Growing Pains: The Preparing for Life Trial at Age 14*

By ORLA DOYLE

University College Dublin

DRAFT

JANUARY 2025

ABSTRACT

This paper examines the impact of *Preparing for Life* (*PFL*), an Irish prenatally commencing home visiting program, ten years after the intervention ended. The intervention involved bi-weekly visits from a trained home visitors from pregnancy until school entry to support parents around child development and parenting. Previous reports of the *PFL* trial, find that the program was effective in boosting children's cognitive skills, with smaller effects on some dimensions of health and socio-emotional skills. The Age 14 Follow-up finds the program has a sustained and long-term effect on children's cognitive development, with large effect sizes of 0.70 SDs. Significant effects are also found on working memory, attention, and educational expectations, however there are relatively few effects on health or socio-emotional outcomes. There is some evidence that the program reduced children's waist-to-height ratio, and improved parental-child relationships. All results are estimating using permutation-based hypothesis testing which account for attrition using inverse probability weighting and multiple hypothesis testing using the stepdown procedure. While 43% of the original sample recruited during pregnancy participated at the Age 14 Follow-up, the treatment groups are still balanced on all key baseline characteristics. This is one of the few experimental home visiting programs that tracks participants into adolescence and finds evidence that *PFL* continues to have a significant impact on important dimensions of children's skills ten years after the families have finished the programme.

JEL Classification: C93, I26, J13

* The evaluation of the *Preparing for Life* programme was funded by the Northside Partnership through the Department of Children and Youth Affairs and The Atlantic Philanthropies. Additional support was provided by Tusla, the Community Foundation of Ireland and the HSE for the Age 14 Follow-up. We would like to thank all those who supported this research, especially the participating families and community organizations and the *PFL* intervention staff. Thanks to the Early Childhood Research Team at the UCD Geary Institute for Public Policy who significantly contributed to this study. 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

1. Introduction

A growing body of evidence highlights the crucial role of early-life circumstances in shaping the skills needed to thrive later in life. Exposure to adverse prenatal and postnatal environments is linked to poorer health, education, and labor market outcomes (Cunha et al., 2006; Heckman, 2006; Almond, Currie, and Duque, 2017). Early childhood intervention programs aimed at mitigating or counteracting these adverse conditions are increasingly recognized as a viable strategy (OECD, 2016). Such interventions are seen as effective from both a biological and economic standpoint (Doyle et al., 2009). Biologically, research indicates that brain plasticity and neurogenesis are most pronounced during the early years, particularly from pregnancy to age three (Thompson and Nelson, 2001; Knudsen et al., 2006). Economically, early investment allows for longer-term benefits from the resulting enhanced skill set (Heckman and Kautz, 2014). This paper assesses the impact of an Irish early childhood intervention program, *Preparing for Life (PFL)* that begins prenatally and focuses on parents as the primary agents of change, when children reach the age of 14.

Previous reports of the *PFL* program have identified some important effects at earlier ages.[1] Doyle (2020), based on data collected at the end of the trial, found that the program had a large 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

[1] Doyle (2013) describes the design of the *PFL* evaluation. Doyle *et al.* (2014) examine the impact of the *PFL* programme on birth outcomes utilising hospital data and identify a significant treatment effect regarding a reduction in caesarean sections, yet no impact on neonatal outcomes. Doyle *et al.* (2017a) examine the impact of the programme on parent reported cognitive and non-cognitive skills at 6, 12, and 18 months, and find no evidence of effects, yet there are significant improvements in the quality of the home environment. Doyle *et al.* (2015) examine the impact of the programme on child health measured at 6, 12, 18, 24, and 36 months and identify a number of significant treatment effects at 24 months in terms of reducing the incidence of asthma, chest infections, and health problems. O'Sullivan, Fitzpatrick and Doyle (2017) examine the impact of the programme on dietary intake at 12, 18, 24, and 36 months, and its mediating effect on cognitive development at 24 and 36 months; and find evidence of improved nutrition at 24 months in terms of increased protein intake. Doyle *et al.* (2017b) examine the impact of the programme on maternal wellbeing using daily data collected over a 24-hour period using the Day Reconstruction Method, finding little evidence of effects on maternal wellbeing. Cote *et al.* (2018) investigate whether the impact of the programme varied according to children's developmental trajectories and find a positive impact on trajectories of cognitive development and number of health clinic visits for all children, whereas positive impacts on externalizing behaviour problems are restricted to children with the most severe problems. Doyle (2020) examines the impact of the programme on children's cognitive and socio-emotional skills at the end of the programme and finds significant effects on all dimensions of children's skills, with large effects for cognitive ability. Coy and Doyle (2024) examine the impact of the programme on the development of health capital in the first 5 years of life. Modest effects on child health were found, mainly driven by reduced hospital attendances attributed to improved earlier health. The low treatment group generated almost twice the hospital costs of the high treatment group, with a mean difference of €1,359 per child, however by age four, no treatment effects were found on any health outcome. Doyle (2024) reports the results of the age 9 follow-up, focusing on cognitive and socio-emotional outcomes. Ssignificant treatment effects on cognitive skills (0.55SD) and school achievement tests (0.30-0.54SD) were found, however, there was no impact on socio-emotional skills and there is little evidence of treatment heterogeneity by gender, birth order, or distribution of ability.

