children earnings 67 percent of average earnings (IE = -24 percent, OECD = 17 percent), and ranks the third lowest for a one-earner married couple with two children earning 100 percent of average earnings (IE = 7 percent, OECD = 26 percent). The negative tax wedge for disadvantaged families (i.e. low earning, lone parent households, which typify our *PFL* sample), shows that low SES working families receive more State benefits relative to taxes paid, compared to every other OECD country. These figures are calculated using the OECD's Taxing Wages database 2017, and are based on the average tax wedge during the period of the study (2008-2015). Regarding general welfare support, in addition to child benefit payments, which is a universal payment made to all families in Ireland currently amounting to €140 per child per month, participants in the *PFL* trial were in receipt of a number of additional mainly means-tested social welfare payments. Appendix Table A1 lists the proportion of *PFL* households receiving non-universal welfare payments when their children were 48 months old. In total, 87 percent of *PFL* households were in receipt of some form of non-universal welfare payment, with the largest categories being Medical Card (78 percent), One-Parent Family Payment (40 percent), and Unemployment Assistance (17 percent).

[9] Care as usual, which is available to all pregnant women and infants in Ireland, involves an initial family doctor (G.P.)/obstetrician appointment at 12 weeks and a further five examinations for first time mothers and six for subsequent pregnancies. Antenatal classes are provided by local public maternity hospitals free of charge. Following birth, a G.P. examination is carried out for the baby at two weeks and the mother and baby at six weeks. All mothers are entitled to free in-patient, out-patient, and accident and emergency/casualty services in public hospitals in respect of the pregnancy and the birth and is not liable for any hospital charges. In addition, checks by a public health nurse are carried out in the home in the weeks after birth and when the infant is nine, 18, and 24 months, but they are not mandatory. A schedule of immunizations is provided free of charge at birth, two, four, six, 12, and 13 months.

(Sandner and Jungmann 2017).[10] Thus, by studying the *PFL* program, this paper can examine whether the impact of ECI varies in a context of extensive welfare supports for vulnerable families.

Fourth, the *PFL* study also benefits from richer baseline data and more frequent assessment points than is typically found in the ECI literature. By collecting a wide range of data capturing parent's personality traits, IQ, parenting knowledge, social support networks, as well as standard socio-demographic and health data, the baseline equivalence of the randomized groups can be established and a comprehensive test of differential attrition can be conducted. In addition, by measuring multiple dimensions of children's skills including general IQ, verbal ability, spatial ability, pictorial reasoning, problem solving, communication, externalizing behaviors, internalizing behavior, socio-emotional competencies, pro-social behaviors, and peer problems, the areas of skill most impacted by early investment can be fully understood.

Fifth, the study embeds a series of innovative design features to test the internal validity of the trial. For example, the use of 'blue-dye' questions[11] permits a direct test for the presence of contamination, and the use of social desirability questions enables a test for performance bias, while using a computerized randomization procedure, with automated recording of treatment assignment, ensured that the randomization procedure was not compromised[12]. The external validity of the study is also assessed by comparing trial participants to eligible non-trial participants. This is a significant contribution as many studies of RCTs, both in the ECI field and more generally, fail to consider those who were eligible for inclusion but did not participate. In addition, trial participants are also compared to a large representative cohort of Irish children, thus testing whether the program was successful at reducing socioeconomic inequalities in children's skills.

Sixth, the study employs a number of methods to address common statistical issues in RCTs. Specifically, exact permutation testing is used to account for non-normality which is frequency associated with small samples, inverse probability weighting utilizing detailed

[10] The *Pro Kind* study benefitted from a larger sample size and direct assessment of children's skills at earlier ages compared to the *PFL* study. They found significant treatment effects at six and 12 months for girls' cognitive development, but not for boys. In addition, the effects had mostly faded by 24 months (Sandner and Jungmann 2017).

[11] 'Blue-dye' questions ask participants in the treatment and control groups specific questions which only the treatment group should be able to answer (as the information is part of the treatment). If the control group correctly answer these questions it is evidence that contamination may have occurred.

[12] This was important given evidence of compromised randomization in some of the most influential early childhood interventions such as the Perry Preschool Program (Heckman *et al.* 2010).

baseline data is used to account for differential attrition, and the stepdown procedure is applied to account for multiple hypothesis testing. These methods have been employed in earlier outcome studies of the *PFL* trial (e.g. Doyle *et al.* 2015; Doyle *et al.* 2017a), and in some recent studies of other ECI programs (e.g. Heckman *et al.* 2010; Campbell *et al.* 2014).

The findings in this paper indicate that the *PFL* intervention has a large and substantive impact on children's cognitive, social, and behavioral development. The program raised general conceptual ability, which is a proxy for IQ, by 0.77 of a standard deviation, indicating the malleability of IQ in the early years. Gains are found across all dimensions of cognitive skill including spatial ability, pictorial reasoning, and language ability. The program significantly reduced the proportion of children scoring below average and increased the proportion of children scoring above average, thus impacting the entire distribution of cognitive skills. These results, based on direct assessment, are supported by significant treatment effects found for parent-reported scores eliciting children's ability from age two onwards. While weaker, the program also impacted several dimensions of non-cognitive skills including externalizing problems such as aggressive behavior, and prosocial behavior such as helping other children. In particular, the program reduced the proportion of children scoring in the clinical range for behavioral problems by 15 percentage points. Contrary to much of the literature, there is little evidence of differential treatment effects by gender. In contrast, the effects are stronger for first born than non-first born children across certain domains, providing some evidence of differential effects by parity status. The size of the treatment effects exceed current meta-analytic estimations in the field (e.g. Sweet and Appelbaum 2004; Gomby 2005; Filene *et al.* 2013) and the results are robust to adjustments made to account for multiple hypothesis testing, differential attrition, baseline differences, contamination, and performance bias. The comparison of the *PFL* treatment groups to a large nationally representative sample of Irish children provides evidence that the program narrowed the socioeconomic gap on some dimensions of children's skills.

The remainder of the paper is structured as follows. Section I describes the study design including the program setting, recruitment and randomization, the intervention under investigation, data, baseline analysis, and the study sample and attrition. Section II outlines the empirical model and statistical methods. Section III presents the main results and robustness tests. Finally, Section IV concludes.

I Study Description

A. Setting & Program Design

The study took place between 2008 and 2015 in a community in Dublin, Ireland which was developed as a social housing initiative in the 1970s to relocate families from tenement buildings in the city center to newly built low-rise housing estates on the outskirts of the city (Brady *et al.* 2005). The community's disadvantaged status was exacerbated in the 1980s when a Government grant encouraging private home ownership resulted in many of the more advantaged families leaving the community (Threshold 1987). The vacant public housing was then populated by marginalized residents characterized by high rates of welfare dependency and lone parenthood. Census data collected prior to program demonstrates high rates of unemployment (12 percent vs national average of 3.5 percent), low levels of education (7 percent completed college degree vs national average of 19.4 percent), and high rates of public housing (42 percent vs national average of 7.2 percent) (Census 2006). The disadvantaged status of the community was also evidenced by the children who consistently scored below the norm in terms of cognitive and language development, communication and general knowledge, physical health and well-being, social competence, and emotional maturity (Doyle, McEntee, and McNamara 2012).

In an effort to break the intergenerational cycle of disadvantage in the community and to address these low levels of skills, the *Preparing for Life* (*PFL*) program was developed as part of the Government's and The Atlantic Philanthropies' Prevention and Early Intervention Program (Office of the Minister for Children and Youth Affairs 2008), by 28 local agencies and community groups. Based on evidence of the importance of the prenatal environment and the early years, the program aims to improve children's health and development by intervening during pregnancy and working with families for five years. The program is thus characterized by two key principles of effective interventions - programs which begin earlier in the lifecycle and are more intensive are typically more effective (Ramey and Ramey 1992).

B. Recruitment & Randomization

The study's inclusion criteria included all pregnant women residing in the designated *PFL* catchment area during the recruitment period. There were no exclusion criteria within the catchment area in order to avoid the stigmatization which may arise with highly selective inclusion criteria. Participation into the program was voluntary and recruitment took place between the 29th of January 2008 and the 4th of August 2010 through two maternity hospitals and/or self-referral using a community-based marketing campaign. Based on estimates of a two to five point difference on standardized cognitive development scores (i.e., average standardized effect size of 0.184) from a meta-analysis of home visiting programs (Sweet and Appelbaum 2004), a sample size of approximately 117 in each group was required to power the study.[13]

In total, 233 participants were recruited by the *PFL* recruitment officers. This represents a recruitment rate of 52 percent based on the number of live births during the recruitment period. For the remaining 48 percent, initial contact was made with 26 percent in the hospital or in the community, but they could not be subsequently contacted or they refused to join the program, and a further 22 percent never had any contact with the recruiters. To test for selection into the trial, a survey was carried out through the local childcare centres when the children of eligible non-participants were four years old. The survey included questions about the family's current socio-demographic characteristics and retrospective questions relating to their characteristics during the recruitment window. The results presented in Appendix Table B1 suggest that the eligible non-participants are of a somewhat higher SES than the participants who joined the program. While there are no statistically significant differences regarding maternal age, family size, parity, relationship status, or type of employment during pregnancy, trial participants were younger at the birth of their first child, have lower levels of education, were less likely to be employed, and were more likely to be eligible for free medical care compared to non-participants. This implies that there may have been some selection into the trial among lower SES families, assuming that the non-participants who completed the retrospective survey are representative of all non-participants. These findings suggest that the program was effective in recruiting families with the highest level of need.[14]

[13] It was not possible to oversample to capture anticipated attrition due to the low birth rate in the catchment area.
[14] The lower take-up rate among employed mothers may reflect the time intensive nature of the intervention.

Of those who joined the program, an unconditional probability randomization procedure, with no stratification, assigned 115 to a high treatment group and 118 to a low treatment group.[15] During the recruitment meeting, the participants initiated their own randomization by touching the screen of a tablet laptop.[16] This generated an email which was automatically sent to the program manager and the principal investigator (the author) listing the participant's permanent treatment condition and identification code. Any attempts to compromise randomization by reassigning participants would trigger an additional email highlighting any intentional subversion of the randomization process. This procedure ensured that treatment assignment was not exposed to randomization bias.

C. Treatment

Figure 1 describes the treatments provided to the high and low treatment groups. The high treatment consists of three components - a 5 year home visiting program, a baby massage course in the first year, and the Triple P Positive Parenting Program in the second year. The treatments are founded on the theories of human attachment (Bowlby 1969), socio-ecological development (Bronfenbrenner 1979), and social-learning (Bandura 1977). The home visits aimed to promote children's health and development by building a strong mentor-parent relationship and focusing on the identification of developmental milestones, appropriate parenting practices, and encouraging enhanced stimulation. The visits started in the prenatal period and continued until school entry at age four/five.[17] Twice monthly home visits of approximately one hour were prescribed and delivered by mentors from different professional backgrounds including education, social care, and youth studies. The mentors were hired to deliver the *PFL* program on a full-time basis and they received extensive training prior to treatment delivery. Mentor supervision took place

[15] As stated in the trial registry (www.controlled-trials.com/ISRCTN04631728/), 100 parents from a non-randomized external comparison group from another community were also recruited as a quasi-experimental component. This external comparison group is not included here as direct assessment data assessing cognition at the end of the program were not collected from this group.

[16] Actively involving participants in the randomization procedure helped to ensure that they trusted that the procedure was truly random and that a judgement on their parenting ability was not being made. Data capturing participants' automatic response to treatment assignment shows that 98% were 'happy' with their group assignment.

[17] Participants were on average 21 weeks (SD 7.4 weeks; range 5-40 weeks) pregnant when they joined the program, with 13 percent of the cohort joining in the first trimester, 55 percent in the second trimester, and 32 percent in the third trimester.

on a monthly basis to ensure fidelity to the program model.[18] Families were allocated the same mentor over the course of the intervention where possible.[19]

Each visit was structured around 210 *PFL*-developed 'Tip Sheets' which included information on pregnancy, parenting, health, and development (see Appendix C for an example of a Tip Sheet and a list of all Tip Sheets topics). The mentors could choose when to deliver the Tip Sheets based on the age of the child and the needs of the family, yet the full set of Tip Sheets must have been delivered by the end of the program. The mentors used a number of techniques to deliver the intervention including role modelling, coaching, discussion, encouragement, and feedback, as well as directly interacting with the *PFL* child. Each home visit began with an update on the family's situation and a discussion of whether the goals agreed at the previous visit were achieved. The mentor would then guide the parent through the Tip Sheet(s) selected for that visit and following this, new goals would be agreed.[20] While some Tip Sheets targeted multiple aspects of development, an analysis of Tip Sheet content found that 12 percent (n=22) encouraged the development of cognitive skills, such as learning numbers and colours; 14 percent (n=25) focused on language development, such as how to pronounce sounds and reading activities; 16 percent (n=30) encouraged children's development of positive approaches to learning, such as using play to encourage children to learn; 33 percent (n=60) dealt with social and emotional development including issues such as attachment, routine, regulation, and relationships; and finally, the largest majority of Tip Sheets addressed physical wellbeing and motor development (59 percent, n=105), such as general child health, immunization, nutrition, safety, and sleep.[21]

[18] The training included an intensive two-day workshop on the *PFL* program, with a focus on the program manual, and included topics such as the evidence-base for mentoring programs, relationships and activities, outcomes and evaluation, policy and practice alignment, and the *PFL* logic model. They also received 21 other relevant courses conducted over a six month period including child protection, attachment theory, and team building. Mentor supervision during the trial was based on the model commonly used by social workers in Ireland and was provided for two hours per month. Key areas addressed during supervision included participant work, team work, support, administration, and training/development.

[19] There were five mentors in total who had a caseload of 25 families each on average, with a lower caseload assigned to the mentor team leader. Participants were randomly assigned to the mentors by the team leader, yet provisions were made to ensure that all mentors had an equal number of high risk families. There was relatively little mentor turnover over the eight year implementation period, however two mentors left and were replaced before the end of the program, and one was absent for a period due to maternity leave.

[20] While both mothers and fathers were encouraged to participate in the home visits, in the majority of cases, the visits were attended by mothers only.

[21] Note that these figures do not sum to 210 as some Tip Sheets are categorized into more than one area. In addition, 178 Tip Sheets focused on promoting child outcomes and the remainder targeted parental outcomes.

Participants in the high treatment group were also encouraged to take part in a baby massage course in the first ten months of their child's life. The course consisted of five two-hour individual or group sessions delivered by one of the mentors. The purpose of these classes was to equip parents with baby massage skills and to emphasize the importance of reciprocal interactions and communication between parents and infants. A systematic review of 34 RCTs of infant massage found limited effects on child outcomes, although the authors note the low quality of many of the included studies (Bennett, Underdown, and Barlow 2013). Baby massage was included as part of the *PFL* treatment as an enjoyable activity which encouraged early engagement with the program.

When the *PFL* children were between two and three years old, the high treatment group were invited to participate in the *Triple P Positive Parenting Program* (Sanders, Markie-Dadds, and Turner 2003) which was delivered by the mentors. The goal of *Triple P* is to encourage positive, effective parenting practices in order to prevent problems in children's development. The program is based on five principles including providing a safe, engaging environment, the home as a positive place to learn, setting of rules and boundaries, realistic expectations of children, and parental self-care (Sanders 2012). Meta-analysis of the impact of *Triple P* has identified improved parenting practices and child social, emotional, and behavioral outcomes (Sanders *et al.* 2014). *Triple P* consists of five treatment levels of increasing intensity including a media campaign and communication strategy, a positive parenting seminar series, single session discussion groups, intensive small group and individual programs, and intensive family intervention. The high treatment participants were specifically encouraged to take part in the small group program which consisted of five two-hour group discussion sessions and three phone calls.

[Insert Figure 1 here]

In addition to care as usual, both the high and low treatment groups received a supply of developmental toys annually (to the value of ~€100 per year) including a baby gym, safety items, and developmental toys such as puzzles and memory games. They also received four book packs containing between six and eight developmentally appropriate books. The groups were also encouraged to attend community-based public health workshops on stress management and healthy eating, as well as social events such as coffee mornings and Christmas parties organized by the *PFL* staff. Program newsletters and birthday cards were sent annually to each family, in addition to two framed

professional photographs taken shortly after birth and at the end of the program. The low treatment group also had access to a *PFL* support worker who could help them avail of community services if needed, while this function was provided by the mentors for the high treatment group. Finally, all participants received a €20 shopping voucher for participating in each of the research assessments. Note that the low treatment group did not receive the home visiting program, Tip Sheets, baby massage classes, or the Triple P program. Further information on the study design may be found in Doyle (2013).

D. Dosage

There was considerable variability in treatment intensity across families. The average number of home visits delivered to the high treatment group between program entry and program end was 49.7 (SD = 38.1, range 0 - 145), which equates to just less than one visit per month. This represents 38 percent of prescribed visits which is somewhat less than the 50 percent which is typically found in shorter HVPs (Gomby *et al.* 1999). The number of visits decreased over the duration of the program – prenatal period (5.2 visits), birth to 12 months (12.1 visits), 12 to 24 months (9.9 visits), 24 to 36 months (11.0 visits), 36 to 48 months (7.3 visits), and 48 months until school entry (4.3 visits). This may be attributed to participant fatigue or the strategy adopted by mentors to reduce the amount of contact time with families in the later stages of the program to ensure a successful transition to program exit. The average duration of each visit was just under one hour, and on average participants received 50.6 hours of the home visiting treatment.

There was, however, large variability in dosage with 17 percent of high treatment families not participating in any home visits and 16 percent receiving over 90 visits.[22] Restricting the analysis to participants in the estimation samples increases the average number of home visits to 66, 69, 66, and 68 for the 24, 36, 48, and 51 month estimation samples respectively, which equates to approximately 50 percent of all home visits prescribed. Regarding the other high treatment supports, 43 percent of all randomized

[22] In order to test whether the number of home visits received varies as a function of family characteristics, separate bivariate regressions using 50 baseline measures are estimated. In total, nine of the 50 measures (18 percent) are significantly associated with the number of visits and there is some evidence that families with more favorable characteristics engaged in more home visits. In particular, mothers with higher IQ, older mothers, mothers who were employed during pregnancy, mothers with greater knowledge of infant development, and who have more positive parenting beliefs engaged in more home visits, whereas those who have a greater number of domestic risks and know more neighbors in the community engaged in less visits. For the purposes of this paper, an intention-to-treat analysis is conducted in line with other studies in the field.

high treatment families participated in some form of the Triple P program. Of those, the majority took part in the small group Triple P program (86 percent), with smaller proportions participating in the single session discussion groups (42 percent) and the intensive individual program (12 percent). The baby massage course was attended by 62 percent of all randomized high treatment participants.

In terms of the common supports available to both groups, 81 percent of the high treatment group and 77 percent of the low treatment group received at least one developmental pack, and 68 and 52 percent respectively attended a *PFL* social event. Finally, 77 percent of the low treatment group made contact with the *PFL* support worker at least once during the course of the program.

E. Data

Data were collected through face-to-face assessments conducted in participants' homes at baseline and when the children were 6, 12, 18, 24, 36, and 48 months old. Direct assessments of children were conducted in either the family's home, the local community center, or the participant's childcare setting when the child was 51 months old on average. To minimize detection bias, all assessments were conducted by trained researchers who were blind to the treatment condition and not involved in intervention delivery (Eble, Boone, and Elbourne 2016). This paper uses data from baseline, 24, 36, 48, and 51 months. Results on child cognitive and non-cognitive outcomes at 6, 12, and 18 months are reported in Doyle *et al.* (2017a).

Two broad areas of children's development are assessed. Cognitive development captures information processing, conceptual resources, perceptual skill, and language learning and is measured using the Communication and Problem Solving domains of the parent reported *Ages and Stages Questionnaire* (ASQ; Squires *et al.* 1999) and the *Developmental Profile cognitive development score* (DP-3; Alpern 2007) at 24, 36, and 48 months, and by direct assessment using the *British Ability Scales II: Early Years Battery* (BAS II; Elliott *et al.* 1997) at 51 months. The BAS II yields an overall score reflecting general cognitive ability (General Conceptual Ability, GCA), as well as three standardized scores for Verbal Ability, Pictorial Reasoning Ability, and Spatial Ability.

Socio-emotional and behavioral development represents the ability to engage effectively in social interactions, to perceive and interpret social skills accurately, and to

regulate emotional responses. It is assessed using parental reports on the *Child Behavior Checklist for Ages 1½ -5* (CBCL; Achenbach and Rescorla 2000) at 24, 36, and 48 months, the *Brief Infant-Toddler Social and Emotional Assessment* (BITSEA; Briggs-Gowan and Carter 2006) at 24 and 36 months, and two sub-domains of the *Strengths and Difficulties Questionnaire* (Goodman 1997) at 48 months. The CBCL yields a Total Problems Score, an Externalizing Problems score, and an Internalizing Problems score. The BITSEA yields a Problem score and a Competence score. The SDQ subdomains used in this study yield a Prosocial Behavior score and a Peer Problems score.

To facilitate comparability, all continuous outcomes are standardized to have a mean of 100 and a standard deviation of 15. Cut-off scores representing the proportion of children scoring below and/or above average are generated for all instruments based on representative norms. Please see Appendix D for detailed information on all outcomes.

F. Baseline Analysis

Baseline data from 205 participants (representing 90 percent of the high treatment group and 86 percent of the low treatment group) were collected after randomization yet prior to treatment delivery when participants were on average 21.5 weeks pregnant.[23] The baseline variables include 117 measures of socio-demographics, physical and mental health, IQ, parenting attitudes, self-control, self-esteem, among others (see Doyle 2013 for the full list). To assess the effectiveness of the randomization procedure, the baseline characteristics of the high and low treatment groups are compared using separate permutation tests across all 117 measures. At the 10 percent significance level, the two groups differ on 7.7 percent (9/117) of measures, which is consistent with pure chance and indicates the success of the randomization process (see Doyle and *PFL* Evaluation Team 2010).[24] In addition, a joint test of the baseline measures fails to be rejected, again suggesting that the thorough randomization procedure was successful. Regarding the few observed statistically significant differences, there are no systematic patterns in the data.[25]

[23] Of the 233 randomly assigned participants, two (high=one; low=one) miscarried, 19 (high=six; low=13) withdrew from the program before the baseline assessment, and seven (high=four; low=three) did not participate in the baseline but participated in subsequent waves. An analysis of a subset (n = 12) of this group on whom recruitment data but no baseline data are available, implies they do not differ on age, education, employment, and financial status from those who did complete a baseline assessment, however the limited sample size should be taken into consideration.

[24] Given the relatively small sample, a 10 percent significance level is adopted throughout.

[25] High treatment mothers were more likely to be at risk of insecure attachment, reported lower levels of parenting self-efficacy, were more likely to have a physical health condition, and were less considerate of future

The presence of such an extensive range of baseline variables, allows us to test for selection on observables, while minimizing the issue of selection on unobservables.

G. Study Sample and Attrition

Figure 2 depicts the families' participation in the trial between program entry and 51 months. Follow-up data was collected from 166 participants at 24 months (high = 71 percent; low = 71 percent), 150 participants at 36 months (high = 64 percent; low = 64 percent), 147 participants at 48 months (high = 64 percent; low = 62 percent), and 134 participants (high = 62 percent; low = 53 percent) at 51 months. Attrition is defined as either formally dropping out of the study or wave non-response. The level of attrition is largely equivalent across both groups over time and compares favorably with other home visiting programs (e.g., Guttentag *et al.* 2014). The 24 month participation rate of 71 percent is far higher than the 24 month participation rate of 46 percent in the only other equivalent European study (Sandner and Jungmann 2017).

[Insert Figure 2 here]

A re-examination of the comparability of the high and low treatment groups at baseline using the estimation samples is conducted using the same 117 measures. Table 1 presents a selection of the baseline characteristics capturing the main areas assessed i.e., socio-demographics, health and health behaviors, and maternal cognitive and non-cognitive skills. At the 10 percent significance level, the two groups differ on 6.8 percent (8/117) of measures using the 24, 36, and 48 month estimation samples, and on 10.3 percent (12/117) of measures using the 51 month estimation sample.[26] This is largely consistent with pure chance and indicates that the groups remain balanced at each time point, as confirmed by a joint test of all baseline variables for the estimation samples. Yet in order to account for any potential bias which differential attrition across the high and low

consequences, however they also demonstrated greater knowledge of infant development and reported using more community services than the low treatment group. More mothers in the low treatment group reported intentions to use childcare for their child and also intended to start their child in childcare at a significantly younger age than mothers in the high treatment group.