(SD), indicating the malleability of IQ in the early years. Gains were found across all dimensions of cognitive skill including spatial ability, pictorial reasoning, and language ability. Although weaker, the program also impacted several dimensions of children's non-cognitive skills including externalizing problems such as aggressive behaviour, and prosocial behaviour such as helping other children. The program also had an impact on child health (Coy and Doyle, 2024). It reduced the amount of hospital services the children used and improved how families used these services.

The first follow-up of the PFL cohort at age nine found that the program continued to have an impact on cognitive scores, with effect sizes of 0.55 SD on general conceptual ability and 0.30 SD and 0.54 SD on achievement tests of reading and math respectively in second grade (Doyle, 2024). These results could not be attributed to differences in school quality across the high and low treatment groups as there was no evidence that the program impacted school choice. The program, however, had no impact on absenteeism or the use of school resources, and the significant treatment effects observed for children's socio-emotional skills at age four were no longer present at age nine. A mediation analysis found that between 35-38% of the treatment effects on age nine test scores was explained by improvements in early parental beliefs, stimulation, and health investments. The size of these treatment effects identified at the end of the trial and age nine exceed current meta-analytic estimations based on the home visiting literature (e.g. Sweet and Appelbaum 2004; Gomby 2005; Filene *et al.* 2013).

The impact of the PFL program may be attributed to both its duration and intensity. By providing a five-year intervention — covering the critical first 2,000 days of a child's life — the program offered early and sustained support to families during a pivotal period of development. This approach aligns with the "technology of skill formation" framework proposed by Cunha and Heckman (2007), which suggests that early skills serve as a foundation for the development of more advanced skills through a process of self-productivity. In turn, this enhances the effectiveness of later investments through dynamic complementarity (Cunha, Heckman, and Schennach, 2010; Heckman and Mosso, 2014). While genetic factors play a role in skill development (Nisbett et al., 2012), research indicates that environmental conditions can shape and enhance these skills (Weaver et al., 2004).

Empirical research highlights several key aspects of the home environment that predict children's skills, including the quality of the home setting (Todd and Wolpin, 2007), parenting skills (Dooley and Stewart, 2007; Fiorini and Keane, 2014), and parental stimulation (Miller et al., 2014). Yet, socio-economic inequalities in the quality of the home

environment exist, with disadvantaged families often facing financial constraints that limit their capacity to invest in their children. Evidence suggests that parents from low socioeconomic backgrounds are more likely to adopt less effective parenting styles and behaviours (Cunha, Elo, and Culhane, 2013), such as permissive or harsh parenting (Bradley and Corwyn, 2002), and to provide fewer stimulating materials and experiences for their children (Bradley et al., 1989). This may, in part, stem from a knowledge gap regarding optimal parenting practices. Cunha et al. (2013) point to a lack of parenting knowledge and differing beliefs about the importance of parenting among parents from low socioeconomic backgrounds. Disadvantaged homes are also less likely to offer pre-academic stimulation, such as reading to children or helping them recognize letters (Miller et al., 2014). To address these issues, the PFL program sought to enhance parenting knowledge and promote developmentally appropriate activities, thereby mitigating the negative effects of socioeconomic disadvantage on children's skill development. Thus, the program targeted both behavioural frictions and informational frictions.

This paper examines the long-term impact of the *PFL* programme now that the cohort have reached adolescence. It focuses on outcomes measured using directly assessed tests and a self-completion questionnaire. In addition, using pre-existing data from Ireland (Growing Up in Ireland; Health Behaviour of School Age Children) and the UK (Millennium Cohort Study), also collected during adolescence, it situates the PFL cohort alongside these representative cohorts. The remainder of the paper is structured as follows. Section 2 reviews the literature assessing the impact of home visiting programs. Section 3 describes the design of the original study and the Age 14 Follow-up. Section 4 outlines the statistical methods that are used to estimate the results. Section 5 presents the main results. Finally, Section 6 concludes.

2. Literature

The effectiveness of home visiting programs in the short term is well-documented. A meta-analysis of 60 home visiting programs by Sweet and Appelbaum (2004) reported an average effect size (ES)[2] of 0.18 for cognitive skill improvements and 0.10 for non-cognitive skills. Subsequent reviews found similar results, with Miller, Maguire, and Macdonald (2011)

[2] The effect size (ES) represents the magnitude or the size of the difference between the treatment and controls group. While the p-value allows the reader to determine whether or not there is a statistically significant difference between the groups, it does not indicate the strength of the difference. Effect sizes are usually expressed in terms of standard deviation of the outcome variable. Effect sizes are calculated using Cohen's d, where effect sizes of 0.0 to 0.2 are considered small, 0.2 to 0.8 medium, and greater than 0.8 large.

reporting an average ES of 0.30 for cognitive skills across 7 studies, and Filene et al. (2013) reporting an ES of 0.25 based on 51 studies. Collectively, these findings suggest that home visiting programs generally produce small to modest improvements in both cognitive and non-cognitive skills (Gomby, 2005; Peacock et al., 2013; Avellar et al., 2016).

Research regarding their effectivness beyond the lifetime of the intervention has found mixed results. A study by Bailey et al. (2017) examined 67 high-quality early intervention programs in the U.S., including some home visiting programs, and identified a general pattern of declining effect sizes. While the average effect size (ES) at the end of the intervention was 0.23, this dropped to 0.10 by the end of the first year post-intervention and to 0.05 within one to two years after the program concluded. Focusing specifically on home visiting programs, findings on medium- and long-term effects are varied. Bierman et al. (2017) found that children who participated in Early Head Start (EHS) demonstrated improved cognitive ability as well as reading and language skills at ages 7 to 9. Studies on the Healthy Families America (HFA) programme also revealed significant impacts, with children more likely to be enrolled in gifted programs, less likely to require special education, and more likely to excel academically at ages 6 to 7 (DuMont et al., 2010; Kirkland and Mitchell-Herzfield, 2012). In terms of socio-emotional outcomes, two EHS studies found evidence of positive effects on children's behavior, perceived competence, and learning approaches at ages 5, 7, and 9 (Bierman et al., 2017; Chazan-Cohen, Raikes, and Vogel, 2013). Additionally, the Nurse-Family Partnership (NFP) programme reported a reduction in internalizing disorders at age 12 (Kitzman et al., 2010).

Research on the long-term effects of home visiting programs provides more compelling evidence. The Jamaica Study, which involved weekly home visits for children aged 9 to 24 months, found an initial IQ effect size of 0.88 at the end of the intervention. Although this effect diminished by age 7, it re-emerged at ages 11, 17, and 22, with effect sizes ranging from 0.40 to 0.60 (Grantham-McGregor and Smith, 2016). The Abecedarian Program, which provided centre-based care and home visits from infancy to age 5, recorded an initial IQ effect size of 0.74. While this declined to 0.37 on average at ages 8, 12, 15, and 21, the effects were still sustained (Campbell et al., 2001). The NFP programme has also demonstrated significant long-term impacts. Cognitive effect sizes of 0.22 to 0.27 were observed at age 6 (for both boys and girls) and at age 12 (for boys only) (Heckman et al., 2017). At age 18, children of mothers with low psychological resources showed higher receptive language (ES = 0.24) and math achievement (ES = 0.38) (Kitzman et al., 2019). By age 19, girls in the treatment group had fewer children, were less likely to receive Medicaid,

and were less involved in crime (Eckenrode et al., 2010). Even in cases where cognitive effects faded over time, other long-term benefits, such as reductions in criminal behaviour and receipt of social welfare, were observed (Campbell et al., 2014; Heckman et al., 2017).

2.1 Studies with follow-ups during adolescence

Table 1 summarizes the literature on home visiting programs with follow-ups during the same period of adolescence as in the current paper (between 12-16 years). Overall, there are very few studies that test for the sustained effects of home visiting programs during this period. Most studies either stop collecting data directly after the intervention; only revisit the families during adulthood (often using administrative data); or the interventions are still in the infancy thus long-term follow up is not yet possible. The only studies that conducted assessments during adolescence are studies of the NFP program. Follow-ups were conducted at ages 12 -16 for the Memphis trial and age 15 for the Elmira trial in 11 separate papers. With few exceptions there were no impact on child outcomes.