[26] As the group difference for the 51 month estimation sample falls just outside the 10% threshold, analyses conditioning on baseline differences are conducted as a robustness test.

treatment groups may introduce, treatment effects are estimated using the Inverse Probability Weighting procedure detailed below.[27]

As shown in Table 1, the participants represent a fairly typical at-risk cohort as characterized by low levels of education, IQ, and employment, and high rates of risky health behaviors during pregnancy. The sample is predominantly Irish, with approximately half being first time mothers, and an average age of 25 years old.

[Insert Table 1 here]

II Methods

Using an intention-to-treat approach, the standard treatment effect framework defines the observed outcome Y_i of participant $i \in I$ by:

$$(1) \qquad Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \qquad i \in I = \{1 \dots N\}$$

where $I = \{1 \dots N\}$ represents the sample space, D_i represents treatment assignment for participant i ($D_i = 1$ for the high treatment group, $D_i = 0$ for the low treatment group) and ($Y_i(0)$, $Y_i(1)$) are the potential outcomes for participant i. The null hypothesis of no treatment effect on children's skills is tested via:

$$(2) \qquad Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

Given the relatively small sample size, traditional hypothesis testing techniques which are based on large sample assumptions are not appropriate, thus the treatment effects are estimated using exact permutation-based hypothesis testing (see Good 2005). This method has been used in other studies of the *PFL* program (e.g. Doyle *et al.* 2015; Doyle *et al.* 2017a; Doyle *et al.* 2017b). As permutation testing does not depend on the asymptotic behavior of the test statistic, it is a more appropriate method to use when dealing with non-normal data (Ludbrook and Dudley 1998). A permutation test is based on the assumption of exchangeability under the null hypothesis. This means that if the null hypothesis is true, indicating the treatment has no impact, then taking random permutations of the treatment variable does not change the underlying distribution of

[27] As another simple test of attrition, treatment effects for a selection of cognitive and non-cognitive outcomes measured at 6 and 12 months were estimated by restricting the sample to the estimation samples at 24, 36, 48, and 51 months respectively. As shown in Appendix Table E1, the results do not differ depending on the estimation sample used, again suggesting that results are unlikely to be subject to attrition bias.

outcomes for the high or low treatment groups. Permutation testing has been shown to exhibit power advantages over parametric *t* tests in simulation studies, particularly when the degree of skewness in the outcome data is correlated with the size of the treatment effect (e.g. Hayes 1996; Mewhort 2005; Keller 2012). While this method is useful for dealing with non-normal data, it cannot be used to compensate for an under-powered study. Thus, the results from permutation testing may not differ from those using standard tests in a small sample, well-powered study with normally distributed outcomes. As a robustness test, standard OLS regressions are also estimated and noted.

Permutation tests are estimated by calculating the observed *t*-statistic. The data are then repeatedly shuffled so that the treatment assignment of some participants is switched (100,000 replications are used). The observed *t*-statistic is then compared to the distribution of *t*-statistics that result from the permutations. The mid-*p* value is reported and is calculated as follows:

$$MP(t) = P(t^* > t) + 0.5P(t^* = t) \tag{3}$$

where *P(.)* is the probability distribution, t^* is the randomly permuted *t*-statistic, and t is the observed *t*-statistic. Similar to other ECI studies (e.g. Heckman *et al.* 2010; Campbell *et al.* 2014; Gertler *et al.* 2014; Conti, Heckman, and Pinto 2016), one-sided tests with the accepted Type I error rate set at 10 percent are used given the small sample size and the hypothesis that the high treatment will have a positive effect on children's skills. However, results from two-tailed tests are also discussed.

As there was an imbalance in the proportion of girls and boys in the treatment groups at baseline, and given differential developmental trajectories by gender, all analyses control for gender.[28] As the assumption of exchangeability under the null hypothesis may be violated when controls are included, conditional permutation testing is applied. Using this method, the sample is proportioned into subsets, called orbits, each including participants with common background characteristics, in this case, there is one orbit for boys and one for girls. Under the null of no effect, the outcomes of the high and low

[28] The high treatment group has more boys than the low treatment group (54 percent vs 36 percent). As recruitment occurred during pregnancy, this difference cannot be attributed to the treatment. In addition, in Ireland, the majority of parents choose not to find out the gender of the baby until birth, therefore in most cases, recruitment occurred before the mothers knew the gender.

treatment groups have the same distributions within an orbit. The exchangeability assumption is thus limited to strata defined by the control variable - gender.

While the few observed group differences found at baseline are likely to be random, controlling for baseline covariates can improve the precision of treatment effects (Duflo, Glennerster, and Kremer 2008). Thus, as a robustness test, conditional permutation tests are estimated by controlling for key differences on which the high and low treatment groups differ and may also affect child outcomes i.e., maternal knowledge of child development, parenting self-efficacy, maternal attachment, and maternal consideration of future consequences. Partitioning the sample into multiple orbits based on variables such as these can prove difficult, as the strata may become too small leading to a lack of variation within each orbit. To address this, a linear relationship is assumed between the control variables and the outcomes. Each outcome is regressed on the four variables assumed to share a linear relationship with child skills and the predicted residuals are permuted from these regressions within the orbits. This method, known as the Freedman–Lane procedure (Freedman and Lane 1983), has been demonstrated to be statistically sound in a series of Monte Carlo studies (e.g., Anderson and Legendre 1999).

As shown above, the estimation samples are largely balanced in terms of baseline characteristics. Yet in order to investigate this more explicitly, the factors predicting participation in each assessment are tested using bivariate tests with 50 baseline measures.[29] Analyses are conducted separately for the high and low treatment groups to allow for differential attrition processes. In general, evidence of differential attrition is low, with between 12-20 percent of measures predicting attrition from the high treatment group, and between 8-20 percent of measures predicting attrition from the low treatment group depending on the assessment point (in two-tailed tests, with 10 percent significance

[29] Baseline measures are used as predictors of attrition as they cannot be influenced by the treatment. However, it is possible that the decision to remain in the study is influenced by child outcomes. For example, families whose children experience improved early developmental outcomes as a result of the treatment may be more likely to leave the program if they believe their children will not derive any additional benefits from staying. Conversely, such families may be more likely to remain in the study in order to maximize their children's ability. In order to test these hypotheses, measures of children's cognitive and non-cognitive skills measured at 6 and 12 months are used to predict the probability of remaining in the study at each assessment point. As shown in Appendix Table E1, there is very little association between early child outcomes and the probability of remaining in the study. In some cases, children with better skills are more likely to stay, while in other cases children with better skills are more likely to leave. This suggests that attrition is unrelated to the gains made by the children early in the study. Separate tests for the high and low treatment groups also reveal no discernible pattern in the results. A limitation of this analysis is that it is restricted to the sample who participated in the 6 or 12 month assessment, which is already subject to some attrition.

level).[30] In addition, the factors predicting attrition from both groups are largely similar. In line with much of the home visiting literature (see Roggman *et al.* 2008), families with higher risk factors are more likely to drop out of the study or miss an assessment, for example, they are less likely to be employed, have lower levels of education and IQ, are younger, and have poorer self-esteem and parenting skills.

In order to account for any potential bias due to differential attrition or wave non-response, an inverse probability weighting (IPW) technique (Robins, Rotnitzky, and Zhao 1994) is applied. First, logistic models are estimated to generate the predicted probability of participation in each assessment. Given the number of significant predictors from the individual bivariate tests (up to 10) and the relatively small sample size, the Bayesian Information Criterion (BIC; Schwarz 1978) is used to reduce the number of variables included in the logistic models while estimating the model with best fit.[31] The predicted probabilities from these logistic models are then used as weights in the permutation tests so that a larger weight is given to participants that are underrepresented in the sample due to attrition/wave non-response.[32] For completeness, the results from non-IPW adjusted results are also presented to examine the impact of the adjustment.

The issue of testing multiple outcomes at multiple time points, and thus increasing the likelihood of a Type-I error, is mitigated using the stepdown procedure which controls the Family-Wise Error Rate (Romano and Wolf 2005). Using this method the cognitive, socio-emotional, and behavioral outcome measures are placed into a series of stepdown families each representing an underlying construct. In this case the measures in each

[30] At 24 months, 12 and 8 percent of baseline measures significantly predict attrition from the high and low treatment groups respectively. At 36 months, the figures are 20 and 20 percent respectively. At 48 months, 15 and 17 percent. At the 51 month assessment, 14 and 12 percent.

[31] The BIC measures goodness of fit while penalizing for the number of variables included in the model. The procedure implemented in this paper is an iterative process. First, all 50 baseline variables are included in an OLS regression modelling attrition and the BIC is calculated and stored. The process continues by testing each combination of 49 baseline variables in order to determine whether dropping any baseline variable would result in an increase in the predictive power as measured using the BIC. Prior to beginning this iterative process, the 50 baseline variables are placed in ascending order according to their effect size (in terms of predicting attrition). When iterating through the combinations of baseline variables, the order in which variables are excluded depends on the effect size. Variables with the lowest effect size will be excluded first. For each combination of 49 variables, the new BIC is calculated and compared with the stored BIC. If the new BIC is smaller than the stored BIC (i.e. a lower BIC indicates a model with greater predictive power) the new BIC is stored and the excluded variable is dropped. A model resulting in a BIC that is within 2 points of the stored BIC is considered to have similar predictive power. Thus, only when the BIC is more than 2 points smaller is it considered a meaningful improvement in predictive power. This process is then repeated by testing all combinations of 48 baseline variables, and so on, until the optimal set of baseline variables has been found. The set of variables which result in the lowest BIC can be found in the Appendix Table G1. Separate models for the high and low treatment groups are conducted at each time point.

[32] Any participant who did not complete the baseline assessment yet completed assessments at later time points are assigned the average weight.

stepdown family are the same instruments measured at different time points. As the BAS and SDQ scores were only assessed at one time point, separate stepdown families are constructed for these. As a further robustness test, all continuous cognitive scores, all cutoff cognitive scores, all continuous non-cognitive scores, and all cutoff non-cognitive scores are placed in their own stepdown families in order to test whether the treatment has an impact on each type of skill (see Appendix H).[33]

The stepdown procedure is conducted by calculating a *t*-statistic for each null hypothesis in the stepdown family using permutation testing. The results are placed in descending order. The largest *t*-statistic is then compared with the distribution of maxima permuted *t*-statistics. If the probability of observing this statistic is $p \geq 0.1$ we fail to reject the joint null hypothesis. If the probability of observing this *t*-statistic is $p < 0.1$ the joint null hypothesis is rejected, and the most significant outcome is excluded, and the remaining subset of outcomes are tested. This process continues until the resulting subset of hypotheses fails to be rejected or only one outcome remains. By 'stepping down' through the outcomes, the hypothesis that leads to the rejection of the null is isolated.

III Results

A. Cognitive Skills

The IPW-adjusted means, standard deviations, and *p*-values that result from weighted individual and stepdown permutation tests, controlling for gender,[34] are reported in Table 2, alongside the treatment effect (mean difference between the high and low treatment groups) and the effect size (as measured by the ratio of the treatment effect and the standard deviation of the low treatment group).[35] The *p*-values that result from non-IPW weighted individual and stepdown tests are also presented in the final two

[33] In addition, stepdown families by each assessment point (24, 36, 48, and 51 months) are also estimated to test the impact of the treatment over time.

[34] The results excluding gender are largely similar to the main results. In two of the 30 models, outcomes which are statistically significant in the models including gender no longer reach conventional levels of significance in the unconditional models (i.e., ASQ communication score at 36 months and BAS pictorial reasoning ability above average cutoff score). In addition, in six models, results which are significant at the 5 percent level in models controlling for gender are significant at the 10 percent in the unconditional models. Results available upon request.

[35] The results are also estimated using standard OLS regression, controlling for gender and adjusted for IPW. There are no differences in the level of statistical significance between the permutation and OLS results for 28 of the 30 outcomes tested; for the two remaining outcomes, the results are significant at the 5 percent level in the permutation results and at the 10 percent level in the OLS results. Results available upon request.

columns for comparative purposes. As the IPW-adjusted and non-adjusted results are largely equivalent, only the IPW-adjusted results are discussed.[36]

[Insert Table 2 here]

The results indicate that the *PFL* program had a significant impact, both statistically and substantively, on children's cognitive skills from 24 months onwards. The high treatment group have significantly higher DP3 cognitive scores at 24, 36, and 48 months in the individual permutation tests, and the joint null hypothesis is rejected for the overall DP3 score stepdown family. The rejection of the null is driven by significant differences between the high and low treatment groups on DP3 at each time point. In terms of the size of the effects, the program increased DP3 scores by between 0.22 to 0.42 of a standard deviation, indicating that children in the high treatment group are more likely to be successful at tasks such as grouping objects by colour, shape, or size. Similarly, the high treatment group are significantly more likely to score above average on the DP3 at each time point, with effect sizes ranging from 12 to 17 percentage point differences between the groups. The joint null hypothesis is also rejected for the DP3 cutoff stepdown family.

While the DP3 measures general cognitive skills, the ASQ focuses on specific abilities including communication and problem solving skills. Fewer treatment effects are found using these measures. There is one significant treatment effect for communication scores at 36 months, with an effect size equating to 0.25 of a standard deviation. This result survives adjustment for multiple comparisons, suggesting that children in the high treatment group have a greater understanding of language and word combinations. Yet the proportion of children at risk of developmental delay in communication skills is largely equivalent at each time point, with very few children in either group scoring within the clinical range. For problem solving, there are significant treatment effects at 24 and 36 months for both the continuous and cutoffs scores, and the joint null hypothesis is rejected for the problem solving score stepdown family. The size of the effects are between 0.22 and 0.36 of a standard deviation, suggesting that children in the high treatment group are better able to follow instructions, engage in pretense, and solve problems. The differing results for the DP3 and ASQ may be a function of the reliability of the instruments. The

[36] In one case, the IPW-adjusted result reaches conventional levels of significance, whereas the non-IPW results did not (for BAS language ability above average cut-off score), however the opposite is also true (for ASQ communication score cut-off at 36 months).

Cronbach alpha for the DP3 (α=0.79-0.84) is considerably higher than the ASQ communication *(*α=0.49-0.78) and problem solving measures *(*α=0.27-0.55), suggesting greater internal consistency.

The DP3 and ASQ are maternal reported measures of cognitive skills and thus may be subject to social desirability bias, however the results for the BAS at 51 months, which is based on direct assessments of children and is generally considered a more reliable indicator of abilities (Najman *et al.* 2001), are similar and indeed larger. Significant treatment effects are identified for children's general conceptual ability (GCA), as well as their spatial ability, pictorial reasoning ability, and language ability. These effects are significant for both the continuous scores and the below average and above average cut-off scores which are based on a representative norm (an exception being the spatial ability above average score). In addition, the joint null hypothesis of no treatment effect is rejected for the overall BAS score stepdown family, as well as the BAS below average and BAS above average stepdown families. The sizes of the treatment effects are large. For example, the treatment increased children's general conceptual ability by 0.77 of a standard deviation, which demonstrates that the high treatment group are better at thinking logically, making decisions, and learning. These results for overall ability are not driven by one particular type of skill; the program impacted upon all forms of ability including spatial ability (0.65 of a standard deviation) which involves problem solving and coordination, pictorial reasoning (0.56 of a standard deviation) which involves the ability to detect similarities and knowledge of numbers, and also language ability (0.67 of a standard deviation) which involves the ability to understand and express language.

The significant results regarding the proportion of children scoring below average and above average suggest that the program has impacted the entire distribution of children's skills. This is demonstrated in Figure 3 which shows that the distribution of GCA scores for the high treatment group is shifted to the right of the low treatment group. In terms of the substantive effects, larger effects are experienced by those at the bottom of the distribution, with the program reducing the probability of scoring below average by 40 percentage points, and increasing the probability of scoring above average by 17 percentage points.

[Insert Figure 3 here]

In total, 22 of the 30 (73 percent) individual permutation tests and seven of the nine (78 percent) stepdown families reach conventional levels of significance using one-tailed tests.[37] If a more stringent two-tailed test is applied, 14 of the 30 (47 percent) individual tests and five of the nine (56 percent) stepdown tests are still statistically significant. The measures lost are largely confined to the weaker parent-report instruments, while the more objective measures assessed at the end of the program are robust to applying two-tailed tests.[38] Moving away from statistical significance, the high treatment group have more favorable outcomes compared to the low treatment group on 28 of the 30 (93.3 percent) cognitive measures studied, which is significantly different to the 50 percent one would expect if the program was having no impact, according to a two-sided binomial test (p<.0001). In sum, these results suggest that the program has an overall positive impact on children's cognitive ability.

B. Socio-emotional and Behavioral Skills

The IPW-adjusted means, standard deviations, and *p*-values that result from weighted individual and stepdown permutation tests controlling for gender[39] are reported in Table 3, alongside the treatment effects and effect sizes.[40] The *p*-values that results from non-IPW weighted individual and stepdown tests are also presented in the final two columns. Unlike the cognitive results, there are some differences between the IPW-

[37] Appendix Table H1 shows that when all the continuous cognitive scores are placed in one large stepdown family, the joint null hypothesis of no treatment effect is rejected. Similarly, the joint null hypothesis of no treatment effect is also rejected for the one large cutoff score stepdown family. In addition, when the stepdown families are defined by each assessment point, rather than by instrument, eight of the nine stepdown families (89 percent) are statistically significant.

[38] In particular, the DP3 continuous and cut-off scores at 36 and 48 months, the ASQ problem solving score at 36 months, all of the BAS continuous scores at 51 months, three of the four BAS below average cutoff scores, and two of the four above average scores, are statistically significant in two-tailed tests.

[39] The results excluding gender are similar to the main results. In one model, an outcome which is statistically significant in the model controlling for gender no longer reaches conventional levels of significance in the unconditional model (CBCL externalizing score at 36 months) In addition, in three models, results which are significant at the 5 percent level in the gender models are significant at the 10 percent in the unconditional models. Results available upon request.

[40] The socio-emotional and behavioral results are also estimated using standard OLS regression, controlling for gender and adjusted for IPW. There is no difference in the level of statistical significance between the permutation and OLS results for 28 of the 30 outcomes tested; for the two remaining outcomes, the results are significant at the 1 percent level in the permutation results and at the 5 percent level in the OLS results. Results available upon request.

adjusted and non-adjusted results. In general, fewer of the non-adjusted results reach conventional levels of significance. These cases are highlighted below.

[Insert Table 3 here]

The results indicate that the program has a significant impact on several dimensions of children's socio-emotional and behavioral development from 24 months onwards. The CBCL assesses problem behaviors in children regarding externalizing and internalizing behavior, as well as producing a total behavioral problems score. Regarding the continuous scores, the high treatment group have significantly lower total behavioral problems at 36 months and externalizing problems at 36 and 48 months, with effect sizes ranging from 0.21 to 0.31 of a standard deviation. However, the joint null hypothesis of no treatment effect for their respective stepdown families fails to be rejected, and the non-IPW adjusted results for these measures do not reach conventional levels of significance. In contrast, there are a number of significant treatment effects for the CBCL cutoff scores. In particular, the program reduced the proportion of high treatment children at risk of clinically significant problems at every time point for total behavioral problems and externalizing problems, and for two of the three time points for internalizing problems (24 and 48 months). In addition, the joint null hypothesis for the total, externalizing, and internalizing problems stepdown families is rejected, although the stepdown family for internalizing problems fails to be rejected in the non-IPW adjusted results. The size of the treatment effects are also large; the program reduces the probability of being at risk of clinically significant problems by between 7 and 15 percentage points for total problems, between 4 and 16 percentage points for externalizing problems, and between 7 and 17 percentage points for internalizing problems depending on the time point. Thus, the high treatment group is less likely to exhibit both externalizing behaviors, such as aggressive behavior and problems with attention, and internalizing behaviors, such as anxiety and emotionally reactivity.

The BITSEA and the SDQ are used to measure children's socio-emotional problems. The BITSEA consists of two sub-domains measured at 24 and 36 months – 'competencies' which measures areas of attention, compliancy, mastery, motivation, pro-social peer relations, empathy, play skills and social relatedness, and 'problems' which measures externalizing and internalizing behavior and dysregulation. As shown in Table 3, the program has no impact on competencies at either time point, however there is an impact

on problems at 24 months for both the continuous and cutoff scores, with an effect size for the continuous score of 0.24 of a standard deviation. In addition, the stepdown family for the continuous scores survives adjustment for multiple comparisons. The stepdown family for the cutoff scores also survives adjustment in the non-IPW adjusted results.

The SDQ includes two sub-domains measured at 48 months – prosocial behavior which measures sharing and helping other children, and peer problems which measures problematic behavior with peers such as bullying and being solitary. The program has an impact on prosocial behavior for both the continuous and cut-offs scores, with an effect size for the continuous score of 0.35 of a standard deviation, but no impact on peer problems. The joint null hypotheses of no effect for the prosocial stepdown family is rejected in the IPW-adjusted results, but not in the non-adjusted results. Again, there is evidence that the significant treatment effects are mainly restricted to the instruments with greater reliability. For example, the CBCL total score (α=0.95-0.96), BITSEA problem score (α=0.85-0.87), and SDQ prosocial score (α=0.72), have higher internal consistency than the BITSEA competence (α=0.64-0.71) or SDQ prosocial (α=0.48) scores.

In total, 15 of the 30 (50 percent) individual permutation tests and five of the 12 (42 percent) stepdown families reach conventional levels of significance using one-tailed tests.[41] When a more stringent two-tailed test is applied, only seven of the 30 (23 percent) individual tests and four of the 12 (33 percent) stepdown tests are still statistically significant using the 10 percent cutoff. While many of the continuous scores are no longer statistically significant when two-sided tests are applied, the cutoff scores are less sensitive to this stricter criteria.[42] Moving away from statistical significance, the high treatment group have more favorable outcomes compared to the low treatment group on 27 of the 30 (90 percent) socio-emotional and behavioral measures studied, which is statistically significantly different to the 50 percent one would expect if the program was having no impact ($p < .0001$). In sum, these results suggests the program's impact on children's socio-emotional and behavioral skills is lower than on cognitive skills, as demonstrated by the

[41] Appendix Table H2 shows that when all the continuous socio-emotional and behavioral scores are placed in one large stepdown family, the joint null hypothesis of no treatment effect fails to be rejected, while the joint null hypothesis of no treatment effect for the one large cutoff score stepdown family is rejected. In addition, when the stepdown families are defined by each assessment point, two of the six stepdown families (33 percent) are statistically significant, namely the stepdown families for the cutoff scores at 24 and 48 months.

[42] In particular, the CBCL total cutoff scores at 24 and 36 months, the CBCL externalizing and internalizing cutoff scores at 24 and 48 months, and the SDQ prosocial score are statistically significant in two-tailed tests.

smaller effect sizes and less robust results. The findings for socio-emotional skills are mainly concentrated on those at-risk of clinically significant problems.

C. Conditioning on Baseline Differences

As a robustness test, the main results are re-estimated by conditioning on four variables on which there are significant differences between the high and low treatment groups at baseline and may impact child outcomes – namely maternal knowledge of child development, parenting self-efficacy, maternal attachment, and maternal consideration of future consequences. The results, provided in Appendix I, show that the conditional results for both the cognitive (Table H1) and socio-emotional and behavioral (Table H2) outcomes are largely equivalent to the main results (presented in Table 2) with some minor exceptions. For example, regarding the cognitive results, some effects which reached conventional levels of significance in the main results, i.e. ASQ communication score stepdown family, ASQ problem solving score at 24 months, and the ASQ problem solving cutoff at 24 months, are not statistically significant in the conditional results, while the BAS spatial ability above average score and the overall BAS above average stepdown families reach conventional levels of significance in the conditional results but not in main results. Regarding the socio-emotional and behavioral results, some effects which did not reach significance in the main results, such as CBCL total score at 24 months, CBCL internalizing score at 36 months, and the BITSEA problem cutoff stepdown family, are statistically significant in the conditional results. One result which reached significance in the main results, CBCL externalizing score at 48 months, is no longer significant in the conditional results. Thus overall, controlling for baseline differences does not substantially affect the main conclusions of the study.[43]

[43] Although there is no significant difference between the high and low treatment groups regarding maternal IQ scores (as measured using the Weschler Abbreviated Scale of Intelligence (WASI) at 3 months postpartum), given the importance of the intergenerational transmission of IQ, the conditional models are also re-estimated with the inclusion of maternal IQ. Overall, the pattern of results, in terms of both size and significance, is similar to the main results. In a few cases, results which were significant at the 5 percent level are significant at the 10 percent level when controlling for maternal IQ. Results available upon request. The robustness of the results, even controlling for such a large predictor of children's skills, adds confidence to the overall impact of the program.