Of the five papers included in Table 1, three found no effects, and two only found effects on a small number of the outcomes assessed. Kitzman et al. (2010) found effects in the Memphis trial at age 12 on the substance use and internalising disorders, but no effects on cognitive scores, achievement tests, other behavioural problems, and educational outcomes. Olds et al. (1998) found effects in the Elmira trial at age 15 on convictions and probation violations, but no effects on substance use, risky behaviours, behavioural problems, anti-social behaviour, or school behaviour outcomes. Note that two only of the studies (Kitman et al., 2010; Sidora-Arcoleo *et al.,* 2010) assessed cognitive skills, and no effects were found. The majority of the studies assessed socio-emotional outcomes (focusing on behavioural problems), and significant effects were only identified in one. Thus, the main takeaway from the sparce home visiting literature which follows the sample into adolescence is that effects on child outcomes are minimal.

Table 1 *Impact of Home Visiting Programmes on Child Outcomes from Ages 12-16*

Author	Sample Size	Programme	Measures	Significant Finding	Effect	Age (years)	% of original sample retained
Kitzman *et al*. (2010)	635 children	Nurse Family Partnership (Memphis)	GPA, Peabody Individual Achievement Tests, Leiter-R sustained attention test, Group achievement test scores, Placement in special education, ever retained in a grade, conduct grades, externalising disorders, internalising disorders, total problems, Days of substance use in the last 30 days	Incidence of substance use, used cigarettes, alcohol or marijuana in the last 30 days, internalising disorders	Favourable	12	80% parent interviews, 76% child interviews, 85% school records
Sidora-Arcoleo *et al.* (2010)	721 mother and child dyads	Nurse Family Partnership (Memphis)	Peabody Picture Vocabulary Test-Revised, physical aggression (CBCL)	None	None	6-12 years	Not reported
Enoch *et al.* (2016)	559 children	Nurse Family Partnership (Memphis)	Composite externalizing disorders continuous total scores (CBCL)	None	None	12	
Eckenrode *et al.* (2001)	228	Nurse Family Partnership (Elmira)	Number of early onset of problem behaviors & Percentage abused or neglected.	None	None	15	
Olds et al. (1998)	245 children	Nurse Family Partnership (Elmira)	Alcohol and drug impairment, Ever pregnant or made someone pregnant, Incidence of sex partners, cigarettes smoked per day, days drank alcohol, days used drugs, times ran away, Number of acting out problems, Number of externalizing problems , Number of internalizing problems , Number of minor antisocial acts, Ever was person in-need of supervision, Incidence of arrests, Incidence–convictions and probation violations, Incidence–long-term school suspensions, Incidence–sent to youth corrections , Incidence–short-term school suspensions, Number of major delinquent acts	Incidence–convictions and probation violations	Favourable	15	

3. Study Description

In an effort to break the intergenerational cycle of disadvantage, *PFL* was developed as part of the Irish Government's and The Atlantic Philanthropies' Prevention and Early Intervention Programme (Office of the Minister for Children and Youth Affairs 2008). The program was developed by 28 local agencies and community groups who collaborated to design an evidence-based intervention tailored to meet the needs of the local community. The study took place between 2008 and 2015 in a highly disadvantaged Dublin community. The *PFL* program was developed to reduce socioeconomic inequalities in children's skills by working directly with parents to improve their knowledge of child development and parenting.

3.1 Initial recruitment and randomisation

Recruitment into the *PFL* program 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. The inclusion criteria included all pregnant women residing in the designated *PFL* catchment area during this period, regardless of their social or family circumstances. Based on estimates of a two to five point difference on cognitive development scores (i.e., average standardised effect size of 0.18) from a meta-analysis of home visiting programmes (Sweet and Appelbaum 2004), a sample size of approximately 117 in each group was required to power the study.

In total, 233 participants were recruited by the *PFL* recruitment officers. This represents a recruitment rate of 52% based on the number of live births during the recruitment period. Of those who joined the programme, an unconditional probability randomisation procedure, with no stratification, assigned 115 to a high treatment group and 118 to a low treatment group. Baseline data from 205 participants (representing 90% of the high treatment group and 86% of the low treatment group) was collected after randomisation yet prior to treatment delivery.[3] The baseline variables include 117 measures of socio-demographics, physical and mental health, IQ, parenting attitudes, and self-control, among others. To assess the effectiveness of the randomisation procedure, the baseline characteristics of the high and low treatment groups were compared using separate

[3] Of the 233 randomly assigned participants, two (high=1; low=1) miscarried, 19 (high=6; low=13) withdrew from the programme before the baseline assessment, and seven (high=4; low=3) 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 noted.

permutation tests. At the 10% significance level, the two groups differed on 7.7% (9/117) of measures, which is consistent with pure chance and indicates the success of the randomisation process (see Doyle and *PFL* Evaluation Team 2010).

3.2 Treatment

Figure 1 below describes the supports provided to the high and low treatment groups. The high treatment consisted of three primary components - a five year home visiting program, a baby massage course, and the Triple P Positive Parenting Program. The treatments are built upon the theories of human attachment (Bowlby 1969), socio-ecological development (Bronfenbrenner 1979), and social-learning (Bandura 1977). The home visiting program 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. Twice monthly home visits of approximately one hour were prescribed with home visitors from different professional backgrounds including education, social care, and youth studies. The visitors were hired to deliver the program on a full-time basis and they received extensive training prior to treatment delivery. Supervision took place on a monthly basis to ensure fidelity to the programme model, and families were allocated the same home visitor over the course of the intervention where possible.

Each home visit was structured around *PFL*-developed 'Tip Sheets' which included information on pregnancy, parenting, health, and development. The 210 Tip Sheets were developed by the *PFL* implementation team based on pre-existing and publicly available information. The home visitors 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 intervention was delivered using techniques such as 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 home visitor would then guide the parent through the Tip Sheet(s) selected for that visit and following this, new goals would be agreed. The Tip Sheets typically targeted multiple aspects of development.

Participants in the high treatment group were also encouraged to take part in a baby massage course in the first year, which consisted of five two-hour individual or group

sessions delivered by the mentors. The purpose of these classes was to equip parents with baby massage skills and to emphasise the importance of early reciprocal interactions and communication between parents and infants.

When the *PFL* children were between two and three years old, the high treatment group was invited to participate in the *Triple P Positive Parenting Program* (Sanders, Markie-Dadds, and Turner 2003) which was delivered by the home visitors. 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 behavioural outcomes (Sanders *et al.* 2014). The high treatment participants were encouraged to take part in five two-hour group discussion sessions and three phone calls. The home visitors also used the *Triple P* principles and techniques when delivering the home visits to ensure consistent messaging across the program components.

In addition to the standard services available to pregnant women and young children, 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. They also received four book packs containing six to 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 of the child. The low treatment group also had access to a *PFL* support worker who could help them avail of community services if needed, and this function was provided by the home visitor for the high treatment group. Note that the low treatment group did not receive the home visiting programme, Tip Sheets, baby massage classes, or the *Triple P* program.