D. Heterogeneous Effects

To explore potential heterogeneity in the program's impact for girls and boys and firstborn and non-firstborn children, IPW-adjusted OLS models including treatment by gender/parity interactions are estimated. The first set of panels in Tables 4 and 5 report the interaction models by gender. They show that in 29 of the 30 cognitive models and 29 of the 30 socio-emotional and behavioral models, the gender by treatment status interaction term does not reach conventional levels of significance, providing little evidence of differential treatment effects by gender. A comparison of the means scores (not shown but available upon request) shows that high treatment girls have more favorable cognitive outcomes compared to low treatment girls on 28 of the 30 (93 percent) measures studied, and for boys the corresponding figure is 27 of the 30 outcomes (90 percent), both are statistically significantly different to the 50 percent one would expect under the null ($p < 0.0001$ for girls; $p < 0.0001$ for boys). Regarding socio-emotional and behavioral outcomes, high treatment girls have more favorable outcomes compared to low treatment girls on 22 of the 30 (73 percent) non-cognitive measures studied, while high treatment boys perform better on all outcomes (100 percent) than low treatment boys, both are statistically significantly different to the 50 percent one would expect under the null ($p = 0.016$ for girls; $p < 0.0001$ for boys). These results differ from some of the ECI literature which often finds stronger effects for girls than boys.

The second set of panels in Tables 4 and 5 report the interaction models by parity status. They show that in 26 of the 30 cognitive models and 26 of the 30 socio-emotional and behavioral models, the parity by treatment status interaction term does not reach conventional levels of significance. Yet in the remaining eight models, the treatment appears to favor firstborn children.[44] A comparison of the mean scores (not shown) finds that for firstborn children, the high treatment group have more favorable cognitive outcomes compared to the low treatment group on 26 of the 30 (87 percent) measures studied, and for non-firstborn children, the high treatment group have more favorable outcomes on 22 of the 30 measures (73 percent), both are statistically significantly different to the 50 percent one would expect under the null ($p < 0.0001$ for firstborns; $p = 0.016$ for non-firstborns). Regarding socio-emotional and behavioral outcomes, firstborns

[44] In particular, the treatment by parity interaction terms are statistically significant for BAS verbal ability standardized score, BAS verbal ability below average and above average cutoff scores, BAS general conceptual ability above average cut-off score, CBCL total behaviors problems score at 24 and 36 months, CBCL internalizing problems standardized and cutoff scores at 24 months.

in the high treatment group have more favorable outcomes compared to firstborns in the low treatment group on 29 of the 30 (97 percent) measures studied, and for non-firstborns the corresponding figure is 20 of the 30 outcomes (67 percent), both are statistically significantly different to the 50 percent one would expect under the null ($p < 0.0001$ for firstborns; $p = 0.099$ for non-firstborns). This provides some evidence of differential treatment effects by parity. As the majority of home visiting programs target first-time parents only, it is difficult to contextualize these results within the literature.

[Insert Tables 4 & 5 here]

E. Are the Results Driven by Childcare?

Much of the ECI literature which has informed policy investments in the early years is founded on center-based preschool programs e.g. Perry Preschool, which have generated long-term positive returns (e.g., Heckman *et al.* 2010). These programs operate by creating a high quality educational environment for children outside of the family home. One potential explanation for the *PFL* results is that differences in exposure to childcare among the high and low treatment groups may have generated the positive treatment effects, particularly if the program directly encouraged or led high treatment parents to choose higher quality childcare. If this occurred, it may lead to an overestimation of the impact of *PFL*. On the other hand, if the low treatment group accessed higher quality childcare as a compensatory measure, this may lead to an underestimation. In order to examine these hypotheses, tests for differences in childcare use between the groups when they were 6, 12, 18, 24, 36, and 48 months old were conducted.[45] The results, presented in Table 6, reveal no statistically significant differences at any time point regarding the use, type, hours, cost, or quality of childcare between the high and low treatment groups.[46] This

[45] Two-tailed tests are reported given the unknown direction of any potential effect.

[46] The proportion of the sample using childcare (defined as more than 10 hours per week) increases over time, and by 48 months the majority of children in the high and low treatment groups had experienced some form of childcare, with children spending ~20 hours per week in care. Although not statistically significant, up until 24 months, a greater proportion of the low treatment group used 'any' form of childcare, but thereafter, the high treatment group used more childcare. Among those who used childcare, there were no differences in the use of formal childcare, which is defined as center-based care, and by 48 months, almost all children who participated in childcare used formal care. For those who paid for childcare, the average cost was relatively low for both groups (<€2 per hour), which can be attributed to the high level of subsidized childcare places for low SES families in Ireland. In addition, the lower cost at 48 months may reflect the national 'Free Pre-School Year in Early Childhood Care and Education Policy', which provides all children in Ireland with one year of center-based childcare in the year prior to school entry for three hours per day, five days per week, over a 38-week year.

suggests that the treatment effects are unlikely to be attributed to differences in exposure to childcare and that the treatment effects can be attributed to changes generated by the home visiting program, baby massage classes, and Triple P program.[47] This is important as such strategies are likely to generate positive spillovers for other children in the family, unlike preschool programs where only the target child is impacted.

[Insert Table 6 here]

F. Testing for Contamination

The potential for contamination or spillover effects within the *PFL* trial is relatively high as participants live in a small geographical area and randomization was conducted at the participant level rather than clusters of communities. Thus tracking contamination has been a key feature of the *PFL* study design since its inception, and a number of strategies have been used to measure information flows between the two groups (details on these strategies can be found in Doyle and Hickey 2013). Contamination may have occurred if participants in the high treatment group shared any of the materials or advice which they received from their mentors with participants in the low treatment group; resulting in treatment effects which are a lower bound.

Previous studies of the *PFL* program found little evidence of contamination as measured at six months (Doyle *et al.* 2017a) and 24 months (Doyle *et al.* 2015) using 'blue-dye' questions. These questions asked participants in both the high and low treatment groups whether they had heard of particular parenting strategies/behaviors and if they know how to engage in these behaviors with their child i.e. 'mutual gaze, 'circle of security' and 'descriptive praise'. These parenting strategies were discussed by the mentors during the home visits and they were described in the Tip Sheets. These questions may be used as proxies for contamination as, if a large proportion of participants in the low treatment group stated that they had heard of these phrases and they could correctly describe how to engage in these behaviors, it is indicative that they may have accessed material or information intended for the high treatment group only.

[47] As an additional check, the BAS models were estimated with controls for childcare use, age started childcare, and hours spend in childcare. The inclusion of these controls did not affect the statistical or substantive impacts of the main results. Results available upon results.

In this paper, the presence of contamination is tested using a blue-dye question asked at 48 months. Specifically, participants in the high and low treatment groups were asked if they have heard of the *'Feeling Wheel'* and if they knew what it is used for. The *'Feeling Wheel'* is a circular chart with cartoon faces showing different emotions. A Tip Sheet describing the '*Feeling Wheel*' was given to the high treatment group during the home visits between 36 and 48 months. The first row in Table 7 shows that a significantly greater proportion of the high treatment group (29 percent) reported knowledge of the phrase compared to the low treatment group (3 percent).[48] In order to provide a more accurate measure of contamination, participants who stated that they had heard of the phrase, yet incorrectly described it, were treated as reporting not knowing the phrase. The test was then re-estimated using the proportion of participants who accurately described the '*Feeling Wheel'* and the result is presented in the second row of Table 7. As before, it shows that a significantly greater proportion of the high treatment group (23 percent) reported knowledge of the phrase and could accurately describe what it is, compared to the low treatment group (2 percent).

A limitation of these analyses is that they are restricted to several discrete phrases, thus it is possible that the high treatment group may have shared material about other aspects of child development not captured by these particular phrases. Contamination, while often discussed in the context of RCTs, is rarely measured. Thus, in the absence of alternative measures, these proxies suggest that contamination may have been limited in the *PFL* trial. Indeed, minimal contamination may be expected as *PFL* is a complex and holistic intervention which attempts to change multiple aspects of parenting behavior by building long-standing relationships between mentors and families. As it is often difficult to achieve such behavioral change, even if contamination between the two groups exists, it may not be enough to meaningfully affect the results (Howe *et al.* 2007).

[48] The fact that just 29 percent of the high treatment group reported knowledge of '*the feeling wheel'* suggests that either mothers did not retain the information provided by the mentors or did not receive the information in the first place. Tests for contamination at six months found that 49 percent and 59 percent of the high treatment mothers reported knowledge of '*circle of security*' and '*mutual gaze'* respectively (Doyle *et al.* 2017a), while at 24 months, only 33 percent reported knowledge of '*descriptive praise*' (Doyle *et al.* 2015). It is possible that the high treatment groups' ability to retain knowledge of such terms declined as the program continued and more information was provided.

[Insert Table 7 here]

G. Testing for Performance Bias

A limitation of the outcomes assessed at 24, 36, and 48 months is that they are based on maternal reports of the child's abilities (e.g. ASQ, BITSEA, CBCL, SDQ) rather than direct assessments or observations. These subjective indicators may be subject to performance bias (McAmbridge, Witton, and Elbourne 2014; Eble *et al.* 2016) if parents in the high or low treatment groups either overestimate or underestimate their children's skills as a result of participation in the trial due to Hawthorne or John Henry effects. Such misreporting will not affect the results if parents in both treatment groups systematically misreport, however if parents in the high and low treatment groups misreport in different ways, the estimates of treatment effects may be biased. One may hypothesize that parents in the high treatment group may overestimate their children's abilities relative to the low treatment group as they are aware that the advice and materials provided by the mentors aim to specifically promote their children's development. It is also possible that the low treatment group, recognizing that they are not receiving intensive parenting supports, may underestimate their child's skills in an attempt to access additional services.

To address this issue, a number of instruments have been used to measure differential misreporting across the high and low treatment groups over the course of the trial. Doyle *et al.* (2017a) test for differences on the defensive responding subdomain of the Parenting Stress Index (PSI; Abidin 1995) assessed at 6 months, and find that the levels of misreporting was equivalent among parents in both groups. Doyle *et al.* (2015) test for differences on the Social Desirability Scale-17 (Stöber 2001) assessed at 24 months, and also find no evidence of social desirability bias across the high and low treatment groups. Both results showed that parents in the high and low treatment groups engaged in some level of misreporting, but the groups do not systematically differ in the direction or magnitude of misreporting.

In this paper, the defensive responding sub-domain of the PSI measured at 24 and 48 months is used to test for differential misreporting. This measure is based on a well-known social desirability instrument called the Crowne-Marlowe Scale and asks parents questions about their experience of routine parenting issues such as '*I feel trapped by my responsibilities as a parent*'. The rationale underlying this measure is that if parents deny

experiencing these common issues which face most parents, it suggests that they may be engaging in defensive, rather than accurate, responding in order to portray themselves more favorably to the interviewer. A score above 10 is indicative that the participant is engaging in defensive responding. A comparison of the high and low treatment groups on the defensive responding scores finds that, on average, both groups score above 10 at 24 months (*high*: 14.94(4.98), *low*: 15.13(4.82) and 48 months (*high*: 14.18(4.87), *low*: 15.13(4.41), however there are no statistically significant differences between the groups at either time point using two-tailed IPW-adjusted permutation tests controlling for gender (24 months: p = 0.294; 48 months: p = 0.804).[49] This suggests that while a certain proportion of participants attempt to portray themselves in a more positive light, there is no systematic misreporting across the groups, as found in earlier studies.

As a further check, in order to test the sensitivity of the main results based on the subjective outcomes, participants who scored above 10 on the defensive responding score at either 24 or 48 months are excluded from the analysis (n_{HIGH} = 27, n_{LOW} = 19), and the main treatment effects are re-estimated and reported in Appendix Tables J1 and J2. While there are somewhat fewer treatment effects (e.g. the 24 month DP3 score and cutoff score, and the BITSEA problem score and cutoff score), the overall pattern and magnitude of the results are the same, suggesting that the findings are not biased by differential misreporting or performance bias.

Indeed, significant correlations between the BAS score, measured at 51 months using direct assessment, and the 48 month parent-reported cognitive measures, also suggest that these parent reported measures are good proxies for children's underlying skills.[50] The use of parental reports, particularly when measuring children's socio-emotional and behavioral skills, is in line with the majority of the ECI literature, and another home visiting study targeting low income families also found a significant correlation between parent reports and direct assessments (Sandner and Jungmann 2016).

In sum, the estimates of treatment effects using the maternal reported measures should not be affected by performance bias, yet the BAS scores, which were directly measured at the end of the program by independent assessors, are the most reliable estimates of the treatment effects.

[49] At 24 months, 22 percent and 13 percent of the high and low treatment groups respectively score above 10 on the defensive responding measure (p = 0.167), while the corresponding figures at 48 months are 23 percent and 18 percent respectively (p = 0.460).

[50] Correlation between BAS score and DP3 score (r = 0.438; p<.0001), BAS score and ASQ problem solving score (r = 0.422; p<0.0001), and BAS score and the ASQ communication score (r = 0.434; p<0.0001).

H. Comparison with Nationally Representative Cohort

The key goal of the *PFL* program is to reduce socioeconomic inequalities in children's skills. In order to test whether the program was successful, the scores from the high and low treatment groups are compared to those from a nationally representative cohort of Irish children participating in the Growing Up in Ireland (GUI) Infant study.

The GUI is a longitudinal study of 11,134 infants born between December 2007 and May 2008 in Ireland (one-third of all births in this period), who were identified from the Child Benefit Register (Williams *et al.* 2010). GUI assessments were conducted at 9, 36, and 60 months. This cohort serves as a useful comparison for the *PFL* sample as it is a relatively contemporaneous cohort reflecting different social groups, and there is some overlap in the instruments used to measure children's skills.[51] Doyle *et al.* (2017a) present descriptive statistics for the GUI and the *PFL* cohorts at baseline. As expected, mothers in the nationally representative GUI cohort are significantly older than mothers in the *PFL* cohort and are more likely to be married and employed. They are also less likely to have low levels of education or live in public housing, and have less physical and mental health conditions, as well as reporting to engage in better health behaviors during pregnancy. Regarding common instrumentation, at 36 and 60 months, the GUI includes two sub-scales from the British Ability Scales (picture similarity scale and naming vocabulary scale[52]) which are assessed at 51 months in the *PFL* cohort, and two sub-scales from the parent-report Strengths and Difficulties Questionnaire (peer problems and prosocial behavior) which are assessed at 48 months in the *PFL* cohort. If the program is effective, one would expect the gap between the GUI cohort and *PFL* high treatment group to be smaller than the gap between the GUI cohort and the *PFL* low treatment group.

Table 8 compares the GUI cohort at 36 and 60 months and the *PFL* high and low treatment groups at 48 and 51 months across the common measures.[53] As expected, in almost all cases, the GUI sample has significantly better scores than the low treatment group. In particular, the GUI cohort has higher picture similarity, naming vocabulary, and lower peer problem scores, at both 36 and 60 months than the low treatment group at 48/51 months, as well as high prosocial behavior at 60 months. In contrast, the GUI sample

[51] The two to three year lag between the *PFL* and GUI studies is unlikely to affect the results assuming an absence of time trends in children's skills.
[52] The analysis is conducted using the BAS *t*-scores rather than the standardized scores as used in the main results.
[53] Two-tailed unpaired t-tests adjusted for attrition using the IPW generated weights for the *PFL* sample and the representative sample weights for the GUI sample are used.

has significantly lower naming vocabulary scores and prosocial behavior scores at 36 months than the high treatment group at 48/51 months, and there are no significant differences in picture similarity scores or peer problem scores measured at 36 months in the GUI sample and at 48/51 in the *PFL* cohort. The GUI sample has significantly higher picture similarity scores and lower peer problem scores at 60 months compared to the high treatment group at 48/51 months. However, there are no significant differences regarding naming vocabulary scores or the prosocial behavior scores as measured at 60 months for the GUI cohort and 48/51 months for the high treatment group. Indeed for prosocial behavior, the high treatment group has the highest score across all groups.

While the timing of the assessment points differ across the two cohorts, the pattern of the low treatment group consistently scoring below the GUI cohort, and the high treatment group either outperforming or scoring similarly to the GUI cohort, suggests that the *PFL* program was successful in narrowing the socioeconomic gap across some dimensions of children's skills.

[Insert Table 8 here]

IV Conclusions

Much of the evidence base on the effectiveness of early intervention programs is based on US studies, and more recently studies from the developing world. To date, we have limited robust evidence that such programs will be as effective or as cost effective in countries which provide relatively generous social welfare policies and comprehensive supports for women and children as standard practice. Based on evidence that the prenatal and infancy periods are critical for brain development, and that the quality of parenting is influential in the development of children's skills, the aim of this study was to explore the impact of a five-year prenatally commencing home visiting program in Ireland. Specifically, the paper examines the impact of the *PFL* program on children's cognitive, socio-emotional, and behavioral skills from 24 months until the end of the program at school entry.

Compared to other disciplines, Eble *et al.* (2016) demonstrate that RCTs conducted within the economic literature frequently fail to address many common risks of bias. In contrast, this study attempts to address the main risks of bias including *selection bias* by capturing data on eligible non-participants, *randomization bias* by using a tamper-proof randomization procedure, *attrition bias* by using IPW to adjust for differential attrition and

non-response, *performance bias* by testing for differential misreporting by participants, *detection bias* by using independent blinded assessors, and *reporting bias* by registering the study protocol. In addition, as there is minimal evidence of contamination across the high and low treatment groups, the internal validity of the study is high.

The results indicate that the program has a large and substantive impact on multiple aspects of children's skills. General conceptual ability, which is a close proxy for IQ, is increased by 0.77 of a standard deviation. As expected, the IQ scores of the children are above that of their parents (i.e. the Flynn effect), yet the correlation between high treatment children and their mothers is small and not statistically significant (r = 0.07, p = 0.562), compared to the larger and significant correlation between the low treatment children and their mothers (r = 0.31, p = 0.018).[54] Indeed, the correlation for the low treatment group is similar to the correlation of 0.38 between fathers and sons found in Black, Devereux, and Salvanes (2009), thus the program appears to be effective in reducing the intergenerational transmission of IQ scores within the high treatment group. The treatment effects are observed across all measures of cognitive skill including spatial ability, pictorial reasoning, and language ability, in addition to reducing the proportion of children scoring below average and increasing the proportion of children scoring above average. Thus, it is clear that the program shifted the entire distribution of children's cognitive skills. These results, based on direct assessment conducted by trained assessors, are supported by significant treatment effects found for parent-report instruments eliciting children's cognitive ability from age two onwards.

The program also has an impact on children's socio-emotional skills, although the effects are mainly concentrated among those most at risk of developing clinical problems. Children who received the high treatment supports are less likely to exhibit externalizing problems such as aggressive behavior, and are more likely to engage in prosocial behavior such as helping other children. The program also reduced the proportion of children scoring in the clinical range for behavioral problems by 15 percentage points, which is likely to have significant cost saving implications regarding future psychological treatment. The comparison of the treatment groups to a large nationally representative sample of Irish children demonstrates that the *PFL* program helped to close the socioeconomic gap in

[54] Maternal IQ was measured using the Weschler Abbreviated Scale of Intelligence (WASI) which measures cognitive ability across four subscales: vocabulary, similarities of constructs, block design, and matrix reasoning. From this, standardized measures of verbal ability, perceptual reasoning, and a full-scale measure of cognitive functioning, standardized to have a mean of 100 and standard deviation of 15, are generated. The full-scale measure was used in this analysis to correspond with the measure of General Conceptual Ability from the BAS.

children's vocabulary skills and prosociality, although they still lag behind the national average in terms of non-verbal ability and peer problems.

An analysis of heterogeneous effects by gender finds cognitive gains for both girls and boys. This is contrary to some of the existing literature which finds cognitive gains for girls only when measured later in childhood or adulthood (e.g. Anderson 2008; Heckman *et al.* 2010), although Sandner and Jungmann (2017) also find treatment effects at six and 12 months for girls, yet this effect had faded by 24 months. While the present study indicates that there are no differential effects by gender prior to school entry, it is possible that gender effects may emerge later in life. An analysis of heterogeneous effects by parity finds somewhat stronger treatment effects for firstborn children compared to non-firstborn children. As most home visiting programs target first time mothers it is difficult to contextualize these results, however there is some evidence that primiparous mothers derive more benefits from home visiting. For example, the Healthy Families America program which targets all mothers, finds significant treatment effects regarding early parenting practices for primiparous parents only (DuMont *et al*. 2008).

The magnitude of the effects on cognitive, socio-emotional, and behavioral development identified here are generally larger than those found in studies of other home visiting programs. A meta-analysis by Sweet and Appelbaum (2004) find an average standardized effect size of 0.18 for cognitive skills and 0.10 for non-cognitive skills, while Miller, Maguire, and Macdonald (2011) and Filene *et al.* (2013) find average standardized effect sizes of 0.30 and 0.25 respectively for cognitive skills. These compare to a standardized effect size of 0.77 for the general conceptual ability score reported here, and 0.24 for total behavioral problems. In addition, the effects are larger than the German home visiting program, which finds average effect sizes for cognition of 0.20-0.30 SDs for girls only (Sandner and Jungmann 2017). However, it is difficult to fully compare the results from different home visiting studies due to wide variations in program goals, target groups, and implementation practices (Gomby *et al.* 1999). For example, the larger effect sizes identified for the *PFL* program may be due to its greater program length and intensity, especially when compared to many of the other home visiting programs which typically end at age two. This suggests a potential role for sustained investment in parenting beyond the initial critical period of the first 1000 days; although further testing of the optimal timing of intervention is needed.

The *PFL* program is based on the premise that providing support to parents will increase their knowledge of appropriate parenting practices and change their attitudes and parenting behaviors. These positive changes would then impact on children's development as a result of the improved stimulation, interactions, and resources that parents would provide for their children. While very few treatment effects were observed for parental wellbeing measured using global and experienced instruments (see Doyle *et al.* 2017b), parents made a number of important behavioral changes which may have contributed to their children's advanced skills. For example, Doyle *et al.* (2017a) identify significant treatment effects for improved parenting skills at six and 18 months in terms of improving the quality of the home environment, while O'Sullivan *et al.* (2017) find positive treatment effects regarding improved nutrition at 24 months, and Doyle *et al.* (2015) find a number of significant effects on child health up to 36 months in terms of reducing the incidence of asthma, chest infections, and health problems. Previous *PFL* evaluation reports also identify a number of treatment effects for parenting behaviors (see Doyle and *PFL* Evaluation Team 2015 for example). Specifically, parents in the high treatment group were found to spend more time interacting with their children. They also exposed them to a greater variety of activities and provided opportunities for exploration. High treatment parents were also more understanding of their children's behaviors, were less likely to punish them unnecessarily, and were more likely to follow through on any necessary punishments. Their houses and routines were more organized, they were more involved in their children's learning, and their children spent less time watching TV. These practices, interactions, and activities are recognized as key means of stimulating children's cognitive and socio-emotional and behavioral development (Farah *et al.* 2008; Edwards, Sheridan, and Knoche 2010). The one other study to investigate the *PFL* program's impact on cognitive and non-cognitive outcomes found little evidence of treatment effects on child outcomes up to 18 months (Doyle *et al.* 2017a). Thus, cumulative improvements in parenting and parental behaviors over the course of the trial may account for the larger effects identified in this paper.

These changes in parenting may be attributed to the extensive and diverse supports offered to the high treatment group, including intensive mentoring, parent training, and baby massage classes. The *PFL* mentors worked with the participants for a substantial and critical period of their children's lives, therefore it is likely that these positive changes were a result of the strength and quality of the mentor-parent relationship. This is consistent

with the home visiting literature which finds that the relationship between parents and program staff is key for understanding program effects (Wesley, Buysse, and Tyndall 1997). The strength of these relationships, coupled with the high quality information from the Tip Sheets and Triple P, may have facilitated these behavioral changes. It is important to note, however, that as participants were not randomized to receive different components of the treatment bundle, it is not possible to tease out the impact of the three different provisions. The finding of no differences in childcare use across the groups also suggests that the results cannot be attributed to differences in center-based childcare on which much of the ECI literature is based. A full mediation analysis, such as that found in Heckman, Pinto, and Savelyev (2013), is required to fully understand the mechanisms underlying the treatment effects.