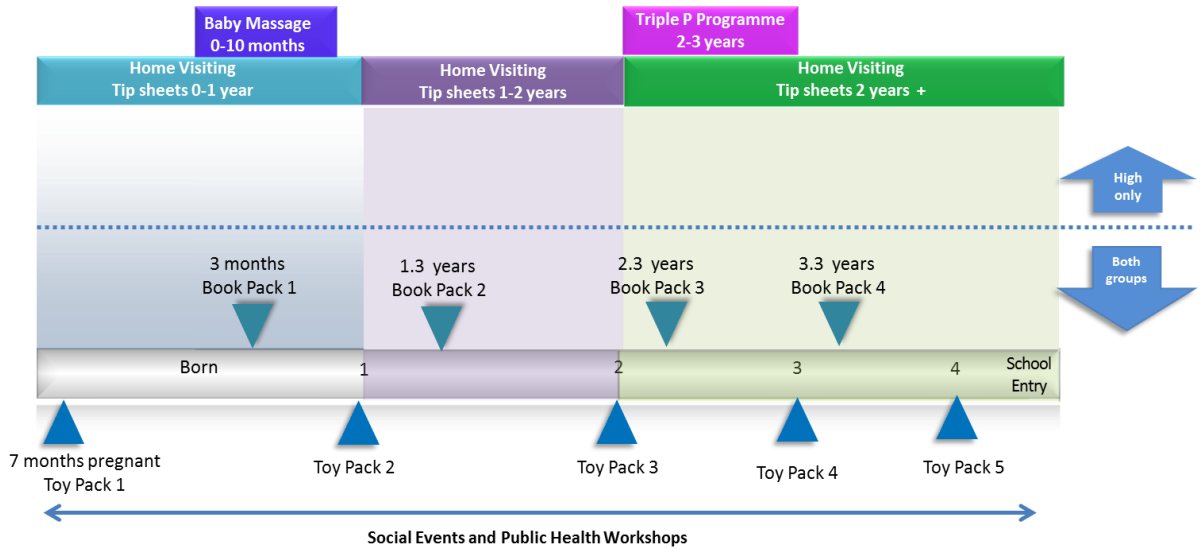


Figure 1 *Timing of PFL treatments*

3.3 Design of the Age 14 Follow-up

Recruitment

The Age 14 follow-up study sought to include as many of the original *PFL* participants as possible. As such, all families that were recruited and randomized in the original *PFL* study between 2008 and 2010 were eligible to take part. However, participants who had officially dropped out or left the study due to death or miscarriage were not contacted. Multiple steps were taken to make contact and reconsent as many families as possible. First, six months prior to formally starting the recruitment process, the *PFL* implementation team attempted to make contact with all families who had not formally withdrawn from the study during previous waves. This informal contact was to inform the families of the upcoming Age 14 Follow-up study and to update their contact details. Second, all families were invited to attend a *PFL* 'Age 14 Birthday Celebration' on 8th October 2023. At this event, which involved talks, games, activities, and food, families were asked if they were interested in hearing about the Age 14 Follow-up study and if they were happy to speak to a researcher about participating. If a family agreed to meet with the researcher, they were informed about the upcoming study and their consent to participate was sought. A total of 25 families were recruited at this event. Third, families who did not attend the event, yet agreed to be contacted by the research team at previous assessments, were attempted to be contacted using all available contact data available (e.g., phone, email, house address etc.). Fourth, the *PFL*

implementation team also followed-up with all potential families to inform them of the study and put them in touch with the research team. Families who left the original study area were still invited to participate. 74 additional families were recruited using a combination of these methods. In total, 99 families were recruited between October 2023 and September 2024.

During the recruitment process, eligible participants were provided with information about the follow-up study. Researchers explained the procedures, articulating all relevant study information, consent details, and participant rights. At this time, we sought consent from the parent for their child to take part in the study, and assent from the child to participate. Parents were provided with an information sheet detailing the study, and children were provided with an age appropriate information booklet. Note, in a few cases, the parent consented for their child to take part in the study, but the child did not assent, thus in these cases, the assessment did not take place.

Age 14 Sample: Age and Gender

The average age of the *PFL* cohort at the time of data collection was 14.3 years, with the youngest participant being aged 12.11 and the oldest aged 16.2. Importantly, there were no statistically significant differences in the age of the participants in the high and low treatment groups at the time of the interview (high treatment=14.4 years, low treatment=14.3 years; *p*-value=0.620). There were also no differences in participant gender – within the high treatment group, 56% of the sample were girls and 44% were boys, and within the low treatment group, 64% of the sample were girls and 36% were boys; *p*-value=0.666). Thus, the sample is balanced in terms of age and gender.

Data Collection Procedure

Data for the Age 14 Follow-up were collected between January and September 2024. The majority of the assessment took place in the participant's secondary school (75%), and the remainder either took place in the participant's house or the village centre. To minimize detection bias, all assessments were conducted by trained researchers who were blind to the treatment condition. There were four assessment components – direct assessment of cognitive skills and executive functioning, a self-completion questionnaire, height and waist measurements, and a saliva sample. These are described in more detail below. When the assessments took place in schools, the assessments were divided into two sessions to reduce fatigue, one before lunch and the other after lunch. Each session lasted about one hour. The

participants received a €20 One4All voucher for each session to compensate them for their time.

Session 1

The participants were invited to take part in assessments of their cognitive skills using the British Ability Scales III: School Age Battery (BAS III; Elliott, Smith and McCulloch 2011) (the same assessment used at the Age 9 Follow-up). The BAS III yields an overall score reflecting general cognitive ability (General Conceptual Ability, GCA), as well as three standardised scores for Verbal Ability, Pictorial Reasoning Ability, and Spatial Ability. The children also conducted several tasks assessing self-regulation/executive functions using the National Institutes of Health Toolbox for Assessment of Neurological and Behavioral Function Cognition Battery (NIH Toolbox; Zelazo and Bauer 2013) (e.g., the Flanker task to assess inhibitory control, the Dimensional Change Card Sort task to assess attention flexibility, and the List Sorting task to assess working memory).

Session 2

The participants were invited to complete a self-completion questionnaire on an iPad. The survey was programmed in Qualtrics. The participant completed the survey on their own, however a researcher was present and the participant was informed that they could ask the researcher any questions they had about the survey or if they did not understand a question. In order to ensure the participant did not lose interest, at two points during the survey, they were prompted on the screen to hand the iPad to the researcher. At the first 'break', the researcher measured the participants' height and waist circumference. At the second 'break', the salvia sample was taken. The survey included a set of standardised instruments and single-item questions capturing measures of socio-emotional development, mental and physical health, health behaviours (diet, substance use), puberty development, self-esteem, attitudes towards school, school absences, educational expectations, life satisfaction, relationship with parents, attitudes towards antisocial behaviour, daily activities, and risk and time preferences.

3.6 Age 14 Follow-up Sample and Attrition

Figure 2 depicts the families' participation in the study between program entry and the Age 14 Follow-up. At the follow-up, data were collected for 99 of the original 233 randomly

assigned participants, representing an overall retention rate of 43%.[4] Despite the substantial time and effort invested in recruitment efforts by both the research and implementation teams, the participation rate was lower than that achieved at Age 9 (50%). At Age 9, a significantly higher proportion of the high treatment group participated compared to the low treatment group (59% vs 42%). However, at Age 14, the level of attrition was relatively similar in both groups (45% vs 40%) and importantly there were no statistically significant differences in the likelihood of participating in the follow-up based on treatment status (p=0.406). An analysis of attrition between Age 9 and 14 shows that a higher proportion of the high treatment group dropped out of the study than the low treatment group at this time point.

It is important to note that the composition of the samples are not the same at both waves e.g., some participants at Age 14 did not participate at Age 9, and vice-versa. For example, 117 participants took part in the Age 9 assessment and 99 took part in the Age 14 assessment. While the majority of participants took part in both assessments (n=82), 34 participants who took part at Age 9 did not participate at Age 14, and 17 participants who took part at Age 14 did not participate at Age 9. In order to test whether participants with certain characteristics were more likely to drop out, we compared the Age 9 cognitive scores of those who participated at Age 9 and Age 14 (mean=84.8) to those who did not participate at Age 14 (mean=83.9), and we found no significant difference (p=0.747) across the groups. In addition, we compared the Age 14 cognitive scores of those who participated at Age 9 and Age 14 (mean=83.1) to those who did not participate at Age 9 (mean=78.5), and also found no significant difference (p=0.190), however, those who re-joined the study at Age 14 had somewhat lower scores. Overall, this implies that the type of participants who took part in the Age 14 assessment are similar to those who took part at Age 9. Therefore, any differences in the results across waves (if found) may not be attributed to differences in the types of participants who took part.

[4] Note that 99 participants completed the direct assessments and 99 completed the self-completion questionnaire, however the samples are not identical. There was one participant who completed the direct assessment, but did not complete the questionnaire, and another participant who completed the questionnaire but not the direct assessment. Therefore, separate weights are used for both samples in the Inverse Probability Weighting procedure.

Figure 2 *Participant flow*

A re-examination of the comparability of the high and low treatment groups at baseline using the Age 14 estimation BAS sample shows that the two groups differ on 11% (13/117) of measures. Using the 10% cut-off level, we would expect 10% of the measures to be statistically significant at random, thus these results are largely consistent with pure chance and indicate that the groups remain largely balanced at the Age14 Follow-up. Table 2 compares the Age 14 participants in the high and low treatment group for a selection of baseline variables. It shows that there are no statistically significant differences across the two groups on all but one of the key socio-demographic and health factors assessed. Specifically, high treatment children who participated in the Age 14 follow-up are more likely to come from families where the mother was married during pregnancy.

Table 2 *Baseline comparison of the Age 14 high and low treatment groups*

	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	p^1
Age	26.69 (5.78)	25.81 (6.03)	0.462
Married	0.23 (0.43)	0.10 (0.31)	**0.090**
No. of children	1.94 (1.19)	1.98 (1.17)	0.891
First time mother	0.48 (0.50)	0.47 (0.50)	0.921
Low education (left $\leq$ age 16)	0.27 (0.45)	0.23 (0.43)	0.779
Weschler Abbreviated Scale of Intelligence (WASI)	85.25 (11.75)	81.61 (14.49)	0.178
Employed	0.48 (0.50)	0.45 (0.50)	0.727
Resides in social housing	0.54 (0.50)	0.55 (0.50)	0.952
Medical card	0.58 (0.50)	0.64 (0.49)	0.462
Prior physical health condition	0.77 (0.43)	0.70 (0.46)	0.453
Prior mental health condition	0.31 (0.47)	0.28 (0.45)	0.767
Smoking during pregnancy	0.40 (0.50)	0.47 (0.50)	0.484
Drinking alcohol during pregnancy	0.33 (0.47)	0.32 (0.47)	0.922
N	52	47	

Notes: All baseline measures were assessed during pregnancy prior to treatment delivery except for WASI which was assessed at 3 months postpartum. Baseline data are missing for four participants in the age 14 assessment. [1] two-tailed *p*-values calculated from permutation tests with 100,000 replications.

Although the estimation samples are largely balanced in terms of baseline characteristics, it is important to test for differential attrition in the high and low treatment group. To investigate this, the factors predicting participation in the Age 14 assessment were tested using bivariate tests with 49 baseline measures. Analyses were conducted separately for the high and low treatment groups to allow for differential attrition processes. In general, there is some evidence of differential attrition, with 11 (22%) baseline measures predicting

attrition from the high treatment group, and 9 (18%) predicting attrition from the low treatment group (in two-tailed tests, using the 10% significance level).[5]

The factors predicting attrition from both groups differ somewhat, however in both cases we find that, consistent with the home visiting literature (see Roggman *et al.* 2008), families who did not take part in the Age14 assessment had more risk factors at baseline, for example, they were less likely to be employed during pregnancy, had lower levels of education and IQ, were younger, and were less likely to smoke during pregnancy. However, within the low treatment group, the children of parents who were not married at baseline and experienced more domestic risk factors were more likely to participate. Figure 2 shows the baseline characteristics that are significant predictors of retention in the high and low treatment group.