While the effects identified here, particularly for the cognitive outcomes at ages four to five, are large, it is possible that they may fade over time. Indeed, some ECI programs demonstrate fade-out on key cognitive outcomes (e.g. Heckman *et al.* 2010, Heckman *et al.* 2013), yet improved social, economic, and health outcomes later in the lifecycle (e.g. Heckman *et al.* 2010; Campbell *et al.* 2014), while other studies do not observe such cognitive fade-outs (e.g. Gertler *et al.* 2014). Thus, the full gains from the *PFL* program may not be realized until adulthood. The *PFL* programs costs approximately $US 2,250 (€2,000) per family per year to be delivered (for a total of $US10,125). Cost–benefit analyses of some of the most well-known US-based home visiting programs finds returns ranging from $US1.61 for the Nurse Family Partnership program, $US3.29 for Parents as Teachers, and $US1.21 for Healthy Families America per $US invested, with total program costs of $US10,049, $US2,688, and $US4,797 respectively (Washington State Institute for Public Policy 2016). In addition, cost-benefit analyses of the Head Start program by Ludwig and Philips (2007) and Deming (2009) find that effect sizes on cognitive skills of 0.10-0.20 SDs and 0.06 SDs respectively, are enough to satisfy cost-benefit tests, based on an average cost per child of ~$US7,000. Therefore, if the significantly larger effects (0.20-0.80 SDs) identified in this study translate into future financial gains both for the individual participants and wider society, the *PFL* program is likely to generate similar positive returns.

In sum, this study finds that a set of parenting interventions provided from pregnancy until age five has positive and statistically significant effects on children's skills. If one accepts the generalization of the results, the *PFL* program may provide a potential

vehicle for reducing the socioeconomic gradient in children's early skills, yet further replication and testing in other sites is needed.

References

Abidin, R.R. 1995. *Manual for the Parenting Stress Index*. Odessa, FL: Psychological Assessment Resources.

Achenbach, T.M., and Rescorla, L. 2000. *ASEBA preschool forms & profiles.* Burlington, VT: University of Vermont, Research Centre for Children, Youth, and Families.

Almond, D., and Currie, J. 2011. "Human Capital Development Before Age Five." *In Handbook of Labor Economics.* Vol. 4B, edited by Orley Ashenfelter and David Card, 1315–1486. Amsterdam: Elsevier, North-Holland.

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

Anderson, M.L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Re-evaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103: 1481-1495.

Anderson, M.J., and Legendre, P. 1999. "An Empirical Comparison of Permutation Methods for Tests of Partial Regression Coefficients in a Linear Model." *Journal of Statistical Computation and Simulation* 62 (3): 271–303.

Attanasio, O., Cattan, S., Fitzsimons, E., Meghir, C., and Rubio-Codina, M. 2015. "Estimating the Production Function for Human Capital: Results from a Randomized Control Trial in Colombia." *NBER Working Paper* No. 20965.

Avellar, S., Paulsell, D., Sama-Miller, E., Del Grosso, P., Akers, L., and Kleinman, R. 2016. *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.

Baker, A. J. L., and Piotrkowski ,C. S. 1996. *Parents and Children Through the School Years: The Effects of the Home Instruction Program for Preschool Youngsters.* New York, NY: National Council of Jewish Women, Center for the Child.

Bakermans-Kranenburg, M. J., van IJzendoorn, M. H., and Bradley, R. H. 2005. "Those Who Have, Receive: The Matthew Effect in Early Childhood Intervention in the Home Environment." *Review of Educational Research* 75: 1–26.

Bandura, A. 1977. "Self-efficacy: Toward a Unifying Theory of Behavioral Change." *Psycholology Review* 84: 191-215.

Becker, G. S. 1965. "A Theory of the Allocation of Time." *The Economic Journal* 75: 493-517.

Becker, G. S. and Tomes, N. 1986. "Human Capital and the Rise and Fall of Families." *Journal of Labor Economics* 4 (3): S1–S39.

Behrman, J. R., Pollak, R. A., and Taubman, P. J. 1982. "Parental Preferences and Provision of Progeny." *Journal of Political Economy* 90 (1): 52–73.

Bernal, R. and Keane, M. P. 2010. "Quasi-structural Estimation of a Model of Childcare Choices and Child Cognitive Ability Production." *Journal of Econometrics* 156 (1): 164–189.

Bennett, C,. Underdown, A., and Barlow, J. 2013. "Massage for Promoting Mental and Physical Health in Typically Developing Infants Under the Age of Six Months. *Cochrane Database of Systematic Reviews* Issue 4. Art. No.: CD005038.

Black, S., Devereux, P., and Salvanes, K. 2009. "Like Father, Like Son? A Note on the Intergenerational Transmission of IQ Scores." *Economics Letters* 105 (1): 138–140.

Bowlby, J. 1969. *Attachment and Loss, Vol. 1: Attachment*. New York: Basic Books.

Bradley, R. H., Caldwell, B. M., Rock, S. L., Ramey, C. T., Barnard, K. E., Gray, C., ... Johnson, D. L. 1989. "Home Environment and Cognitive Development in the First 3 Years of Life: A collaborative Study Involving Six Sites and Three Ethnic Groups in North America." *Developmental Psychology* 25: 217–235.

Bradley, R.H., and Corwyn, R.F. 2002. "Socioeconomic Status and Child Development." *Annual Review of Psychology* 53 (1): 371-399.

Brady, C., Byrne, A., French, S., Larkin, J., Hand, T., Lawlor, A., and Sherlock, E. 2005. *Darndale: A Living History*. Dublin: Darndale Belcamp Resource Centre Ltd.

Briggs-Gowan, M. J., and Carter, A. S. 2006. *BITSEA: Brief Infant-Toddler Social and Emotional Assessment. Examiner's Manual*. Harcourt Assessment.

Bronfenbrenner, U. 1979. *The Ecology of Human Development: Experiments by Nature and design*. Cambridge, MA: Harvard University Press.

Burton, P., Phipps, S., and Curtise, L. 2002. "All in the Family: A Simultaneous Model of Parenting Style and Child Conduct." *The American Economic Review* 92 (2): 368-372.

Campbell, F., Conti, G., Heckman, J.J., Moon, S.H., Pinto, R., Pungello, E., et al. 2014. "Early Childhood Investments Substantially Boost Adult Health". *Science* 343, 1478-1485.

Carneiro, P., and Heckman, J. 2003. "Human Capital Policy, in Inequality in America: What Role for Human Capital Policies." *Inequality in America: What Role for Human Capital Policies*.

Census. 2006. Retrieved from http://www.cso.ie/en/census/census2006reports/.

Cobb-Clarke, D., Salamanca, N., and Zhu, A. 2016. "Parenting Style as an Investment in Human Development." *IZA DP* No. 9686.

Connell, A., Bullock, B. M., Dishion, T. J., Shaw, D., Wilson, M., and Gardner, F. 2008. "Family Intervention Effects on Co-occurring Early Childhood Behavioral and Emotional Problems: A Latent Transition Analysis Approach." *Journal of Abnormal Child Psychology* 36 (8): 1211–1225.

Conti, G., Heckman, J.J, and Pinto, R. 2016. "The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behaviors." *The Economic Journal* 126: F28-F65.

Council of Economic Advisors. 2014. The Economics of Early Childhood Investments. https://www.whitehouse.gov/sites/default/files/docs/early_childhood_report1.pdf

Cunha, F. 2015. "Subjective Rationality, Parenting Styles, and Investments in Children." In Paul R. Amato, Alan Booth, Susan M. McHale, Jennifer Van Hook (Eds.) *Families in an Era of Increasing Inequality* (pp. 83-94). Cham: Springer.

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

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

Cunha, F., Heckman, J.J., Lochner, L.J., and Masterov, D.V. 2006. "Interpreting the Evidence on Life Cycle Skill Formation" in *Handbook of the Economics of Education*, eds Hanushek EA, Welch F. North–Holland, Amsterdam, 697–812.

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

Del Boca, D., Flinn, C., and Wiswall, M. 2014. "Household Choices and Child Development." *The Review of Economic Studies* 81, 137–185.

Del Bono, E., Francesconi, M., Kelly, Y., and Sacker, A. 2016. "Early Maternal Time Investment and Early Child Outcomes." *Economic Journal* 126: 96-135.

Deming, D., 2009. "Early Childhood Intervention and Life-cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1 (3): 111–134.

Doepke, M., and Zilibotti, F. 2014. "Parenting with Style: Altruism and Paternalism in Intergenerational Preference Transmission." NBER Working Paper, No. 20214.

Dooley, M., and Stewart, J. 2007. "Family Income, Parenting Styles and Child Behavioural–Emotional Outcomes." *Health Economics* 16 (2): 145-162.

Doyle, O. 2013. "Breaking the Cycle of Deprivation: An Experimental Evaluation of an Early Childhood Intervention." *Journal of the Statistical and Social Inquiry Society of Ireland* Vol. XLI, 92-111.

Doyle, O., Delaney, L., O'Farrelly, C., Fitzpatrick, N., and Daly, M. 2017b. "Can Early Intervention Policies Improve Well-being? Evidence from a randomized controlled trial." *PLoS ONE* 12 (1): e0169829.

Doyle, O., Fitzpatrick, N., Rawdon, C., and Lovett J. 2015. "Early Intervention and Child Health: Evidence from a Dublin-based Trial". *Economics and Human Biology* 19: 224-245.

Doyle, O., Harmon, C., Heckman, J., and Tremblay R. 2009. "Investing in Early Human Development: Timing and Economic Efficiency". *Economics and Human Biology* 7 (1): 1-6.

Doyle, O., Harmon, C., Heckman, J., Logue, C., and Moon. S. 2017a "Measuring Investment in Human Capital Formation: An Experimental Analysis of Early Life Outcomes." *Labour Economics* 45: 40-58.

Doyle, O., and Hickey C. 2013. "The Challenges of Contamination in Evaluations of Childhood Interventions." *Evaluation* 19: 180-191.

Doyle, O., McEntee, L., and McNamara, K. 2012. "Skills, Capabilities and Inequality at School Entry in a Disadvantaged Community." *European Journal of the Psychology of Education* 27: 133-154.

Doyle, O., McGlanaghy, E., Palamaro Munsell, E., and McAuliffe, F. 2014. "Home Based Educational Intervention to Improve Perinatal Outcomes for a Disadvantaged Community: A Randomised Control Trial". *European Journal of Obstetrics and Gynaecology* 180: 162-167.

Doyle, O., and *PFL* Evaluation Team. 2010. *Assessing the Impact of Preparing for Life: Baseline Report.* Report to Preparing for Life Programme. Atlantic Philanthropies & Department of Children and Youth Affairs.

Doyle, O., and *PFL* Evaluation Team. 2015. *Assessing the Impact of Preparing for Life at 48 Months.* Report to Preparing for Life Programme. Atlantic Philanthropies & Department of Children and Youth Affairs.

Drazen, S. M., and Haust, M. 1993. "Raising Readiness in Low-income Children by Parent Education." Paper presented at the annual meeting of the American Psychological Association.

Duflo, E., Glennerster, R., and Kremer, M. 2008. *Using Randomization in Development Economics Research: A Toolkit. Handbook of Development Economics*, Elsevier.

DuMont, K.A., Mitchell-Herzfeld, C., Greene, R., Lee, E., Lowenfels, A., Rodriguez, M., and Dorabawila, V. 2008. "Healthy Families New York Randomized Trial: Effects on early child abuse and neglect." *Child Abuse and Neglect* 32: 295-315.

Eble, A., Boone, P., Elbourne, D. 2016. "On Minimizing the Risk of Bias in Randomized Controlled Trials in Economics". The World Bank Economic Review, 0: 1-12.

Eckenrode, J., Campa, M., Luckey, D. W., Henderson, C. R., Cole, R., Kitzman, H., . . . & Olds, D. 2010. "Long Term Effects of Prenatal and Infancy Nurse Home Visitation on the Life Course of Youths: 19-year Follow-up of a Randomized-Controlled Trial." *JAMA Pediatrics* 164: 9-15.

Edwards, C. P., Sheridan, S. M. D., and Knoche, L. 2010. "Parent-child Relationships in Early Learning." In E. Baker, P. Peterson, & B. McGaw (Eds.), *International Encyclopedia of Education* (pp. 438-443). Oxford, England: Elsevier.

Elliott, C., Smith, P., and McCulloch, K. 1997. *British Ability Scales II.* London: NFER-Nelson.

Farah, M. J., Betancourt, L., Shera, D. M., Savage, J. H., Giannetta, J. M., Brodsky, N. L., … and Hurt, H. 2008. "Environmental Stimulation, Parental Nurturance and Cognitive Development in Humans." *Developmental Science* 11 (5): 793-801.

Fergusson, D. M., Grant, H., Horwood, L. J., and Ridder, E. M. 2005. "Randomized Trial of the Early Start Program of Home Visitation." *Pediatrics* 116 (6): e803-e809.

Fernald, A., Marchman, V. A., and Weisleder, A. 2013. "SES Differences in Language Processing Skill and Vocabulary are Evident at 18 Months." *Developmental Science* 16: 234–248.

Filene, J. H., Kaminski, J. W., Valle, L. A., and Cachat, P. 2013. "Components Associated with Home Visiting Program Outcomes: A Meta-analysis. *Pediatrics 132*(Supplement): S100–S109.

Fiorini, M., and Keane, M. 2014. "How the Allocation of Children's Time Affects Cognitive and Non-cognitive Development." *Journal of Labor Economics* 32 (4): 787-836.

Freedman, D., and Lane, D. 1983. "A Nonstochastic Interpretation of Reported Significance Levels." *Journal of Business Economics and Statistics* 1 (4), 292–298.

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., Chang, S.M. and Grantham-McGregor, S. 2014. "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica." *Science* 344 (6187): 998-1001.

Gomby, D. S. 2005. *Home Visitation in 2005: Outcomes for Children and Parents* (Vol. 7). Invest in Kids Working Paper No. 7. Committee for Economic Development: Invest in Kids Working Group.

Gomby, D. S., Culross, P. L., and Behrman, R. E. 1999. "Home Visiting: Recent Program Evaluations: Analysis and Recommendations." *The Future of Children* 9 (1): 4.

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

Goodman, R. 1997. "The Strengths and Difficulties Questionnaire: A Research Note." *Journal of Child Psychology and Psychiatry* 38 (5): 581-586.

Guttentag, C.L., Landry, S.H., Williams, J.M., Baggett, K.M., Noria, C.W., Borkowski, J.G., et al. 2014. ""My Baby & Me": Effects of an Early, Comprehensive Parenting Intervention on At-risk Mothers and Their Children." *Developmental Psychology* 50: 1482-96.

Halfon, N., Shulman, E., and Hochstein, M. 2001. "Brain Development in Early Childhood." In: Halfon, N., Shulman, E., Hochstein, M. (Eds.), *Building Community Systems for Young Children.* UCLA Center for Healthier Children, Families and Communities.

Hayes, A. 1996. "Permutation Test is Not Distribution-free: Testing h0 :) = 0." *Psychological Methods* 1: 184–198.

Heckman, J.J. 2006. "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science* 312 (5782): 1900-1902.

Heckman, J. J. and Kautz T. 2014. "Fostering and Measuring Skills Interventions that Improve Character and Cognition." In J. J. Heckman, J. E. Humphries, and T. Kautz (Eds.), The GED Myth: Education, Achievement Tests, and the Role of Character in American Life, Chapter 9. Chicago, IL: University of Chicago Press.

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

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

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

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

Jones Harden, B., Chazan-Cohen, R., Raikes, H., and Vogel, C. 2012. "Early Head Start Home Visitation: The Role of Implementation in Bolstering Program Benefits." *Journal of Community Psychology* 40(4): 438–455.

Karoly, L. A., Kilburn M. R, and Cannon, J. S. 2005. *Early Childhood Interventions: Proven Results, Future Promise*. Santa Monica, CA: RAND Corporation.

Keller, B. 2012. "Detecting Treatment Effects with Small Samples: The Power of Some Tests Under the Randomization Model." *Psychometrika* 77: 324-338.

Knudsen, E.I., Heckman J.J., Cameron J. L, and Shonkoff, J. P. 2006. "Economic, neurobiological, and behavioral perspectives on building America's future workforce." *Proceedings of the National Academy of Science USA* 103 (27): 10155–10162.

Landsverk, J., Carrilio, T., Connelly, C. D., Ganger, W., Slymen, D., Newton, R., … Jones, C. 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.

Lareau, A. 2011. *Unequal Childhoods: Class, Race, and Family Life* (2 ed.). Berkeley, CA: University of California Press.

Lecerf, M. 2016. "*Child Poverty in the European Union The Crisis and its Aftermath*". EPRS | European Parliamentary Research Service. European Union, 2016.

Ludbrook, J., and Dudley, H. 1998. "Why Permutation Tests are Superior to t and F Tests in Biomedical Research." *American Statistician* 52 (2): 127-132.

Ludwig, J., and Phillips, D. 2007. "The Benefits and Costs of Head Start." NBER Working Paper 12973.

Madden, J., O'Hara, J., and Levenstein, P. 1984. "Home Again: Effects of the Mother-child Home Program on Mother and Child." *Child Development* 55 (2): 636.

Martins, L., and Veiga, P. (2010). "Do Inequalities in Parents' Education Play an Important Role in PISA Students' Mathematics Achievement Test Score Disparities?". *Economics of Education Review* 29 (6): 1016-1033.

Maternal and Child Health Bureau. 2016. "Demonstrating Improvement in the Maternal, Infant, and Early Childhood Home Visiting Program: A Report to Congress". Health Resources and Services Administration (HRSA) and the Administration for Children and Families (ACF).

McCambridge, J., Witton J., and Elbourne D. R. 2014. "Systematic Review of the Hawthorne Effective: New Concepts are Needed to Study Research Participation Effects." *Journal of Clinical Epidemiology* 67 (3): 267–77.

Mewhort, D.J.K. 2005. "A Comparison of the Randomization Test with the F test when Error is Skewed." *Behavior Research Methods* 37: 426–435.

Michael, R. T., and Becker, G. S. 1973. "On the New Theory of Consumer Behavior." *The Swedish Journal of Economics* 75 (4): 378-396.

Miller, E. B., Farkas, G., Vandell, D. L., and Duncan, G. J. 2014. "Do the Effects of Head Start Vary by Parental Preacademic Stimulation?" *Child Development* 85: 1385–1400.

Miller, S., Maguire, L.K., and Macdonald, G. 2011. "Home-based Child Development Interventions for Preschool Children from Socially Disadvantaged Families." *Cochrane Database of Systematic Reviews* 12, CD008131.

Najman, J.M., Williams, G.M., Nikles, J., Spence, S., Bor, W., O'Callaghan, M., LeBrocque, R., Andersen, M.J., and Shuttlewood, G.J. 2001. "Bias Influencing Maternal Reports of Child Behaviour and Emotional State." *Social Psychiatry and Psychiatric Epidemiology* 36 (4): 186–194.

Necoechea, D. M. 2007. Children At-risk for Poor School Readiness: The Effect of an Early Intervention Home Visiting Program on Children and Parents. *Dissertations Abstracts International Section A: Humanities and Social Sciences,* 68 (6-A), 2311. (Dissertation Abstract: 2007-99230-512)

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F, and Turkheimer, E. 2012. "Intelligence: New Findings and Theoretical Developments." *American Psychologist* 67(2): 130–159.

O'Sullivan, A., Fitzpatrick, N., and Doyle, O. 2017. "Effects of Dietary Recommendations During Early Childhood on Cognitive Functioning: A Randomized Controlled Trial." *Public Health Nutrition* 20 (1): 154-164.

OECD. 2016. *Enhancing Child Well-being to Promote Inclusive Growth*. OECD Publishing, Paris.

Office of the Minister for Children and Youth Affairs. 2008. *Forum on Prevention and Early Intervention for Children and Youth*. The Stationery Office, Dublin.

Olds, D. L., Kitzman, H., Cole, R., Robinson, J., Sidora, K., Luckey, D. W., ... Holmberg, J. 2004. "Effects of Nurse Home-visiting on Maternal Life Course and Child Development: Age 6 Follow-up Results of a Randomized Trial." *Pediatrics* 114 (6): 1550-1559.

Olds, D. L., Henderson, C. R., and Kitzman, H. 1994. "Does Prenatal and Infancy Nurse Home Visitation have Enduring Effects on Qualities of Parental Caregiving and Child Health at 25 to 50 Months of Life?" *Pediatrics* 93 (1): 89-98.

Olds, D. L., Henderson, C. R., Jr., Kitzman, H. J., Eckenrode, J. J., Cole, R., and Tatelbaum, C. 1999. "Prenatal and Infancy Home Visitation by Nurses: Recent Findings." *The Future of Our Children* 9 (1): 44–65.

Peacock, S., Konrad, S., Watson, E., Nickel, D., and Muhajarine, N. 2013. "Effectiveness of Home Visiting Programs on Child Outcomes: A Systematic Review." *BMC Public Health* 13 (1): 17.

Ramey, S.L., and Ramey, C.T. 1992. "Early Educational Intervention with Disadvantaged Children - To What Effect?" *Applied and Preventive Psychology* I, 131-140.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. 1994. "Estimation of Regression Coefficients when Some Regressors are Not Always Observed." *Journal of the American Statistical Association* 89 (427): 846-866.

Roggman, L. A., Boyce, L. K., and Cook, G. A. 2009. "Keeping Kids on Track: Impacts of a Parenting-focused Early Head Start Program on Attachment Security and Cognitive Development." *Early Education & Development* 20 (6): 920–941.

Roggman L.R,. Cook, G.A., Peterson, C.A., and Raikes, H.H. 2008. "Who Drops Out of Early Head Start Home Visiting Programs?" *Early Education And Development* 19 (4).

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

Sanders, M. R. 2012. "Development, Evaluation, and Multinational Dissemination of the Triple P-Positive Parenting Program." *Annual Review of Clinical Psychology* 8(1): 345–379.

Sanders, M.R., Kirby, J.N., Tellegen, C.L., and Day, J.J. 2014. "The Triple P-Positive Parenting Program: A Systematic Review and Meta-analysis of a Multi-level System of Parenting Support." *Clinical Psychology Review* 34: 337-357.

Sanders, M.R., Markie-Dadds. C., and Turner. K. 2003. "Theoretical, Scientific and Clinical Foundations of the Triple P-Positive Parenting Program: A Population Approach to the Promotion of Parenting Competence." *Parenting Research and Practice Monograph* 1: 1-21.

Sandner, M., and Jungmann, T. 2016. "How Much Can We Trust Maternal Ratings of Early Child Development in Disadvantaged Samples?" *Economics Letters* 141: 73-76.

Sandner, M., and Jungmann, T. 2017. "Gender-specific Effects of Early Childhood Intervention: Evidence from a Randomized Controlled Trial." *Labour Economics* 45: 59-78.

Schwarz G. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6: 461-464.

Shaw, D. S., Connell, A., Dishion, T. J., Wilson, M. N., and Gardner, F. 2009. "Improvements in Maternal Depression as a Mediator of Intervention Effects on Early Childhood Problem Behavior." *Development and Psychopathology* 21 (2): 417.

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

Stöber, J. (2001). "The Social Desirability Scale – 17 (SDS-17): Convergent Validity, Discriminant Validity, and Relationship with Age." *European Journal of Psychological Assessment* 17: 222-232.

Sweet, M. A., and Appelbaum, M. I. 2004. "Is Home Visiting an Effective Strategy? A Meta-Analytic Review of Home Visiting Programs for Families with Young Children." *Child Development* 75: 1435-1456.

The Lancet 2016. "The Lancet Early Childhood Development Series: Advancing Early Childhood Development: from Science to Scale." *The Lancet* 389: 10064.

Thompson, R. A., and Nelson, C. A. 2001. "Developmental Science and The Media: Early Brain Development." *American Psychologist* 56 (1): 5–15.

Threshold. 1987. *Policy Consequences: A Study of the £5,000 Surrender Grant in the Dublin Housing Area*. Dublin: Threshold.

Todd, P.E. and Wolpin, K.I. 2007. "The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps." *Journal of Human Capital* 1 (1): 91-136.

Walker, S., Chang, S., Vera-Hernandez, M., and Grantham-McGregor, S. 2011. "Early Childhood Stimulation Benefits Adult Competence and Reduces Violent Behavior." *Pediatrics* 127 (5): 849{857).

Washington State Institute for Public Policy. 2016. *Benefit-Cost Results – Public Health and Prevention*. http://www.wsipp.wa.gov/BenefitCost

Weaver, I.C.G., Cervoni, N., Champagne, F.A., D'Alessio, A.C., Sharma, S., Seckl, J.R., et al. 2004. "Epigenetic Programming by Maternal Behavior." *Nature Neuroscience* 7: 847–854.

Wesley, P. W., Buysse, V., and Tyndall, S. 1997. "Family and Professional Perspectives on Early Intervention: An Exploration Using Focus Groups." *Topics in Early Childhood Special Education* 17 (4): 435-456.

Williams, J., Greene, S., McNally, S., Murray, A., and Quail, A. 2010. *Growing up in Ireland National Longitudinal Study of Children. The Infants and their Families*. The Stationery Office, Ireland.

Figure 1 *Timing of PFL treatments*

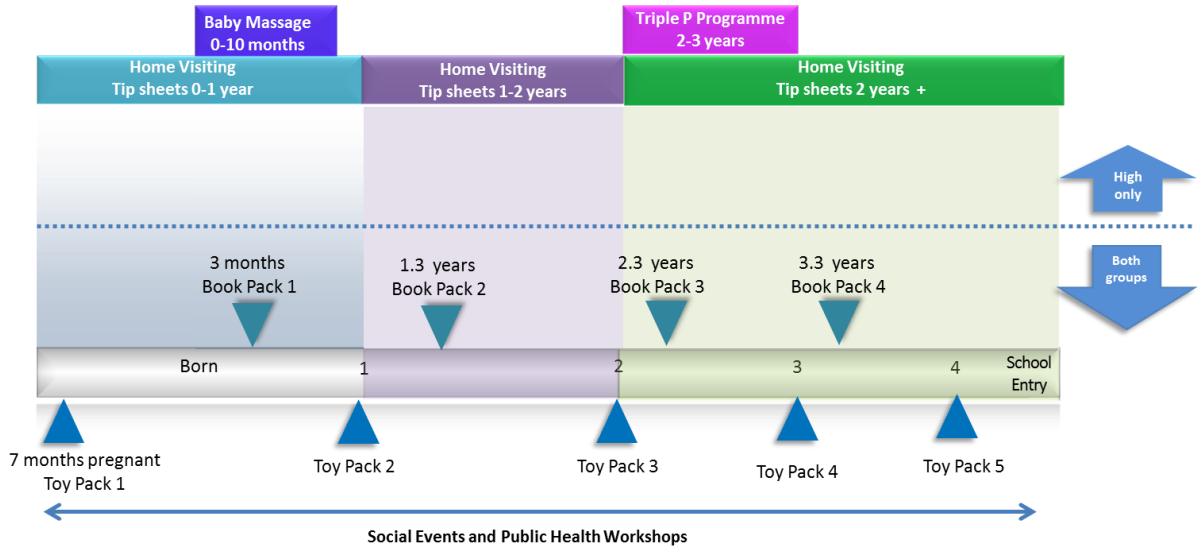

Figure 2 *Participant flow*

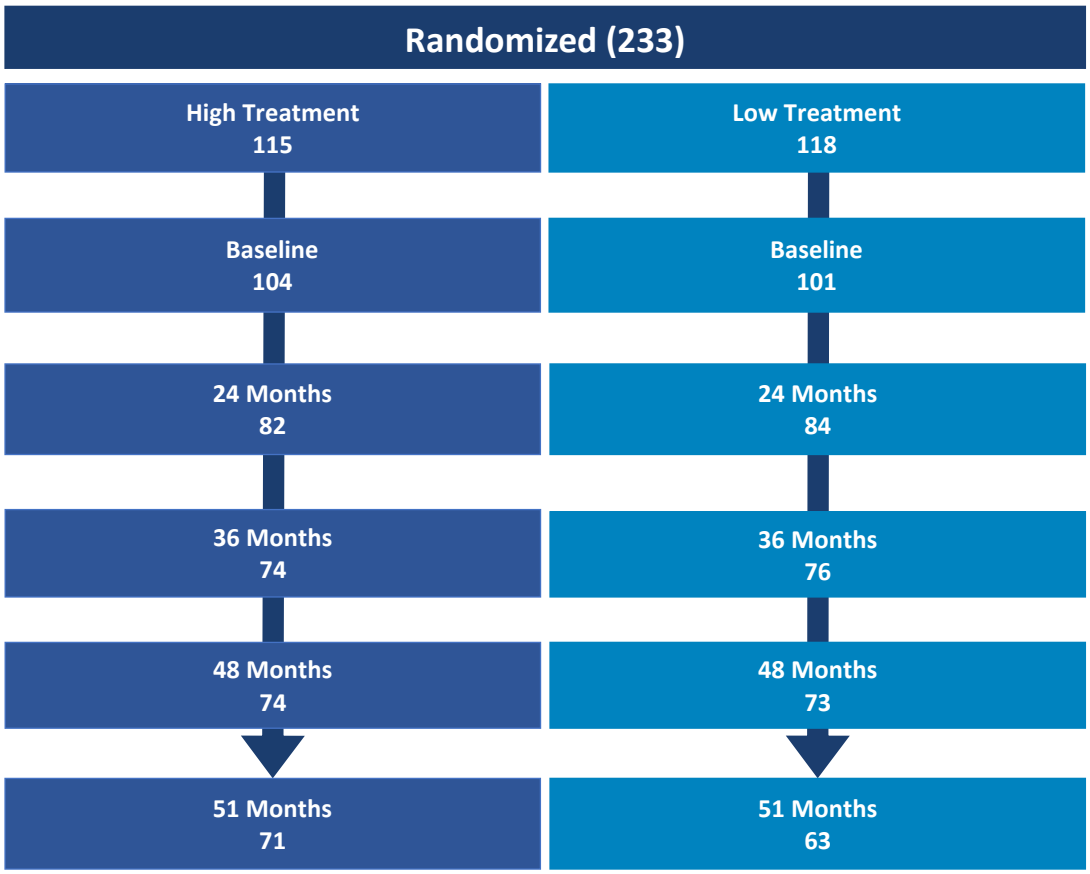

Figure 3 *Distribution of BAS GCA cognitive scores*

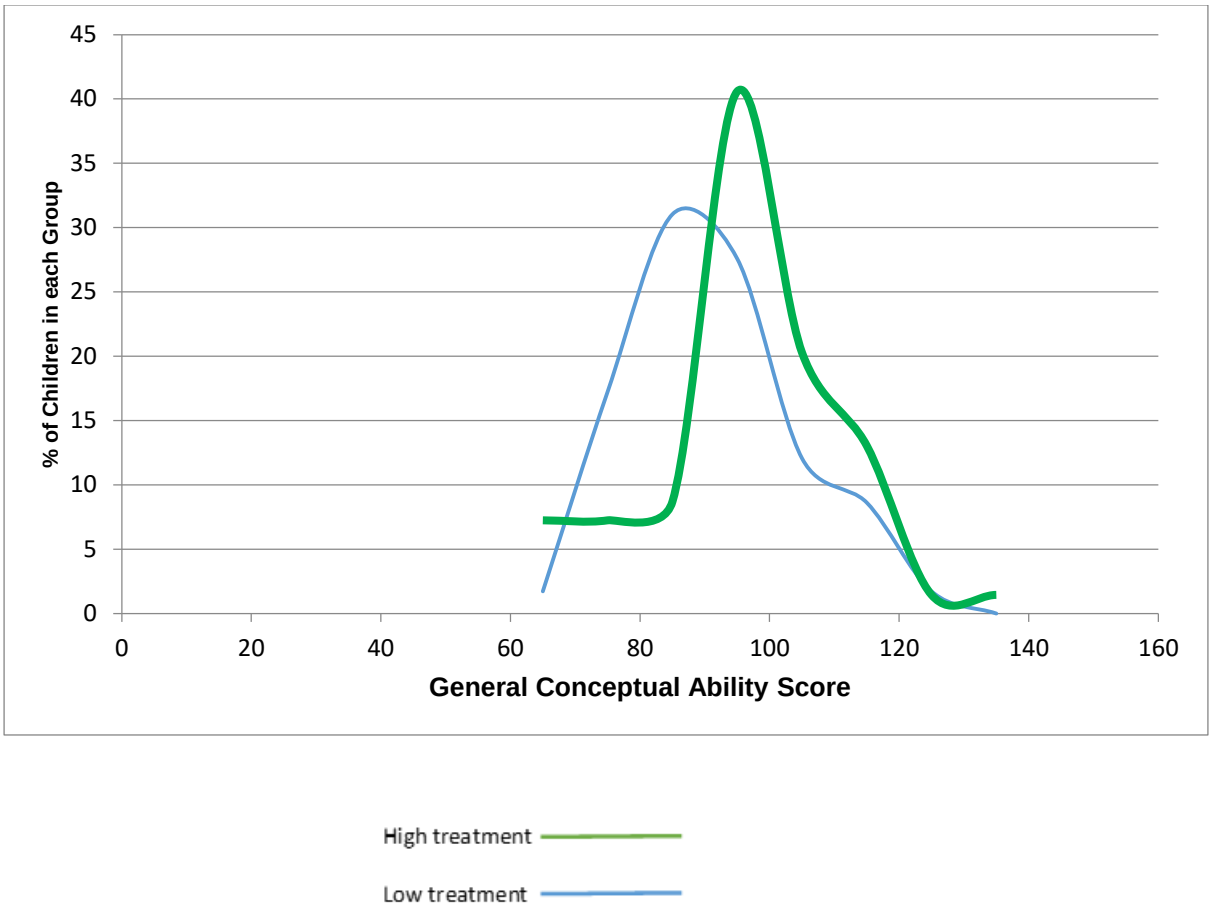

Table 1 *Baseline comparison of high and low treatment groups: Estimation samples*

	24 Month Sample			36 Month Sample			48 Month Sample			51 Month Sample		
	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	p^1	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	p^1	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	p^1	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	p^1
Age	25.85 (5.86)	25.60 (6.25)	0.790	25.64 (5.69)	25.96 (5.98)	0.744	26.33 (5.85)	25.96 (5.92)	0.707	26.49 (5.86)	26.13 (5.89)	0.727
Married	0.16 (0.37)	0.18 (0.39)	0.794	0.15 (0.36)	0.16 (0.37)	0.734	0.16 (0.37)	0.15 (0.36)	0.946	0.17 (0.38)	0.16 (0.37)	0.964
Irish	0.99 (0.11)	0.99 (0.11)	0.747	0.99 (0.12)	0.99 (0.12)	0.826	0.99 (0.12)	0.99 (0.12)	0.787	0.99 (0.12)	0.98 (0.13)	0.822
First time mother	0.52 (0.50)	0.46 (0.50)	0.403	0.55 (0.50)	0.45 (0.50)	0.201	0.49 (0.50)	0.43 (0.50)	0.412	0.49 (0.50)	0.42 (0.50)	0.432
Low education (left $\leq$ age 16)	0.29 (0.46)	0.36 (0.48)	0.330	0.29 (0.46)	0.32 (0.47)	0.666	0.32 (0.47)	0.33 (0.47)	0.792	0.30 (0.46)	0.34 (0.48)	0.642
IQ[a]	83.15 (12.36)	81.31 (12.46)	0.346	83.62 (12.25)	81.93 (12.20)	0.400	83.63 (12.49)	80.93 (12.90)	0.202	83.99 (12.04)	80.71 (13.17)	0.141
Equalized household income (€)	241.18 (106.97)	264.62 (150.47)	0.298	243.56 (110.77)	272.22 (156.00)	0.244	233.38 (107.96)	272.32 (155.78)	0.111	234.21 (103.74)	267.09 (149.34)	0.186
Employed	0.43 (0.50)	0.41 (0.49)	0.823	0.43 (0.50)	0.45 (0.50)	0.815	0.44 (0.50)	0.42 (0.50)	0.786	0.46 (0.50)	0.42 (0.50)	0.666
Resides in public housing	0.54 (0.50)	0.54 (0.50)	0.942	0.53 (0.50)	0.54 (0.50)	0.936	0.53 (0.50)	0.53 (0.50)	0.902	0.54 (0.50)	0.56 (0.50)	0.796
Prior physical health condition	0.76 (0.43)	0.64 (0.48)	0.111	0.76 (0.43)	0.64 (0.48)	**0.088**	0.75 (0.43)	0.65 (0.48)	0.176	0.76 (0.43)	0.66 (0.48)	0.218
Prior mental health condition	0.27 (0.45)	0.25 (0.44)	0.775	0.29 (0.46)	0.28 (0.45)	0.930	0.26 (0.44)	0.28 (0.45)	0.787	0.27 (0.45)	0.31 (0.46)	0.611
Body Mass Index	24.32 (5.03)	24.11 (4.73)	0.802	24.19 (5.14)	24.22 (4.92)	0.970	24.40 (5.00)	24.51 (4.90)	0.901	24.35 (4.95)	24.53 (5.12)	0.853
Smoked during pregnancy	0.50 (0.50)	0.46 (0.50)	0.591	0.51 (0.50)	0.47 (0.50)	0.682	0.49 (0.50)	0.47 (0.50)	0.805	0.50 (0.50)	0.48 (0.50)	0.795
Alcohol during pregnancy	0.30 (0.46)	0.29 (0.46)	0.790	0.29 (0.46)	0.30 (0.46)	0.984	0.33 (0.47)	0.31 (0.46)	0.839	0.33 (0.47)	0.35 (0.48)	0.780
Drugs during pregnancy	0.01 (0.120	0.01 (0.12)	0.721	0.02 (0.12)	0.01 (0.12)	0.853	0.02 (0.13)	0.02 (0.12)	0.812	0.02 (0.13)	0.02 (0.13)	0.800
Pearlin self-efficacy score	2.92 (0.50)	3.02 (0.51)	0.214	2.93 (0.49)	3.05 (0.54)	0.148	2.93 (0.50)	3.03 (0.53)	0.218	2.94 (0.49)	3.01 (0.52)	0.473
Rosenberg self-esteem score	13.05 (2.65)	12.80 (2.86)	0.583	13.06 (2.61)	12.76 (2.98)	0.540	12.91 (2.66)	12.86 (2.92)	0.915	13.00 (2.70)	12.56 (2.92)	0.384
TIPI Emotional Stability	3.81 (1.62)	4.13 (1.55)	0.195	3.87 (1.63)	4.01 (1.58)	0.602	3.99 (1.64)	4.09 (1.65)	0.696	3.89 (1.64)	4.04 (1.69)	0.591
TIPI Conscientiousness	5.49	5.47	0.903	5.46	5.49	0.919	5.47	5.52	0.818	5.41	5.43	0.916

	(1.28)	(1.30)		(1.31)	(1.30)		(1.29)	(1.31)		(1.30)	(1.33)	
TIPI Openness to Experience	4.97	5.12	0.442	5.00	5.27	0.182	4.96	5.26	0.144	4.96	5.20	0.241
	(1.23)	(1.26)		(1.22)	(1.20)		(1.25)	(1.22)		(1.24)	(1.16)	
TIPI Agreeableness	5.71	5.80	0.650	5.71	5.84	0.510	5.68	5.82	0.474	5.68	5.87	0.363
	(1.16)	(1.21)		(1.17)	(1.21)		(1.15)	(1.19)		(1.17)	(1.22)	
TIPI Extraversion	5.15	5.19	0.846	5.19	5.27	0.721	5.18	5.23	0.821	5.17	5.12	0.840
	(1.29)	(1.41)		(1.19)	(1.34)		(1.21)	(1.39)		(1.24)	(1.42)	
N		165		149				145			132	

Notes: All baseline measures were assessed during pregnancy prior to treatment delivery except for the measure of IQ which was assessed at 3 months postpartum using Weschler Abbreviated Scale of Intelligence (WASI). Baseline data are missing for two participants who participated in later waves but did not complete the baseline assessment. [1] two-tailed *p*-value calculated from permutation tests with 100,000 replications.

Table 2 *Cognitive skills results*

	N (HIGH/LOW)	*IPW* M_{HIGH} (*SD*)	*IPW* M_{LOW} (*SD*)	*IPW Treatment Effect*	*IPW Effect Size*	p^1	p^2	p^3	p^4
DP3 Scores									
24 Months	166 (82/84)	101.64 (13.61)	98.16 (15.62)	3.48	0.22	**0.034**	**0.034**	**0.019**	**0.019**
36 Months	150 (74/76)	102.64 (14.90)	96.64 (14.24)	6.00	0.42	**0.006**	**0.013**	**0.006**	**0.012**
48 Months	147 (74/73)	102.23 (13.19)	97.32 (15.42)	4.91	0.32	**0.017**	**0.025**	**0.008**	**0.017**
DP3 Cutoffs - Above Average %									
24 Months	166 (82/84)	0.66 (0.48)	0.54 (0.50)	0.12	0.24	**0.031**	**0.031**	**0.031**	**0.031**
36 Months	150 (74/76)	0.53 (0.50)	0.36 (0.48)	0.17	0.36	**0.008**	**0.023**	**0.008**	**0.023**
48 Months	147 (74/73)	0.34 (0.48)	0.19 (0.40)	0.15	0.37	**0.012**	**0.022**	**0.012**	**0.022**
ASQ Communication Scores									
24 Months	166 (82/84)	100.41 (15.05)	100.59 (14.44)	0.17	-0.01	0.345	0.345	0.381	0.381
36 Months	150 (75/75)	101.38 (14.17)	97.30 (16.40)	4.08	0.25	**0.073**	**0.091**	**0.091**	0.171
48 Months	147 (74/73)	101.10 (13.20)	99.63 (14.94)	1.47	0.10	0.104	0.202	0.137	0.232
ASQ Communication Cutoffs – Below Average %									
24 Months	166 (82/84)	0.10 (0.30)	0.07 (0.25)	0.03	0.13	0.633	0.633	0.684	0.684
36 Months	150 (75/75)	0.04 (0.21)	0.07 (0.25)	0.02	0.09	0.148	0.305	**0.076**	0.171
48 Months	147 (74/73)	0.04 (0.20)	0.05 (0.22)	0.01	0.03	0.238	0.395	0.186	0.319
ASQ Problem Solving Scores									
24 Months	166 (82/84)	101.67 (15.19)	98.39 (14.83)	3.28	0.22	**0.080**	0.137	**0.085**	0.118
36 Months	147 (73/74)	102.28 (13.58)	96.77 (15.14)	5.51	0.36	**0.021**	**0.041**	**0.018**	**0.032**
48 Months	147 (74/73)	100.55 (14.52)	100.04 (16.69)	0.50	0.03	0.303	0.303	0.227	0.227

ASQ Problem Solving Cutoffs - Below Average %									
24 Months	166 (82/84)	0.07 (0.25)	0.14 (0.35)	0.07	0.21	**0.066**	0.173	**0.094**	0.143
36 Months	147 (73/74)	0.11 (0.31)	0.19 (0.39)	0.08	0.20	**0.083**	0.163	**0.040**	**0.098**
48 Months	147 (74/73)	0.05 (0.22)	0.06 (0.25)	0.02	0.06	0.296	0.296	0.408	0.408
BAS Scores @ 51 Months									
General Conceptual Ability	128 (69/59)	104.87 (15.18)	94.58 (13.30)	10.29	0.77	**<0.001**	**0.001**	**0.001**	**0.006**
Spatial Ability	129 (69/60)	104.48 (14.58)	95.91 (13.11)	8.57	0.65	**<0.001**	**0.001**	**0.002**	**0.006**
Pictorial Reasoning Ability	132 (71/61)	103.53 (15.31)	96.33 (12.85)	7.20	0.56	**0.001**	**0.001**	**0.011**	**0.028**
Language Ability	134 (71/63)	104.16 (15.67)	94.21 (14.77)	9.95	0.67	**0.002**	**0.002**	**0.022**	**0.022**
BAS Cutoffs - Below Average @ 51 Months %									
General Conceptual Ability	128 (69/59)	0.20 (0.40)	0.60 (0.49)	0.40	0.81	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Spatial Ability	129 (69/60)	0.31 (0.47)	0.60 (0.49)	0.29	0.58	**0.001**	**0.004**	**0.002**	**0.005**
Pictorial Reasoning Ability	132 (71/61)	0.29 (0.46)	0.46 (0.50)	0.16	0.33	**0.043**	**0.043**	**0.065**	**0.097**
Language Ability	134 (71/63)	0.26 (0.44)	0.46 (0.50)	0.20	0.40	**0.015**	**0.031**	0.111	0.111
BAS Cutoffs - Above Average @ 51 Months %									
General Conceptual Ability	128 (69/59)	0.25 (0.44)	0.08 (0.27)	0.17	0.64	**0.016**	**0.031**	**0.098**	0.222
Spatial Ability	129 (69/60)	0.14 (0.35)	0.09 (0.29)	0.05	0.16	0.138	0.138	0.166	0.166
Pictorial Reasoning Ability	132 (71/61)	0.17 (0.38)	0.09 (0.29)	0.08	0.27	**0.057**	0.100	**0.095**	0.198
Language Ability	134 (71/63)	0.25 (0.43)	0.08 (0.27)	0.17	0.62	**0.016**	**0.018**	**0.039**	**0.087**

Notes: 'N' indicates the sample size. 'IPW M' indicates the IPW-adjusted mean. 'IPW SD' indicates the IPW-adjusted standard deviation. [1] one-tailed (right-sided) conditional *p*-value from individual IPW-adjusted permutation test with 100,000 replications. [2] one-tailed (right-sided) conditional *p*-value from IPW-adjusted stepdown permutation test with 100,000 replications. [3] one-tailed (right-sided) conditional *p*-value from individual permutation test with 100,000 replications. [4] one-tailed (right-sided) conditional *p*-value from stepdown permutation test with 100,000 replications. 'Treatment effect' is the difference in means between the high and low treatment group. 'Effect size' is the ratio of the treatment effect to the standard deviation of the low treatment group.

Table 3 *Socio-emotional and behavioral skills results*

	N (HIGH/LOW)	*IPW* M_{HIGH} (*SD*)	*IPW* M_{LOW} (*SD*)	*IPW Treatment Effect*	*IPW Effect Size*	p^1	p^2	p^3	p^4
CBCL Total Scores									
24 Months	164 (81/83)	98.74 (13.53)	101.81 (16.58)	3.06	0.18	0.108	0.108	0.172	0.258
36 Months	150 (74/76)	98.20 (13.50)	101.92 (15.60)	3.71	0.24	**0.064**	0.109	0.121	0.210
48 Months	146 (74/72)	100.42 (12.64)	105.55 (21.04)	5.13	0.24	0.139	0.184	0.324	0.324
CBCL Total Cutoffs %									
24 Months	164 (81/83)	0.00 (0.00)	0.09 (0.29)	0.09	0.32	**<0.001**	**<0.001**	**0.004**	**0.011**
36 Months	150 (74/76)	0.01 (0.11)	0.08 (0.27)	0.07	0.25	**0.015**	**0.015**	**0.026**	**0.042**
48 Months	146 (74/72)	0.02 (0.15)	0.17 (0.38)	0.15	0.39	**0.028**	**0.028**	**0.068**	**0.068**
CBCL Externalizing Scores									
24 Months	164 (81/83)	99.10 (13.44)	100.89 (16.26)	1.79	0.11	0.232	0.232	0.403	0.403
36 Months	150 (74/76)	98.32 (12.49)	101.76 (16.31)	3.44	0.21	**0.064**	0.119	0.122	0.240
48 Months	146 (74/72)	99.98 (13.12)	106.82 (22.13)	6.85	0.31	**0.097**	0.111	0.224	0.356
CBCL Externalizing Cutoffs %									
24 Months	164 (81/83)	0.00 (0.00)	0.04 (0.20)	0.04	0.21	**0.009**	**0.016**	**0.038**	**0.044**
36 Months	150 (74/76)	0.01 (0.11)	0.07 (0.25)	0.05	0.21	**0.021**	**0.021**	**0.030**	**0.030**
48 Months	146 (74/72)	0.00 (0.00)	0.16 (0.37)	0.16	0.43	**0.005**	**0.005**	**0.018**	**0.022**
CBCL Internalizing Scores									
24 Months	164 (81/83)	100.03 (14.78)	101.17 (15.68)	1.13	0.07	0.303	0.303	0.311	0.431
36 Months	150 (74/76)	98.26 (15.42)	101.37 (14.29)	3.11	0.22	0.132	0.263	0.157	0.242
48 Months	146 (74/72)	101.90 (13.69)	103.23 (17.57)	1.33	0.08	0.279	0.452	0.596	0.596
CBCL Internalizing Cutoffs %									

24 Months	164 (81/83)	0.02 (0.15)	0.09 (0.29)	0.07	0.24	**0.041**	**0.067**	0.112	0.193
36 Months	150 (74/76)	0.07 (0.26)	0.07 (0.26)	0.00	0.00	0.513	0.513	0.435	0.435
48 Months	146 (74/72)	0.03 (0.18)	0.20 (0.40)	0.17	0.41	**0.023**	**0.025**	**0.044**	0.114
BITSEA Competency Score									
24 Months	166 (82/84)	99.26 (15.29)	100.12 (14.35)	0.86	0.06	0.541	0.541	0.563	0.563
36 Months	151 (75/76)	100.53 (14.93)	98.57 (14.70)	1.97	0.13	0.175	0.254	0.126	0.198
BITSEA Competency Cutoffs %									
24 Months	166 (82/84)	0.11 (0.32)	0.09 (0.29)	0.02	0.07	0.310	0.433	0.357	0.476
36 Months	151 (75/76)	0.13 (0.34)	0.17 (0.38)	0.04	0.10	0.621	0.621	0.694	0.694
BITSEA Problems Score									
24 Months	166 (82/84)	98.61 (11.72)	101.88 (17.49)	3.27	0.19	**0.054**	**0.093**	**0.039**	**0.065**
36 Months	151 (75/76)	99.06 (12.52)	100.25 (16.81)	1.20	0.07	0.244	0.244	0.217	0.217
BITSEA Problems Cutoffs %									
24 Months	166 (82/84)	0.13 (0.34)	0.23 (0.43)	0.10	0.24	**0.056**	0.103	**0.038**	**0.073**
36 Months	151 (75/76)	0.15 (0.36)	0.19 (0.39)	0.03	0.09	0.335	0.335	0.231	0.231
SDQ Scores @ 48 Months									
Prosocial Behavior Score	147 (74/73)	101.44 (13.91)	95.32 (17.71)	6.13	0.35	**0.034**	**0.080**	0.122	0.197
Peer Problems	147 (74/73)	99.11 (14.22)	103.83 (19.35)	4.71	0.24	0.157	0.157	0.273	0.273
SDQ Cutoffs @ 48 Months %									
Prosocial Behavior Score	147 (74/73)	0.08 (0.27)	0.18 (0.39)	0.10	0.27	**0.091**	0.229	0.104	0.160
Peer Problems	147 (74/73)	0.09 (0.29)	0.16 (0.37)	0.07	0.20	0.255	0.255	0.449	0.449

Note: 'N' indicates the sample size. 'IPW M' indicates the IPW-adjusted mean. 'IPW SD' indicates the IPW-adjusted standard deviation. [1] one-tailed (right-sided) conditional *p*-value from individual IPW-adjusted permutation test with 100,000 replications. [2] one-tailed (right-sided) conditional *p*-value from IPW-adjusted stepdown permutation test with 100,000 replications. [3] one-tailed (right-sided) conditional *p*-value from individual permutation test with 100,000 replications. [4] one-tailed (right-sided) conditional *p*-value from stepdown permutation test with 100,000 replications. 'Treatment effect' is the difference in means between the high and low treatment group. 'Effect size' is the ratio of the treatment effect to the standard deviation of the low treatment group.

Table 4 *Test for heterogeneous effects - cognitive skills results*

	Gender[1]			**Parity Status**[2]		
	Treatment X Gender	*Treatment*	*Gender*	*Treatment X Parity*	*Treatment*	*Parity*
DP3 Scores						
24 Months	-0.489	4.473	4.595	-0.461	4.279	2.569
	(4.767)	(3.841)	(3.579)	(4.627)	(3.143)	(3.398)
36 Months	0.112	6.717	4.545	0.442	**6.629****	0.442
	(5.178)	(4.102)	(3.739)	(5.328)	**(3.128)**	(5.328)
48 Months	2.617	3.800	2.336	7.041	1.950	-2.599
	(5.625)	(4.725)	(4.774)	(5.541)	(3.148)	(4.714)
DP3 Cutoffs - Above Average %						
24 Months	-0.030	0.148	0.118	-0.282	0.133	0.132
	(0.164)	(0.125)	(0.118)	(0.162)	(0.118)	(0.115)
36 Months	**-0.305***	**0.390*****	**0.255****	0.034	**0.196***	0.023
	(0.169)	**(0.114)**	**(0.118)**	(0.177)	**(0.116)**	(0.126)
48 Months	0.023	0.146*	0.133	0.054	0.128	0.070
	(0.136)	(0.086)	(0.081)	(0.141)	(0.084)	(0.090)
ASQ Communication Scores						
24 Months	-0.156	0.927	**6.240***	0.052	0.943	-1.559
	(4.815)	(4.261)	**(3.520)**	(4.551)	(3.144)	(3.114)
36 Months	8.776	-0.509	-1.043	4.334	2.543	-3.936
	(6.111)	(4.865)	(5.087)	(6.166)	(3.392)	(5.127)
48 Months	-2.198	3.378	**9.552*****	5.804	-0.541	-3.056
	(4.591)	(3.985)	**(3.513)**	(4.617)	(2.632)	(3.541)
ASQ Communication Cutoffs - Below Average %						
24 Months	-0.015	0.025	-0.095	-0.023	0.028	0.011
	(0.092)	(0.083)	(0.066)	(0.085)	(0.061)	(0.054)
36 Months	0.025	-0.057	-0.114	-0.015	-0.034	-0.000
	(0.089)	(0.086)	(0.073)	(0.075)	(0.053)	(0.055)
48 Months	-0.004	-0.012	-0.082	0.014	-0.020	-0.006
	(0.071)	(0.069)	(0.056)	(0.065)	(0.047)	(0.048)
ASQ Problem Solving Scores						
24 Months	4.200	1.068	-1.026	6.482	0.316	**-7.503****
	(4.986)	(3.833)	(3.554)	(4.824)	(3.235)	**(3.278)**
36 Months	-1.876	**6.363***	-0.402	4.425	2.670	-4.392
	(4.978)	**(3.523)**	(3.724)	(4.866)	(3.890)	(3.489)
48 Months	3.378	-0.787	4.739	6.739	-1.978	-5.515
	(7.169)	(6.711)	(6.343)	(6.569)	(3.820)	(5.672)

ASQ Problem Solving Cutoffs -Below Average %						
24 Months	-0.001	-0.071	0.014	-0.095	-0.020	**0.141***
	(0.095)	(0.071)	(0.079)	(0.093)	(0.047)	**(0.076)**
36 Months	-0.016	-0.074	-0.013	0.039	-0.095	-0.023
	(0.124)	(0.097)	(0.100)	(0.118)	(0.088)	(0.091)
48 Months	0.028	-0.036	-0.074	-0.053	0.001	0.073
	(0.101)	(0.095)	(0.088)	(0.096)	(0.040)	(0.083)
BAS Scores @ 51 Months						
General Conceptual Ability	2.915	**9.001****	5.021	7.798	**6.669****	-4.392
	(5.828)	**(3.737)**	(4.187)	(5.997)	**(2.940)**	(4.646)
Spatial Ability	3.137	**7.134***	4.001	2.402	8.015	-3.640
	(5.350)	**(4.128)**	(3.891)	(5.334)	(3.069)	(3.979)
Pictorial Reasoning Ability	3.807	5.517	4.939	5.325	4.900	-1.845
	(5.198)	(3.899)	(3.574)	(5.163)	(3.309)	(3.657)
Language Ability	1.740	**9.126***	1.584	**15.415****	**1.900****	-5.405
	(6.652)	**(4.648)**	(5.070)	**(6.416)**	**(3.456)**	(5.200)
BAS Cutoffs - Below Average @ 51 Months %						
General Conceptual Ability	0.033	**-0.432*****	-0.234	-0.164	**-0.341*****	0.231
	(0.184)	**(0.133)**	(0.156)	(0.174)	**(0.105)**	(0.145)
Spatial Ability	-0.263	-0.145	-0.019	0.0201	**-0.312*****	0.040
	(0.207)	(0.167)	(0.170)	(0.212)	**(0.115)**	(0.178)
Pictorial Reasoning Ability	-0.091	-0.118	-0.035	-0.113	-0.121	0.136
	(0.209)	(0.155)	(0.174)	(0.214)	(0.116)	(0.180)
Language Ability	-0.071	-0.171	-0.126	**-0.392****	-0.007	0.190
	(0.204)	(0.151)	(0.173)	**(0.197)**	(0.125)	(0.165)
BAS Cutoffs - Above Average @ 51 Months %						
General Conceptual Ability	-0.008	**0.185***	**0.132****	**0.308****	0.008	-0.011
	(0.163)	**(0.104)**	**(0.057)**	**(0.142)**	(0.067)	(0.068)
Spatial Ability	-0.123	**0.123****	**0.152****	-0.138	0.129	0.010
	(0.103)	**(0.055)**	**(0.060)**	(0.115)	(0.090)	(0.074)
Pictorial Reasoning Ability	-0.008	**0.095****	**0.148****	0.030	0.070	0.031
	(0.115)	**(0.047)**	**(0.059)**	(0.125)	(0.074)	(0.076)
Language Ability	-0.124	**0.239****	0.059	**0.328****	0.014	-0.004
	(0.149)	**(0.113)**	(0.149)	**(0.132)**	(0.061)	(0.067)

Notes: [1]Estimated using IPW-adjusted OLS regressions including a gender by treatment status interaction term, gender (girl=1), and treatment status. [2] Estimated using IPW-adjusted OLS regressions including a parity by treatment status interaction term, parity status (firstborn=1), treatment status, and gender (not shown).

Table 5 *Test for heterogeneous effects – socio-emotional and behavioral skills results*

	Gender			**Parity Status**		
	Treatment X Gender	*Treatment*	*Gender*	*Treatment X Parity*	*Treatment*	*Parity*
CBCL Total Scores						
24 Months	2.901	-4.832	-1.987	**-8.538**	1.086	6.371
	(5.227)	(4.075)	(4.235)	**(4.963)**	(3.284)	(3.892)
36 Months	2.817	-5.444	-1.777	**-8.727**	1.036	**7.023**
	(5.258)	(4.207)	(4.100)	**(4.947)**	(3.515)	**(3.645)**
48 Months	13.700	-13.097	-14.034	2.647	-7.172	0.606
	(9.442)	(8.850)	(8.897)	(8.778)	(7.409)	(8.019)
CBCL Total Cutoff %						
24 Months	0.097	**-0.156**	-0.097	-0.100	-0.053	0.101
	(0.086)	**(0.074)**	(0.086)	(0.078)	(0.035)	(0.078)
36 Months	0.062	-0.114	-0.087	-0.088	-0.031	0.064
	(0.085)	(0.080)	(0.082)	(0.070)	(0.042)	(0.066)
48 Months	0.280	**-0.314**	**-0.324**	0.036	-0.179	-0.023
	(0.184)	**(0.182)**	**(0.181)**	(0.163)	(0.145)	(0.156)
CBCL Externalizing Scores						
24 Months	**0.123**	**-0.123**	**-0.123**	0.005	-0.054	-0.004
	(0.069)	**(0.069)**	**(0.069)**	(0.049)	(0.034)	(0.048)
36 Months	0.131	**-0.146**	**-0.155**	-0.061	-0.036	0.038
	(0.085)	**(0.083)**	**(0.085)**	(0.063)	(0.043)	(0.057)
48 Months	0.300	**-0.335**	-0.300	0.055	-0.196	-0.048
	(0.186)	**(0.184)**	(0.186)	(0.163)	(0.147)	(0.159)
CBCL Externalizing Cutoff %						
24 Months	1.489	-2.653	-0.750	-5.227	0.871	4.473
	(5.198)	(4.256)	(4.219)	(4.777)	(3.342)	(3.675)
36 Months	2.280	-5.088	-2.865	-6.225	-0.266	5.169
	(5.591)	(4.851)	(4.722)	(4.884)	(3.732)	(3.832)
48 Months	11.071	-13.402	-12.844	8.216	-11.584	-1.052
	(10.908)	(10.426)	(10.268)	(9.526)	(8.269)	(8.700)
CBCL Internalizing Scores						
24 Months	0.094	**-0.125**	-0.050	**-0.141**	-0.004	0.144
	(0.891)	**(0.069)**	(0.083)	**(0.084)**	(0.036)	(0.078)
36 Months	0.050	-0.029	-0.012	-0.085	0.045	0.045
	(0.089)	(0.065)	(0.065)	(0.087)	(0.065)	(0.060)
48 Months	0.254	**-0.314**	-0.274	0.071	-0.210	-0.035
	(0.188)	**(0.182)**	(0.183)	(0.166)	(0.144)	(0.158)

CBCL Internalizing Cutoff %						
24 Months	1.103	-1.967	-1.704	**-10.249**	3.636	7.102
	(5.321)	(4.102)	(4.096)	**(5.021)**	(3.284)	(3.708)
36 Months	3.100	-4.778	-0.597	-8.587	1.759	**6.701**
	(5.289)	(3.863)	(3.710)	(5.360)	(3.598)	**(3.716)**
48 Months	8.368	-6.264	-9.449	-1.757	-1.148	1.777
	(6.925)	(5.955)	(6.050)	(6.904)	(5.161)	(5.916)
BITSEA Competency Score						
24 Months	-1.754	0.729	4.460	-3.898	1.776	0.681
	(4.694)	(3.425)	(3.185)	(4.605)	(2.768)	(3.113)
36 Months	-2.495	3.870	3.704	-1.060	2.869	-1.745
	(5.252)	(3.895)	(3.751)	(4.920)	(3.121)	(3.363)
BITSEA Competency Cutoff %						
24 Months	0.033	0.005	0.005	0.109	-0.033	0.015
	(0.097)	(0.068)	(0.064)	(0.095)	(0.054)	(0.065)
36 Months	-0.067	0.021	0.138	0.124	-0.087	-0.028
	(0.117)	(0.072)	(0.084)	(0.122)	(0.076)	(0.094)
BITSEA Problems Score						
24 Months	4.381	-6.363	-5.869	-0.534	-3.875	4.252
	(5.004)	(4.175)	(4.214)	(4.695)	(3.114)	(3.879)
36 Months	4.835	-4.691	-6.082	-0.500	-1.456	4.426
	(5.317)	(4.350)	(4.350)	(5.082)	(3.175)	(4.086)
BITSEA Problems Cutoff %						
24 Months	0.019	-0.115	-0.013	-0.062	-0.081	0.148
	(0.131)	(0.100)	(0.103)	(0.127)	(0.070)	(0.100)
36 Months	0.178	-0.135	-0.069	0.017	-0.030	0.080
	(0.136)	(0.100)	(0.104)	(0.130)	(0.075)	(0.097)
SDQ Scores @ 48 Months						
Prosocial Behavior Score	-10.519	**12.143**	8.889	1.019	6.026	0.550
	(6.439)	**(4.622)**	(5.429)	(6.589)	(5.080)	(5.568)
Peer Problems	3.342	-7.318	-11.019	-0.156	-5.686	-2.405
	(10.226)	(9.814)	(9.725)	(8.239)	(7.765)	(7.464)
SDQ Cutoff @ 48 Months %						
Prosocial Behavior Score	0.066	-0.137	0.013	-0.226	0.001	0.122
	(0.144)	(0.104)	(0.127)	(0.143)	(0.097)	(0.128)
Peer Problems	0.109	-0.144	-0.191	0.028	-0.102	0.121
	(0.205)	(0.199)	(0.194)	(0.161)	(0.155)	(0.144)

Notes: [1]Estimated using IPW-adjusted OLS regressions including a gender by treatment status interaction term, gender (girl=1), and treatment status. [2] Estimated using IPW-adjusted OLS regressions including a parity by treatment status interaction term, parity status (firstborn=1), treatment status and gender (not shown). Figures in bold indicate statistically significance at the 10% or below.

Table 6 *Childcare use among the high and low treatment group from 6 – 48 months*

	N (HIGH/LOW)	*IPW* M_{HIGH} (*SD*)	*IPW* M_{LOW} (*SD*)	p^1	p^2
6 Months					
Uses any type of childcare	172 (82/90)	0.18 (0.38)	0.33 (0.47)	0.201	0.539
Age started this childcare (months)	37 (15/22)	3.69 (1.88)	2.51 (2.79)	0.765	0.848
Uses formal childcare	37 (15/22)	0.26 (0.46)	0.17 (0.39)	0.657	0.657
Hours per week in childcare	37 (15/22)	22.46 (11.39)	19.78 (9.31)	0.489	0.782
12 Months					
Uses any type of childcare	163 (80/83)	0.30 (0.46)	0.37 (0.49)	0.455	0.915
Age started this childcare (months)	61 (24/37)	6.66 (2.76)	6.74 (3.13)	0.916	0.916
Uses formal childcare	63 (25/38)	0.35 (0.49)	0.46 (0.51)	0.414	0.881
Hours per week in childcare	26 (9/17)	18.20 (5.96)	18.55 (3.37)	0.870	0.983
Childcare costs per hour (€)	26 (9/17)	1.62 (0.72)	1.91 (1.65)	0.620	0.922
18 Months					
Uses any type of childcare	153 (79/74)	0.36 (0.48)	0.45 (0.50)	0.340	0.692
Age started this childcare (months)	58 (27/31)	9.72 (6.10)	10.24 (4.59)	0.756	0.927
Uses formal childcare	59 (27/32)	0.56 (0.51)	0.73 (0.45)	0.253	0.649
Hours per week in childcare	58 (27/31)	21.58 (7.67)	21.28 (7.94)	0.883	0.883
Childcare costs per hour (€)	43 (19/24)	1.49 (0.93)	2.23 (1.84)	0.114	0.458
24 Months					
Uses any type of childcare	165 (81/84)	0.41 (0.50)	0.45 (0.50)	0.623	0.852
Age started this childcare (months)	75 (35/40)	14.27 (7.41)	13.33 (5.84)	0.563	0.958
Uses formal childcare	76 (36/40)	0.80 (0.40)	0.85 (0.36)	0.577	0.923
Hours per week in childcare	75 (35/40)	18.64 (8.93)	22.14 (8.48)	**0.083**	0.379
Childcare costs per hour (€)	69 (33/36)	2.21 (1.55)	2.13 (1.57)	0.823	0.823
36 Months					
Uses any type of childcare	150 (74/76)	0.79 (0.41)	0.76 (0.43)	0.727	0.922
Age started this childcare (months)	111 (58/53)	23.15 (10.13)	20.01 (11.25)	0.300	0.735
Uses formal childcare	112 (58/54)	0.96 (0.19)	0.94 (0.24)	0.619	0.943
Hours per week in childcare	111 (57/54)	20.21 (6.98)	20.42 (6.98)	0.877	0.877
Childcare costs per hour (€)	101 (54/47)	2.21 (2.31)	1.71 (1.14)	0.207	0.675
Attends high quality accredited center	106 (56/50)	0.65 (0.48)	0.55 (0.50)	0.410	0.764
48 Months					
Uses any type of childcare	147 (74/73)	0.79 (0.41)	0.68 (0.47)	0.322	0.638
Age started this childcare (months)	117 (59/58)	30.86 (12.19)	31.32 (13.79)	0.860	0.956

Uses formal childcare	119 (60/59)	0.98 (0.12)	0.99 (0.12)	0.927	0.927
Hours per week in childcare	117 (59/58)	16.92 (7.04)	15.94 (6.22)	0.414	0.844
Childcare costs per hour (€)	39 (21/18)	1.52 (0.79)	1.93 (1.80)	0.520	0.948
Attends high quality accredited center	117 (59/58)	0.71 (0.46)	0.64 (0.48)	0.480	0.916

Notes: 'N' indicates the sample size. 'IPW M' indicates the IPW-adjusted mean. 'IPW SD' indicates the IPW-adjusted standard deviation. [1] two-tailed conditional *p*-value from individual IPW-adjusted permutation test with 100,000 replications.

Table 7 *Testing for contamination across groups*

	N $_{(HIGH\,/_{LOW})}$	M_{HIGH} (SD)	M_{LOW} (SD)	p^1
Heard the phrase the '*Feeling Wheel'* %	147 (74/73)	0.29 (0.46)	0.03 (0.17)	**<0.001**
Heard the phrase the '*Feeling Wheel'* & accurately reports what it is %	140 (68/72)	0.23 (0.44)	0.02 (0.15)	**0.001**

Notes: 'N' indicates the sample size. 'M' indicates the IPW-adjusted mean. 'SD' indicates the IPW-adjusted standard deviation.
[1]two-tailed *p*-value from an individual IPW-adjusted permutation test with 100,000 replications.

Table 8 *Comparison of Irish nationally representative GUI cohort and PFL cohort*

	M_{GUI_3YRS} (*SD*)	$M_{GUI\text{-}5YRS}$ (*SD*)	M_{HIGH_4YRS} (*SD*)	M_{LOW_4YRS} (*SD*)	$GUI_{\_3YRS}$ v High p^1	$GUI_{\_5YRS}$ v High p^1	$GUI_{\_3YRS}$ v Low p^1	$GUI_{\_5YRS}$ v Low p^1	High v Low p^1
BAS Picture Similarity T-Score	52.76 (10.76)	58.48 (10.72)	51.51 (9.37)	49.59 (7.15)	0.327	**<0.001**	**0.020**	**<0.001**	0.203
BAS Naming Vocabulary T-Score	50.78 (12.78)	55.24 (12.05)	53.29 (11.18)	45.95 (11.21)	**0.097**	0.174	**0.003**	**<0.001**	**<0.001**
SDQ Peer Problems	1.21 (1.40)	1.01 (1.33)	1.32 (1.41)	1.79 (1.92)	0.496	**0.046**	**0.001**	**<0.001**	**0.094**
SDQ Prosocial Behavior	7.94 (1.77)	8.43 (1.65)	8.49 (1.60)	7.79 (2.03)	**0.007**	0.733	0.476	**0.001**	**0.021**
N	9,179-9,786	8,886-8,998	71-74	63-73					

Notes: The BAS T-scores are standardized to have a mean of 50 and a standard deviation of 10. M' indicates the weighted mean. 'SD' indicates the weighted standard deviation. [1] two-tailed *p*-value from an unpaired t test with weights applied.

Appendix A

Table A1 *PFL households in receipt of social welfare payments at 48 months*

Unemployment Payments	Jobseeker's Benefit	13.6%
	Jobseeker's Allowance or Unemployment Assistance	17.0%
Employment Supports	Family Income Supplement	15.0%
	Back to Work Enterprise Allowance	0.7%
	Farm Assist	0.0%
	Part-time Job Incentive Scheme	2.0%
	Back to Work Allowance (Employees)	0.0%
	Back to Education Allowance	2.0%
	Supplementary Welfare Allowance (SWA)	4.1%
One-Parent Family/Widower Payments	Widow's or Widower's (Contributory) Pension	2.0%
	Deserted Wife's Allowance	0.7%
	Deserted Wife's Benefit	0.7%
	Prisoner's Wife Allowance	0.0%
	Widowed Parent Grant	0.0%
	One-parent Family Payment	39.5%
	Widow's or Widower's (Non-contributory) Pension	0.7%
Child Related Payments	Maternity Benefit	2.7%
	Health and Safety Benefit	0.0%
	Adoptive Benefit	0.0%
	Guardian's Payment (Contributory)	0.7%
	Guardian's Payment (Non-Contributory)	0.0%
Disability and Caring Payments	Illness Benefit	3.4%
	Injury Benefit	0.0%
	Invalidity Pension	1.4%
	Incapacity Supplement	0.0%
	Disability Allowance	4.8%
	Disablement Benefit	0.7%
	Blind Pension	0.0%
	Medical Care Scheme	1.4%
	Carer's Benefit	2.0%
	Medical Card	77.6%
	GP Visit Card	6.8%
	Constant Attendance Allowance	0.0%
	Domiciliary Care Allowance	2.7%
	Death Benefits (Survivor's Benefits)	0.0%
	Partial Capacity Benefit	0.0%
	Carer's Allowance	3.4%
	Mobility Allowance	0.0%
	Dependent Persons Pension	0.0%
Retirement Payments	State Pension (Transition)	0.0%
	State Pension (Non-Contributory)	1.4%
	State Pension (Contributory)	0.7%
	Pre-Retirement Allowance	0.0%
% Receiving any benefits		87%
N		147

Appendix B

Table B1 *Comparison of PFL participants and eligible non-participants at baseline*

	Total *N* *(Part./Non-part.)*	*PFL participants* $M_{(SD)}$	*Non-participants* $M_{(SD)}$	p^i
Gender of study child – Girl (%)	301 (199/102)	0.54 (0.50)	0.47 (0.50)	0.232
Number of older siblings the study child has	286 (210/76)	0.93 (1.24)	0.99 (1.05)	0.729
Mother's age when had first child	311 (210/101)	21.65 (4.16)	23.53 (5.39)	**0.003**
First-time mother when had study child (%)	312 (210/102)	0.51 (0.50)	0.55 (0.50)	0.578
Mother eligible for free medical care (%)	306 (210/96)	0.63 (0.48)	0.47 (0.50)	**0.009**
Married (%)	306 (210/96)	0.16 (0.37)	0.24 (0.43)	0.103
Partner (%)	306 (210/96)	0.81 (0.39)	0.85 (0.36)	0.385
Single (%)	306 (210/96)	0.19 (0.39)	0.15 (0.36)	0.385
Age left full-time education	282 (191/91)	17.41 (2.78)	17.81 (2.08)	0.187
Finished full-time education (%)	253 (152/101)	0.86 (0.35)	0.90 (0.30)	0.349
Leaving Cert education or higher (%)	312 (210/102)	0.46 (0.50)	0.59 (0.50)	**0.037**
Left school before the age of sixteen (%)	282 (191/91)	0.19 (0.39)	0.04 (0.21)	**<0.001**
Employed (%)	306 (210/96)	0.39 (0.49)	0.67 (0.47)	**<0.001**
Engaged in skilled work (%)	133 (76/57)	0.71 (0.46)	0.61 (0.49)	0.256

Notes: The *PFL* participants include the high and low treatment groups. All baseline measures pertain to when the participant was pregnant with the study child measured during pregnancy for the *PFL* participants, and when the study child was four years old for the eligible non-participants. 'M' indicates the mean. 'SD' indicates the standard deviation. [i] two-tailed *p*-value from permutation test with 100,000 replications.

Appendix C

Table C1 *List of PFL Tip Sheets*

Tip Sheets	**Pre-birth – 12 months**	**12 – 24 months**	**24 – 48 Months**
Cognitive Development	Milestones 0-6 months, Milestones 6-12 months; Cognitive Development 0-3 months; Cognitive Development 3-6 months; Cognitive Development 6-12 months; Playing and learning; Hand-eye coordination 0-6 months; Hand-eye coordination 6-12 months; Language development 0-3 months; Language development 3-6 months; Language development 6-12 months; Developing movement 0-6 months; Developing movement 6-12 months	Milestones 12-24 months; Movement; Listening and Talking; Listening and Talking 2; First steps towards learning to read; Stories and books; First steps towards learning to write; First steps towards learning numbers; Learning through play; Messy play; Playing outdoors; Action rhymes 2	Getting Ready for Maths; Getting Ready for Writing; Children and Art 1; Children and Art 2; Children and Art 3; Basic Skills for School: Using Scissors; Basic Skills for School: Drawing Shapes; Basic Skills for School: Getting Dressed; Basic Skills for School: Hop, Skip and Jump; Basic Skills for School: Managing a Lunch Box; Basic Skills for School: Tying Shoelaces; Encouraging your Toddler's Play; Play; Sand Play; Water Play; Play Dough; Developing your Child's Language; Reading Together; Music and Learning; Milestones for 2 Years; Milestones for 3 Years; Developing Vocabulary1; Developing Vocabulary2; Developing Vocabulary3; Developing Vocabulary4; Developing Vocabulary5; Developing Vocabulary6
Social & Emotional Development	Circle of repair, Circle of trust; Circle of security; Getting to know your baby pre-birth; Getting to know your baby 0-3 months; Attachment; Secure base; Social and emotional development confidence 0-12 months; Getting to know your baby0-3 months communicating; Getting to know your baby 0-3 months regulation; Mutual gaze; Getting to know your baby 0-3 months tired signs; Getting to know your baby 0-3 months siblings; Social and emotional development 6-12 months	Child parent relationship; Self-awareness; Fear; Self-assertion; Temper tantrums; Learning to play; secure base; What is it like to be 12 months; What is it like to be 13 months; What is it like to be 14 months; What is it like to be 15 months; What is it like to be 16 months; What is it like to be 17 months; What is it like to be 18 months; What is it like to be 19 months ; What is it like to be 20 months; What is it like to be 21 months' What is it like to be 22 months; What is it like to be 23 months; What is it like to be 24 months	Caring and Sharing; Emotions; Expressing Emotions; List of Feeling Words; Creative Play; Social Skills; Disobedience; Friendships; Hurting Others; Giving Praise; Lies; Nightmares; Role Play 1; Role Play 2; Self Esteem; Separation Problems; Tantrums; The Toddler Years; Whining; Being Three; Being Four; ADD & ADHD; Sharing; Biting; Feeling Wheel
Rest & Routine / Parenting supports	Routine, Rest during pregnancy; Crying, Sleep 0-6 months; Cot death; Sleep chart; Daily routine; Sleep 6-12 months Family planning; Extra supports for parents; Support agencies 1; Support agencies 2; Relationships mam dad baby; Relationships quality time; Relationships mam and dad; Relationships making changes; Postnatal depression; Preparing for labor; Labor; Labor	Routine 1; Routine 2; Daily routine; Sleeping and crying; Exercise; Looking after yourself 1-2 years; Especially for Mams and Dads; Supports	Bedtime Routine; Sleep Diary; Toilet Training

	birth plan; Labor and delivery; After the birth; Different types of families; Work, leave and entitlements		
Nutrition	Nutrition during pregnancy – portion size; Nutrition during pregnancy – weight gain; Nutrition during pregnancy – nutrients; Food safety; Managing common complaints; Breastfeeding; Breastfeeding patterns; Breastfeeding getting started; Breastfeeding expressing; Storing breastmilk; Formula feeding how much; Formula feeding advance preparation; Weaning to solids introduction; Weaning to solids chart; Weaning to solids tips; Weaning to solids drinks; Spoon feeding questions	Allergies and constipation; Food groups; Fussy eating; General freezing and thawing; Getting the balance right; Hygiene in the kitchen; Iron and calcium; Making most of mealtimes; Recipes for children; Sample meal planner; Shopping guide; Smart drinks for smart kids; Suitable snacks; The food pyramid; Pureed recipes for children; A diary of food; Twelve ways to disguise vegetables, Be sugar smart	Food Groups 1; Food Groups 2; Food Groups 3; Shopping and Labels; The Food Pyramid; Iron; Healthy Eating Recipes; Meal Planner; Healthy Eating for Teeth; Healthy Lifestyle for Children; Mealtimes
Safety & Supervision	Smoking; Alcohol; Drug use; Domestic violence; Immunizing; Baby health; Travelling in a car, Caring for your baby, Childhood illness 0-6 month, Temperature; Keeping baby safe 0-6 months; Teething; Keeping baby safe 6 months – 2 years; Kid safe rooms; Childhood illness 6-24 months	Travelling in the car; Baby's health; Teething; Keeping baby safe 6 months – 2 years; Kid safe rooms; Childhood illness 6-24 months; Basic first aid; Caring for your child's teeth; Playing with toys; Teaching your child safety; Head lice; Soothers	Television 1; Television 2; Television 3; Soothers; Thumb-sucking; Passive Smoking; Family Holidays

Example of a Tip Sheet

Listening and Talking

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

Things you can do to help your child:

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

Sounds around us

Indoors:	Outdoors:
✓ tap running	✓ plane overhead
✓ radio and TV	✓ car, bus, train
✓ baby crying	✓ wind in the trees
✓ children playing	✓ someone calling
✓ washing machine	✓ birds or insects

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

Appendix D

Description of Outcome Measures

Cognitive Outcomes

Developmental Profile 3- Cognitive Section
Children's cognitive development during the program was assessed at 24, 36, and 48 months using the Developmental Profile-3 (DP-3; Alpern 2007). The DP-3 is a maternal report measure of child development from birth to age 12 years and 11 months. The cognitive section is a 38-item scale measuring cognitive abilities ($\alpha = 0.79 - 0.84$), starting at number 1 and continuing until the stop rule is satisfied (i.e. when five consecutive no responses are recorded). Each of the items refers to tasks which require cognitive skill and were arranged in order of difficulty. For example, 'Does your child say size words (large or big, and little or small) correctly'. For each item, participants were asked whether their child had carried out the task and responded *yes* or *no* accordingly. The *yes* responses were tabulated to create a continuous score whereby higher values indicated greater cognitive development. These scores were standardized by age according to the DP3 normative sample, with a mean of 100 and standard deviation of 15. In addition, a binary variable was created to indicate those scoring above average, that is, a score of above 115.

Ages and Stages Questionnaire
Children's communication and problem solving skills during the program were assessed at 24, 36, and 48 months using maternal reports on the *Ages and Stages Questionnaire* (ASQ; Squires *et al*. 1999). The ASQ consists of 19 screening questionnaires at specific age intervals ranging from 4 to 60 months of age. Each questionnaire consists of a 30-item instrument for identifying children at risk for developmental delay. The ASQ measures five domains of development including Communication, Problem Solving Skills, Gross Motor Skills, Fine Motor Skills, and Personal-Social Skills. The current paper uses the Communication *(*$\alpha = 0.49 - 0.78$) sub-domain which measures the child's understanding of language, naming of items and word combinations, and the Problem Solving *(*$\alpha = 0.27 - 0.55$) sub-domain which measures the child's ability to follow instruction, pretense, and problem solving. During the interview, the interviewer asked the mother questions related to different activities her child was capable of at that time. The mother responded by indicating whether her child exhibited the behavior *regularly, sometimes,* or *not yet.* If the mother did not know whether her child was capable of the behavior, where possible, the interviewer asked her to test the behavior with the child during the interview using the ASQ toolkit. Domain scores represent the sum of all six items in that domain, resulting in a possible range of 0 to 60 with higher scores indicative of more advanced development. The scores were standardized to have a mean of 100 and standard deviation of 15. In addition, age-specific standardized cut-off points for each domain were used to derive cutoff scores indicating if the child was considered to be at risk of developmental delay in that domain.

British Ability Scales II
Children's cognitive development was measured at the end of the program by direct assessment of the children using the British Ability Scales II: Early Years Battery (BAS II; Elliott *et al.* 1997). Assessments were conducted in either the participant's home (33 percent), in a local community centre (27 percent), or in the child's childcare setting (40 percent) by trained assessors who were blind to the children's treatment assignment. On average, the participants were 50.5 months when they completed the assessment. Each assessment lasted approximately 30 minutes and children received a gift as a thank you for their time. The BAS II early years battery was designed as an assessment of children's abilities in clinical, educational, and research settings for children ages 3 years and 6 months to 5 years 11 months. The upper level battery consists of six subscales: verbal comprehension, naming vocabulary, picture similarities, early number concepts, pattern construction, and copying. These sub-scales yield an overall score reflecting General Conceptual Ability (GCA) which is a proxy for IQ, as well as three cluster scores for Verbal Ability, Pictorial Reasoning Ability, and Spatial

Ability. The GCA score assesses overall cognitive ability such as thinking logically, making decisions, and learning. The Spatial Ability score assesses problem solving and coordination. The Pictorial Reasoning score assesses non-verbal reasoning such as the ability to detect similarities and knowledge of numbers. The Verbal Ability score assesses children's overall ability to understand (using listening skills) and express language. The *T* scores are calculated for each domain and standardized to have a mean of 100 and a standard deviation of 15, as well as cutoff scores indicating whether the child scores below or above average for the GCA and cluster scores.

Socio-emotional and Behavioral Outcomes

Child Behavior Checklist

Children's behavioral skills were assessed at the 24, 36 and 48 month assessments using maternal reports on the *Child Behavior Checklist for Ages 1½ -5* (CBCL; Achenbach and Rescorla 2000). The CBCL is a 100 item instrument for assessing externalizing behavior and internalizing behavior in children aged 18 months to age five. It includes three possible response options, *not true, somewhat/sometimes true, or very true/often true,* which correspond to 0, 1, and 2 respectively. The CBCL consists of seven syndromes - emotionally reactive, anxious/depressed, somatic complaints, withdrawn, sleep problems, attention problems, aggressive behavior, and one 'other problems' category. These eight categories map onto two subscales, Internalizing (a= 0.90 - 0.91) and Externalizing Problems (a= 0.90 - 0.92), and also a Total Problems score (a= 0.95 - 0.96) by generating standardized *T* scores for each. The scores were standardized to have a mean of 100 and standard deviation of 15. In addition, for each scale the clinical cutoff *T* score was used to index children with more significant problems. Missing data for individual items were imputed using the mean plus a random residual value and was approved by the instrument's developer. If more than eight items were missing, participants were excluded from the analysis.

Brief Infant-Toddler Social and Emotional Assessment

Children's socio-emotional skills were assessed at 24 and 36 months using maternal reports on the *Brief Infant-Toddler Social and Emotional Assessment* (BITSEA; Briggs-Gowan and Carter 2006). The BITSEA is a 42-item screening tool for social-emotional/behavioral problems and delays in competence in children. The BITSEA yields a Problem score (α= 0.85 - 0.87) and a Competence score (α = 0.64 -0.71). Problem behavior items include externalizing, internalizing, and dysregulation problems. Higher values on the Problem score indicate greater levels behavioral problems. Competencies include areas of attention, compliancy, mastery, motivation, pro-social peer relations, empathy, play skills and social relatedness. Lower values on the Competence score indicate possible delays. All scores are normed by child gender. The scores were standardized to have a mean of 100 and standard deviation of 15.

Strengths and Difficulties Questionnaire: Peer Problems and Prosocial Subscales

Children's socio-emotional skills were assessed at the 48 month assessment using maternal reports on the *Strengths and Difficulties Questionnaire* (SDQ; Goodman, 1997). The SDQ is a 25-item questionnaire assessing behaviors, emotions, and relationships of four to 16 year olds. The questionnaire covers five dimensions: conduct problems, emotional symptoms, hyperactivity, peer problems, and prosocial behavior. The 5-item Peer Problems (α=0.48) and 5-item Prosocial (α=0.72) subscales were assessed in the *PFL* study. Items were scored 0 for *not true*, 1 for *somewhat true* and 2 for *certainly true*. Two items from the Peer Problems subscale were reverse scored. The five items for each subscale were summed giving a total score of 0 to 10 for each subscale. The scores were standardized to have a mean of 100 and standard deviation of 15.

Appendix E

Table E1 *Testing for attrition: early childhood outcomes using later estimation samples*

	N (HIGH/LOW)	M_{HIGH} (SD)	M_{LOW} (SD)	p^1	p^2
Original Estimation Sample					
ASQ Communication Score 6M	173 (83/90)	101.23 (14.37)	98.86 (15.56)	0.123	0.720
ASQ Communication Score 12M	165 (82/83)	99.78 (15.18)	100.21 (14.91)	0.470	0.966
ASQ Problem Solving Score 6M	173 (83/90)	99.44 (14.60)	100.51 (15.42)	0.733	0.967
ASQ Problem Solving Score 12M	165 (82/83)	100.05 (14.20)	99.95 (15.84)	0.390	0.947
DP3 Cognitive Score 12M	165 (82/83)	100.54 (13.79)	99.47 (16.18)	0.249	0.925
Difficult Temperament Score 6M	173 (83/90)	11.70 (5.71)	12.21 (5.50)	0.354	0.911
ASQ Socio-emotional Score 6M	173 (83/90)	14.76 (10.68)	15.17 (13.75)	0.408	0.958
BITSEA Competency Score 12M	165 (82/83)	101.19 (14.61)	98.83 (15.38)	0.903	0.903
BITSEA Problems Score	165 (82/83)	99.89 (14.10)	100.11 (15.93)	0.450	0.965
24 Month Estimation Sample					
ASQ Communication Score 6M	162 (80/82)	102.13 (13.53)	99.52 (15.84)	0.115	0.655
ASQ Communication Score 12M	159 (80/79)	99.52 (15.27)	100.85 (13.83)	0.622	0.969
ASQ Problem Solving Score 6M	162 (80/82)	100.04 (14.20)	100.43 (15.75)	0.639	0.964
ASQ Problem Solving Score 12M	159 (80/79)	100.09 (14.08)	100.34 (15.06)	0.458	0.970
DP3 Cognitive Score 12M	159 (80/79)	100.19 (13.76)	99.72 (15.77)	0.329	0.971
Difficult Temperament Score 6M	162 (80/82)	11.60 (5.72)	12.37 (4.99)	0.215	0.760
ASQ Socio-emotional Score 6M	162 (80/82)	14.69 (10.74)	14.57 (12.15)	0.513	0.981
BITSEA Competency Score 12M	159 (80/79)	100.85 (14.59)	98.85 (15.69)	0.870	0.870
BITSEA Problems Score	159 (80/79)	100.10 (14.18)	100.37 (15.80)	0.435	0.970
36 Month Estimation Sample					
ASQ Communication Score 6M	147 (73/74)	102.63 (13.09)	99.32 (15.60)	**0.075**	0.510
ASQ Communication Score 12M	145 (73/72)	99.76 (15.58)	99.96 (15.37)	0.465	0.949
ASQ Problem Solving Score 6M	147 (73/74)	100.16 (14.49)	100.11 (16.23)	0.573	0.968
ASQ Problem Solving Score 12M	145 (73/72)	100.42 (14.29)	98.85 (16.42)	0.216	0.897
DP3 Cognitive Score 12M	145 (73/72)	100.40 (13.98)	98.07 (16.36)	0.125	0.778
Difficult Temperament Score 6M	147 (73/74)	11.44 (5.57)	11.96 (5.20)	0.332	0.884
ASQ Socio-emotional Score 6M	147 (73/74)	14.79 (10.97)	13.99 (11.59)	0.688	0.949
BITSEA Competency Score 12M	145 (73/72)	100.51 (15.05)	97.46 (15.72)	0.933	0.933
BITSEA Problems Score	145 (73/72)	99.70 (13.96)	99.80 (15.59)	0.516	0.978
48 Month Estimation Sample					

ASQ Communication Score 6M	142 (70/72)	101.62 (14.45)	99.80 (14.90)	0.217	0.883
ASQ Communication Score 12M	137 (68/69)	98.71 (15.82)	100.26 (15.59)	0.674	0.674
ASQ Problem Solving Score 6M	142	99.32 (15.15)	100.43 (15.81)	0.718	0.996
ASQ Problem Solving Score 12M	137 (68/69)	100.40 (14.61)	99.40 (15.83)	0.315	0.946
DP3 Cognitive Score 12M	137 (68/69)	99.33 (14.09)	99.27 (16.76)	0.426	0.981
Difficult Temperament Score 6M	142 (70/72)	11.24 (5.68)	11.69 (5.09)	0.361	0.942
ASQ Socio-emotional Score 6M	142 (70/72)	14.64 (10.91)	13.82 (11.64)	0.654	0.985
BITSEA Competency Score 12M	137 (68/69)	100.20 (14.81)	98.94 (15.17)	0.777	0.966
BITSEA Problems Score	137 (68/69)	98.83 (13.86)	99.42 (15.71)	0.425	0.965
51 Month Estimation Sample					
ASQ Communication Score 6M	130 (68/62)	101.80 (14.63)	99.30 (16.29)	0.153	0.797
ASQ Communication Score 12M	125 (66/59)	99.21 (15.69)	99.36 (16.03)	0.438	0.943
ASQ Problem Solving Score 6M	130 (68/62)	98.94 (15.27)	98.55 (17.12)	0.480	0.942
ASQ Problem Solving Score 12M	125 (66/59)	100.37 (14.74)	98.45 (16.48)	0.201	0.822
DP3 Cognitive Score 12M	125 (66/59)	100.05 (13.64)	97.96 (17.25)	0.169	0.822
Difficult Temperament Score 6M	130 (68/62)	11.26 (5.72)	11.98 (5.14)	0.254	0.856
ASQ Socio-emotional Score 6M	130 (68/62)	14.63 (10.84)	14.44 (11.28)	0.537	0.888
BITSEA Competency Score 12M	125 (66/59)	100.48 (14.99)	97.60 (15.53)	0.924	0.924
BITSEA Problems Score	125 (66/59)	99.26 (13.82)	100.92 (15.71)	0.279	0.815

Notes: These models estimate treatment effects at 6 and 12 months using the 24, 36, 48, and 51 month estimation samples**.** 'N' indicates the sample size. 'M' indicates the mean. 'SD' indicates the standard deviation. [1]one-tailed (right-sided) conditional *p*-value from individual permutation test with 100,000 replications. [2]one-tailed (right-sided) conditional *p*-value from stepdown permutation test with 100,000 replications. Child gender included in all analyses.

Appendix F

Table F1 *Testing for attrition: early childhood outcomes & later study participation*

6 Month Outcomes	24M Stayer	24M Non-stayer	p^1	36M Stayer	36M Non-stayer	p^1	48M Stayer	48M Non-stayer	p^1	51M Stayer	51M Non-stayer	p^1
ASQ Communication Score	100.81	88.11	**0.008**	100.96	94.55	**0.075**	100.70	96.79	0.222	100.61	98.16	0.323
	(14.76)	(14.08)		(14.46)	(17.05)		(14.66)	(16.34)		(15.43)	(13.61)	
ASQ Problem Solving Score	100.23	96.54	0.463	100.14	99.23	0.754	99.88	100.55	0.800	98.75	103.77	**0.019**
	(14.96)	(15.93)		(15.34)	(13.18)		(15.44)	(12.98)		(16.11)	(10.23)	
ASQ Socio-emotional Score	14.63	20.00	0.432	14.39	18.27	0.272	14.23	18.39	0.181	14.62	16.05	0.587
	(11.44)	(21.91)		(11.26)	(17.14)		(11.26)	(16.20)		(10.94)	(15.94)	
Difficult Temperament Score	11.99	11.64	0.894	11.70	13.46	0.202	11.47	14.23	**0.023**	11.58	13.14	0.135
	(5.36)	(8.66)		(5.38)	(6.59)		(5.37)	(6.09)		(5.40)	(6.03)	
N	162	11		147	26		142	31		130	43	
12 Month Outcomes	24M Stayer	24M Non-stayer	p^1	36M Stayer	36M Non-stayer	p^1	48M Stayer	48M Non-stayer	p^1	51M Stayer	51M Non-stayer	p^1
DP3 Cognitive Score	99.96	101.18	0.896	99.24	105.49	**0.044**	99.30	103.41	0.128	99.01	103.09	0.110
	(14.74)	(22.54)		(15.20)	(12.41)		(15.44)	(12.33)		(15.39)	(13.41)	
ASQ Communication Score	100.18	95.25	0.647	99.86	101.02	0.694	99.49	102.48	0.233	99.28	102.25	0.216
	(14.54)	(25.93)		(15.42)	(11.75)		(15.67)	(11.08)		(15.79)	(12.12)	
ASQ Problem Solving Score	100.21	94.31	0.581	99.64	102.59	0.327	99.90	100.51	0.836	99.37	101.98	0.302
	(14.53)	(25.90)		(15.35)	(12.14)		(15.19)	(14.30)		(15.48)	(13.36)	
BITSEA Competency Score	99.86	103.82	0.410	99.00	107.27	**0.001**	99.57	102.12	0.425	99.19	102.52	0.199
	(15.13)	(11.26)		(15.41)	(8.92)		(14.95)	(15.33)		(15.33)	(13.79)	
BITSEA Problems Score	100.23	93.80	0.341	99.75	101.82	0.607	99.12	104.29	0.113	99.89	100.34	0.874
	(14.96)	(16.11)		(14.74)	(17.07)		(14.77)	(15.64)		(14.78)	(15.85)	
N	159	6		145	20		137	28		125	40	

Notes: Mean and standard deviation of children's skills at 6 and 12 months reported for those who participated and those who did not in the 24, 36, 48, and 51 month assessments respectively. [1] two-tailed *p*-value from individual IPW-adjusted permutation test with 100,000 replications.

Appendix G

Table G1 *Baseline predictors of attrition*

	High Treatment Group	**Low Treatment Group**
24 Months	WASI *perceptual reasoning* score (-), AAPI *parental expectations of children* score (-), AAPI *parental empathy towards children's needs* score (+), support from relatives (+), drinks alcohol during pregnancy (-), knows neighbors (+) (6 variables)	Eats healthily (-), exercises regularly (-), has ever taken illegal drugs (+), satisfaction with neighborhood (+), Irish national (-) (5 variables)
36 Months	WASI *perceptual reasoning* score (-), AAPI *parental expectations of children* score (-), AAPI *parental empathy towards children's needs* score (-), AAPI *children's power and independence* score (+), support from relatives (-), satisfaction with neighborhood (-) (6 variables)	WASI *verbal ability* score (-), TIPI *agreeableness* score (-), TIPI *conscientiousness* score (+), TIPI *openness* score (-), AAPI *parental expectations of children* score (-), AAPI *parental empathy towards children's needs* score (-), KIDI score (-), age (-), married (+), experience financial difficulty (+), prior physical health condition (-), exercises regularly (-), has ever used drugs (+), satisfaction with neighborhood (+) (14 variables)
48 Months	WASI *perceptual reasoning* score (-), AAPI *parental responsiveness* score (-), AAPI *parental empathy towards children's needs* score (+),drinks alcohol during pregnancy (-), Irish national (-) (5 variables)	WASI *verbal ability* score (-), TIPI *openness* score (-), AAPI *parental expectations of children* score (-), AAPI *parental empathy towards children's needs* score (+), AAPI *parental responsiveness* score (-), AAPI *children's power and independence* score (-), low education (+), Irish national (-), took folic acid during pregnancy (+), has a medical card (+), ever used drugs (+) (11 variables)
51 Months	WASI performance score (-), age (-), took folic acid during pregnancy (-), AAPI *parental empathy towards children's needs* score (+),AAPI *parental responsiveness* score (-), receives social welfare payments (-), support from relative (-),support from friends (-), low education (+), employed during pregnancy (-), drank alcohol during pregnancy (-), Irish national (-) (12 variables)	Took folic acid during pregnancy (+), Pearlin *mastery* score (-), VASQ *insecure attachment* score (-), activities impaired by illness (-), has a medical card (-), TIPI *agreeableness* score (-), TIPI *openness* score (-), Consideration of Future Consequences score (+),AAPI *parental expectations of children* score (-), AAPI *parental empathy towards children's needs* score (+), low education (+), saves money regularly (-), experience financial difficulty (+), resides in social housing (-), no. of health services used (-), ever used drugs (+), knows neighbors (+), no. of services used (-), Irish national (-) (19 variables)

Note: The table includes the set of variables which resulted in the lowest BIC in models of attrition and are included in the logistic model used to generate the IPW weights. (+) and (-) indicates a participant with this characteristic has a higher/lower probability of dropping out.

Appendix H

Table H1 *Cognitive skills stepdown family results*

	Continuous Scores	*Cutoff Scores*
	IPW Stepdown p[1]	*IPW Stepdown p*[1]
DP3 24 Months	0.157	0.278
DP3 36 Months	**0.047**	**0.073**
DP3 48 Months	0.089	0.123
ASQ Communication 24 Months	0.345	0.633
ASQ Communication 36 Months	0.181	0.463
ASQ Communication 48 Months	0.336	0.395
ASQ Problem Solving 24 Months	0.171	0.287
ASQ Problem Solving 36 Months	**0.069**	0.264
ASQ Problem Solving 48 Months	0.451	0.505
BAS General Conceptual Ability 51 months	**0.001**	~
BAS Spatial Ability 51 months	**0.003**	~
BAS Pictorial Reasoning Ability 51 months	**0.017**	~
BAS Language Ability 51 months	**0.003**	~
BAS General Conceptual Ability below average cutoff 51 months	~	**<0.001**
BAS Spatial Ability below average cutoff 51 months	~	**0.013**
BAS Pictorial Reasoning Ability below average cutoff 51 months	~	0.165
BAS Language Ability below average cutoff 51 months	~	**0.081**
BAS General Conceptual Ability above average cutoff 51 months	~	**0.060**
BAS Spatial Ability above average cutoff 51 months	~	0.457
BAS Pictorial Reasoning Ability above average cutoff 51 months	~	0.301
BAS Language Ability above average cutoff 51 months	~	**0.052**

Notes: [1] one-tailed (right-sided) conditional *p*-value from IPW-adjusted stepdown permutation test with 100,000 replications including all cognitive outcomes.

Table H2 *Socio-emotional and behavioral skills stepdown family results*

	Continuous Scores	*Cutoff Scores*
	IPW Stepdown p[1]	*IPW Stepdown p*[1]
CBCL Total Scores 24 Months	0.367	**0.060**
CBCL Total Scores 36 Months	0.275	0.328
CBCL Total Scores 48 Months	0.312	**0.059**
CBCL Externalizing Scores 24 Months	0.516	0.298
CBCL Externalizing Scores 36 Months	0.317	0.347
CBCL Externalizing Scores 48 Months	0.183	**0.015**
CBCL Internalizing Scores 24 Months	0.427	0.343
CBCL Internalizing Scores 36 Months	0.370	0.724
CBCL Internalizing Scores 48 Months	0.575	**0.058**
BITSEA Competency Score 24 Months	0.541	0.690
BITSEA Competency Score 36 Months	0.547	0.709
BITSEA Problems Score 24 Months	0.323	0.359
BITSEA Problems Score 36 Months	0.500	0.629
SDQ Prosocial Behavior Score 48 months	0.178	0.349
SDQ Peer Problems 48 months	0.319	0.499

Notes: [1] one-tailed (right-sided) conditional *p*-value from IPW-adjusted stepdown permutation test with 100,000 replications including all socio-emotional and behavioral outcomes.

Appendix I

Table I1 *Cognitive skills results – conditioning on baseline covariates*

	N (HIGH/LOW)	*IPW* M_{HIGH} (*SD*)	*IPW* M_{LOW} (*SD*)	*IPW Treat. Effect*	*IPW Effect Size*	p^1	p^2
DP3 Scores							
24 Months	163 (82/81)	101.64 (13.61)	98.20 (15.85)	3.15	0.20	**0.048**	**0.048**
36 Months	147 (74/73)	102.64 (14.90)	96.39 (14.40)	5.16	0.36	**0.015**	**0.034**
48 Months	143 (73/70)	102.35 (13.23)	97.40 (15.65)	4.18	0.27	**0.037**	**0.054**
DP3 Cutoffs - Above Average %							
24 Months	163 (82/81)	0.64 (0.48)	0.53 (0.50)	0.10	0.20	**0.076**	**0.076**
36 Months	147 (74/73)	0.52 (0.50)	0.31 (0.47)	0.18	0.38	**0.013**	**0.027**
48 Months	143 (73/70)	0.31 (0.47)	0.17 (0.38)	0.13	0.36	**0.013**	**0.031**
ASQ Communication Scores							
24 Months	163 (82/81)	100.41 (15.05)	100.92 (13.73)	-0.97	-0.07	0.483	0.483
36 Months	147 (75/72)	101.38 (14.17)	97.33 (16.11)	3.83	0.24	**0.093**	0.117
48 Months	143 (73/70)	100.98 (13.23)	99.52 (15.03)	1.13	0.08	0.138	0.259
ASQ Communication Cutoffs – Below Average %							
24 Months	163 (82/81)	0.10 (0.30)	0.06 (0.24)	-0.05	-0.19	0.749	0.749
36 Months	147 (75/72)	0.04 (0.21)	0.06 (0.24)	0.01	0.06	0.216	0.427
48 Months	143 (73/70)	0.04 (0.21)	0.05 (0.22)	0.01	0.04	0.245	0.400
ASQ Problem Solving Scores							
24 Months	163 (82/81)	101.67 (15.19)	98.78 (14.95)	1.96	0.13	0.183	0.298
36 Months	144 (73/71)	102.28 (13.58)	96.86 (14.92)	4.99	0.33	**0.034**	**0.067**
48 Months	143 (73/70)	100.42 (14.56)	99.94 (16.94)	-0.38	-0.02	0.413	0.413
ASQ Problem Solving Cutoffs – Below Average %							
24 Months	163 (82/81)	0.07 (0.25)	0.13 (0.34)	0.06	0.17	0.123	0.227
36 Months	144 (73/71)	0.11 (0.31)	0.17 (0.38)	0.07	0.20	**0.087**	0.276
48 Months	143 (73/70)	0.05 (0.22)	0.07 (0.25)	0.01	0.03	0.367	0.367
BAS Scores @ 51 Months							
General Conceptual Ability	119 (63/56)	104.97 (15.25)	94.54 (13.33)	10.45	0.78	**<0.001**	**<0.001**
Spatial Ability	120 (63/57)	104.69 (14.65)	95.76 (13.09)	9.06	0.69	**<0.001**	**0.001**
Pictorial Reasoning Ability	123 (65/58)	103.77 (15.68)	96.31 (12.77)	8.01	0.63	**<0.001**	**<0.001**
Language Ability	125 (65/60)	103.69 (15.76)	94.22 (14.96)	9.15	0.61	**0.001**	**0.002**
BAS Cutoffs - Below Average @							

51 Months %							
General Conceptual Ability	124 (68/56)	0.20 (0.40)	0.61 (0.49)	0.39	0.79	**<0.001**	**<0.001**
Spatial Ability	125 (68/57)	0.30 (0.46)	0.60 (0.50)	0.28	0.58	**0.001**	**0.002**
Pictorial Reasoning Ability	128 (70/58)	0.29 (0.46)	0.46 (0.50)	0.14	0.27	**0.046**	**0.046**
Language Ability	130 (70/60)	0.26 (0.44)	0.45 (0.50)	0.16	0.32	**0.017**	**0.038**
BAS Cutoffs - Above Average @ 51 Months %							
General Conceptual Ability	124 (68/56)	0.26 (0.44)	0.08 (0.28)	0.16	0.58	**0.021**	**0.036**
Spatial Ability	125 (68/57)	0.14 (0.35)	0.08 (0.28)	0.06	0.21	**0.095**	**0.095**
Pictorial Reasoning Ability	128 (70/58)	0.17 (0.38)	0.10 (0.30)	0.09	0.30	**0.045**	**0.081**
Language Ability	130 (70/60)	0.25 (0.44)	0.08 (0.28)	0.14	0.51	**0.022**	**0.035**

Notes: 'N' indicates the sample size. 'IPW M' indicates the IPW-adjusted mean. 'IPW SD' indicates the IPW-adjusted standard deviation. [1]one-tailed (right-sided) conditional *p*-value from individual IPW-adjusted permutation test with 100,000 replications. [2]one-tailed (right-sided) conditional *p*-value from IPW-adjusted stepdown permutation test with 100,000 replications. The conditioning set includes maternal knowledge of child development, parenting self-efficacy, maternal attachment, and maternal consideration of future consequences, as well as child gender which is included in all analyses. 'Treatment effect' is the difference in means between the high and low treatment group. 'Effect size' is the ratio of the treatment effect to the standard deviation of the low treatment group.

Table I2 *Socio-emotional and behavioral skills results – conditioning on baseline covariates*

	N (HIGH/LOW)	*IPW* M_{HIGH} (*SD*)	*IPW* M_{LOW} (*SD*)	*IPW Treat. Effect*	*IPW Effect Size*	p^1	p^2
CBCL Total Scores							
24 Months	161 (81/80)	98.74 (13.53)	101.57 (16.60)	3.18	0.19	**0.079**	0.219
36 Months	147 (74/73)	98.20 (13.50)	101.59 (15.73)	3.76	0.24	**0.052**	0.191
48 Months	142 (73/69)	100.17 (12.51)	105.39 (21.36)	3.82	0.18	0.177	0.177
CBCL Total Cutoffs %							
24 Months	161 (81/80)	0.00 (0.00)	0.10 (0.30)	0.08	0.28	**0.007**	**0.010**
36 Months	147 (74/73)	0.01 (0.11)	0.08 (0.28)	0.07	0.24	**0.022**	**0.022**
48 Months	142 (73/69)	0.02 (0.15)	0.18 (0.38)	0.13	0.35	**0.052**	**0.057**
CBCL Externalizing Scores							
24 Months	161 (81/80)	99.10 (13.44)	100.52 (16.28)	2.23	0.14	0.166	0.166
36 Months	147 (74/73)	98.32 (12.49)	101.34 (16.48)	3.43	0.21	**0.057**	0.112
48 Months	142 (73/69)	99.70 (12.94)	106.64 (22.47)	5.60	0.25	0.121	0.150
CBCL Externalizing Cutoffs %							
24 Months	161 (81/80)	0.00 (0.00)	0.04 (0.20)	0.05	0.23	**0.005**	**0.010**
36 Months	147 (74/73)	0.01 (0.11)	0.07 (0.25)	0.05	0.21	**0.027**	**0.027**
48 Months	142 (73/69)	0.00 (0.00)	0.17 (0.38)	0.15	0.39	**0.041**	**0.041**
CBCL Internalizing Scores							
24 Months	161 (81/80)	100.03 (14.78)	101.18 (15.85)	1.42	0.09	0.236	0.371
36 Months	147 (74/73)	98.26 (15.42)	101.29 (14.40)	3.59	0.25	**0.073**	0.155
48 Months	142 (73/69)	101.67 (13.61)	103.17 (17.85)	0.64	0.04	0.321	0.321
CBCL Internalizing Cutoffs %							
24 Months	161 (81/80)	0.02 (0.15)	0.10 (0.30)	0.07	0.22	**0.060**	**0.091**
36 Months	147 (74/73)	0.07 (0.26)	0.07 (0.26)	0.01	0.02	0.465	0.465
48 Months	142 (73/69)	0.03 (0.18)	0.21 (0.41)	0.14	0.35	**0.040**	**0.048**
BITSEA Competency Score							
24 Months	163 (82/81)	99.26 (15.29)	100.24 (13.77)	-0.71	-0.05	0.540	0.540
36 Months	148 (75/73)	100.53 (14.93)	98.56 (14.81)	2.41	0.16	0.165	0.227
BITSEA Competency Cutoffs %							
24 Months	163 (82/81)	0.11 (0.32)	0.08 (0.28)	0.04	0.13	0.201	0.289
36 Months	148 (75/73)	0.13 (0.34)	0.17 (0.38)	-0.06	-0.15	0.745	0.745
BITSEA Problems Score							
24 Months	163 (82/81)	98.61 (11.72)	102.19 (17.63)	3.79	0.21	**0.025**	**0.042**

36 Months	148 (75/73)	99.06 (12.52)	100.42 (17.04)	1.86	0.11	0.165	0.165
BITSEA Problems Cutoffs %							
24 Months	163 (82/81)	0.13 (0.34)	0.24 (0.43)	0.11	0.26	**0.034**	**0.062**
36 Months	148 (75/73)	0.15 (0.36)	0.18 (0.39)	0.03	0.08	0.354	0.354
SDQ Scores @ 48 Months							
Prosocial Behavior Score	143 (73/70)	101.39 (13.98)	95.03 (17.85)	6.40	0.36	**0.021**	**0.059**
Peer Problems	143 (73/70)	98.67 (13.70)	103.87 (19.69)	4.70	0.24	0.136	0.136
SDQ Cutoffs @ 48 Months %							
Prosocial Behavior Score	143 (73/70)	0.08 (0.27)	0.19 (0.39)	0.11	0.28	**0.088**	0.225
Peer Problems	143 (73/70)	0.08 (0.27)	0.17 (0.38)	0.09	0.25	0.170	0.170

Notes: 'N' indicates the sample size. 'IPW M' indicates the IPW-adjusted mean. 'IPW SD' indicates the IPW-adjusted standard deviation. [1] one-tailed (right-sided) conditional *p*-value from individual IPW-adjusted permutation test with 100,000 replications. [2] one-tailed (right-sided) conditional *p*-value from IPW-adjusted stepdown permutation test with 100,000 replications. The conditioning set includes maternal knowledge of child development, parenting self-efficacy, maternal attachment, and maternal consideration of future consequences, as well as child gender. 'Treatment effect' is the difference in means between the high and low treatment group. 'Effect size' is the ratio of the treatment effect to the standard deviation of the low treatment group.

Appendix J

Table J1 *Cognitive skills results – Misreporters removed*

	N (HIGH/LOW)	*IPW* M_{HIGH} (*SD*)	*IPW* M_{LOW} (*SD*)	*IPW Treat. Effect*	*IPW Effect Size*	p^1	p^2
DP3 Scores							
24 Months	120 (55/65)	99.73 (14.04)	98.03 (15.81)	1.70	0.11	0.210	0.210
36 Months	104 (47/57)	101.54 (15.41)	96.37 (14.52)	5.17	0.36	**0.056**	**0.074**
48 Months	104 (49/55)	100.87 (14.75)	95.66 (15.44)	5.21	0.34	**0.054**	0.113
DP3 Cutoffs - Above Average %							
24 Months	120 (55/65)	0.57 (0.50)	0.51 (0.50)	0.06	0.12	0.241	0.241
36 Months	104 (47/57)	0.53 (0.50)	0.37 (0.49)	0.15	0.31	**0.082**	0.131
48 Months	104 (49/55)	0.29 (0.46)	0.12 (0.33)	0.17	0.52	**0.017**	**0.058**
ASQ Communication Scores							
24 Months	166 (82/84)	100.41 (15.05)	100.59 (14.44)	-0.17	-0.01	0.345	0.345
36 Months	150 (75/75)	101.38 (14.17)	97.30 (16.40)	4.08	0.25	**0.073**	**0.091**
48 Months	147 (74/73)	101.10 (13.20)	99.63 (14.94)	1.47	0.10	0.104	0.202
ASQ Communication Cutoffs – Below Average %							
24 Months	120 (55/65)	0.13 (0.34)	0.04 (0.20)	-0.09	-0.44	0.925	0.925
36 Months	104 (48/56)	0.05 (0.22)	0.04 (0.20)	-0.01	-0.03	0.448	0.711
48 Months	104 (49/55)	0.05 (0.22)	0.04 (0.19)	-0.01	-0.07	0.475	0.671
ASQ Problem Solving Scores							
24 Months	120 (55/65)	100.79 (14.56)	97.88 (14.44)	2.91	0.20	0.126	0.213
36 Months	101 (46/55)	102.26 (14.77)	95.61 (15.33)	6.64	0.43	**0.029**	**0.054**
48 Months	104 (49/55)	99.90 (16.41)	99.85 (17.45)	0.06	0.00	0.340	0.340
ASQ Problem Solving Cutoffs – Below Average %							
24 Months	120 (55/65)	0.07 (0.26)	0.14 (0.34)	0.06	0.18	0.147	0.251
36 Months	101 (46/55)	0.11 (0.31)	0.21 (0.41)	0.11	0.26	**0.051**	0.182
48 Months	104 (49/55)	0.08 (0.27)	0.08 (0.28)	0.00	0.01	0.349	0.349

Notes: Participants who scored above 10 on the PSI Defensive Responding Scale at either 24 or 48 months are excluded from the analysis. 'N' indicates the sample size. 'IPW M' indicates the IPW-adjusted mean. 'IPW SD' indicates the IPW-adjusted standard deviation. [1] one-tailed (right-sided) conditional *p*-value from individual IPW-adjusted permutation test with 100,000 replications. [2] one-tailed (right-sided) conditional *p*-value from IPW-adjusted stepdown permutation test with 100,000 replications. 'Treatment effect' is the difference in means between the high and low treatment group. 'Effect size' is the ratio of the treatment effect to the standard deviation of the low treatment group.

Table J2 *Socio-emotional and behavioral development results – misreporters removed*

	N (HIGH/LOW)	*IPW* M_{HIGH} (*SD*)	*IPW* M_{LOW} (*SD*)	*IPW Treat. Effect*	*IPW Effect Size*	p^1	p^2
CBCL Total Scores							
24 Months	119 (55/64)	101.16 (13.18)	104.58 (16.26)	3.43	0.21	0.141	0.203
36 Months	104 (47/57)	100.54 (13.98)	103.79 (16.17)	3.26	0.20	0.153	0.153
48 Months	103 (49/54)	103.05 (12.75)	109.19 (21.70)	6.15	0.28	0.140	0.198
CBCL Total Cutoffs %							
24 Months	119 (55/64)	0.00 (0.00)	0.10 (0.31)	0.10	0.34	**0.001**	**0.004**
36 Months	104 (47/57)	0.02 (0.14)	0.10 (0.31)	0.08	0.27	**0.021**	**0.021**
48 Months	103 (49/54)	0.03 (0.19)	0.22 (0.42)	0.18	0.44	**0.037**	**0.037**
CBCL Externalizing Scores							
24 Months	119 (55/64)	100.71 (12.76)	103.85 (15.18)	3.14	0.21	0.148	0.148
36 Months	104 (47/57)	99.17 (12.74)	104.01 (16.86)	4.84	0.29	**0.039**	**0.095**
48 Months	103 (49/54)	101.90 (12.09)	110.95 (22.83)	9.05	0.40	**0.065**	**0.074**
CBCL Externalizing Cutoffs %							
24 Months	119 (55/64)	0.00 (0.00)	0.04 (0.20)	0.04	0.20	**0.018**	**0.037**
36 Months	104 (47/57)	0.02 (0.14)	0.08 (0.28)	0.07	0.23	**0.025**	**0.025**
48 Months	103 (49/54)	0.00 (0.00)	0.20 (0.41)	0.20	0.50	**0.013**	**0.013**
CBCL Internalizing Scores							
24 Months	119 (55/64)	102.90 (14.19)	102.95 (15.93)	0.05	0.00	0.498	0.498
36 Months	104 (47/57)	101.15 (16.57)	102.45 (15.00)	1.30	0.09	0.439	0.545
48 Months	103 (49/54)	103.88 (14.14)	105.73 (17.79)	1.85	0.10	0.289	0.569
CBCL Internalizing Cutoffs %							
24 Months	119 (55/64)	0.03 (0.17)	0.12 (0.33)	0.09	0.27	**0.050**	**0.082**
36 Months	104 (47/57)	0.11 (0.32)	0.09 (0.29)	-0.02	-0.06	0.616	0.616
48 Months	103 (49/54)	0.05 (0.23)	0.25 (0.44)	0.20	0.45	**0.032**	**0.038**
BITSEA Competency Score							
24 Months	120 (55/65)	96.94 (16.10)	99.35 (13.79)	-2.40	-0.17	0.721	0.721
36 Months	105 (48/57)	98.45 (16.05)	98.43 (14.00)	0.02	0.00	0.492	0.595
BITSEA Competency Cutoffs %							
24 Months	120 (55/65)	0.14 (0.35)	0.10 (0.31)	0.04	0.12	0.261	0.349
36 Months	105 (48/57)	0.17 (0.38)	0.16 (0.37)	0.01	0.03	0.304	0.304
BITSEA Problems Score							
24 Months	120 (55/65)	101.65 (12.29)	104.03 (17.56)	2.38	0.14	0.133	0.213

36 Months	105 (48/57)	102.51 (13.32)	101.88 (18.21)	-0.63	-0.03	0.454	0.454
BITSEA Problems Cutoffs %							
24 Months	120 (55/65)	0.19 (0.39)	0.28 (0.45)	0.10	0.21	0.119	0.201
36 Months	105 (48/57)	0.20 (0.41)	0.23 (0.42)	0.03	0.06	0.361	0.361
SDQ Scores @ 48 Months							
Prosocial Behavior Score	104 (49/55)	100.13 (14.35)	92.87 (17.62)	7.26	0.41	**0.039**	**0.096**
Peer Problems	104 (49/55)	100.75 (15.28)	105.79 (19.87)	5.04	0.25	0.220	0.220
SDQ Cutoffs @ 48 Months %							
Prosocial Behavior Score	104 (49/55)	0.09 (0.29)	0.20 (0.41)	0.11	0.27	0.167	0.365
Peer Problems	104 (49/55)	0.11 (0.31)	0.19 (0.40)	0.09	0.22	0.293	0.293

Note: Participants who scored above 10 on the PSI Defensive Responding Scale at either 24 or 48 months are excluded from the analysis. 'N' indicates the sample size. 'IPW M' indicates the IPW-adjusted mean. 'IPW SD' indicates the IPW-adjusted standard deviation. [1]one-tailed (right-sided) conditional *p*-value from individual IPW-adjusted permutation test with 100,000 replications. [2] one-tailed (right-sided) conditional *p*-value from IPW-adjusted stepdown permutation test with 100,000 replications. 'Treatment effect' is the difference in means between the high and low treatment group. 'Effect size' is the ratio of the treatment effect to the standard deviation of the low treatment group.