Figure 2 *Baseline factors predicting retention at Age 14*

High Treatment Group (11/49)	**Low Treatment Group (9/49)**
• WASI IQ Score (+) • Married (+) • Use of health services (-) • Employed (+) • Smoke during pregnancy (-) • Age (+) • Vulnerable attachment style questionnaire score (-) • Drink during pregnancy (+) • Birth control (-) • Satisfied with neighbourhood (+) • WHO5 well-being score (+)	• Low education (<16) (-) • TIPI Openness personality (+) • Regular exercise (+) • Non traveller (+) • Impaired by illness (+) • Married (-) • Birth control (+) • Domestic risk factors (+) • Consideration of future consequence scale score (-)

Table 3 compares a selection of baseline characteristics of those who participated in the Age 14 assessment - 'stayers' - to those who did not - 'non-stayers'. It shows that high treatment participants who completed the Age 14 assessment had parents who were older, and more likely to be married at baseline, employed, and drink at baseline, however they were less likely to smoke. Low treatment participants who completed the Age 14 assessment had parents who were less likely to be married and to have low education. In order to account

[5] This analysis is based on the BAS sample. For the self-completion questionnaire sample, 12 (24%) baseline measures predicting attrition from the high treatment group, and 9 (18%) predicting attrition from the low treatment group.

for differential attrition across the high and low treatment groups, treatment effects are estimated using the Inverse Probability Weighting procedure detailed in the Methods section.

Table 3 *Baseline characteristics predicting attrition from the Age 14 sample*

	High Treatment Group			*Low Treatment Group*		
	M_{STAYER} (*SD*)	$M_{NON\text{-}STAYER}$ (*SD*)	p^1	M_{STAYER} (*SD*)	$M_{NON\text{-}STAYER}$ (*SD*)	p^1
Age	26.69 (5.78)	24.23 (5.71)	**0.031**	25.81 (6.03)	24.85 (5.97)	0.425
Married	0.23 (0.43)	0.06 (0.24)	**0.011**	0.11 (0.31)	0.24 (0.43)	**0.067**
First time mother	0.48 (0.50)	0.60 (0.50)	0.236	0.47 (0.50)	0.52 (0.50)	0.612
No. of children	1.94 (1.19)	1.94 (1.42)	1.000	1.98 (1.17)	1.85 (1.14)	0.593
Low education (left $\leq$ age 16)	0.27 (0.45)	0.40 (0.50)	0.159	0.23 (0.43)	0.54 (0.50)	**0.001**
Weschler Abbreviated Scale of Intelligence (WASI)	80.90 (11.32)	77.38 (11.41)	0.116	79.04 (12.78)	77.41 (10.46)	0.476
Employed	0.48 (0.50)	0.25 (0.44)	**0.012**	0.45 (0.50)	0.35 (0.48)	0.327
Medical card	0.58 (0.50)	0.62 (0.49)	0.693	0.64 (0.49)	0.69 (0.47)	0.617
Prior physical health condition	0.77 (0.43)	0.73 (0.45)	0.656	0.70 (0.46)	0.56 (0.50)	0.137
Prior mental health condition	0.31 (0.47)	0.25 (0.44)	0.514	0.28 (0.45)	0.20 (0.41)	0.387
Smoking during pregnancy	0.40 (0.50)	0.62 (0.49)	**0.029**	0.47 (0.50)	0.48 (0.50)	0.919
Drinking alcohol during pregnancy	0.33 (0.47)	0.17 (0.38)	**0.070**	0.30 (0.46)	0.24 (0.43)	0.538
N		52			47	

Notes: All baseline measures were assessed during pregnancy prior to treatment delivery except for WASI which was assessed at 3 months postpartum. Baseline data are missing for four participants in the age 9 assessment. [1] two-tailed *p*-values calculated from permutation tests with 100,000 replications.

3.7 Power

Attrition has two consequences. First, as discussed above, it can reduce the comparability of the high and low treatment groups which means that the assumption of baseline equivalence no longer holds i.e., the results could be attributed to underlying differences between the two groups rather than differences 'caused' by the programme. However, another consequence of attrition is reduced statistical power. Power is the ability to identify statistically significant differences between the high and low treatment group if the program really did have an impact. More formally, it is the probability of not rejecting the null hypothesis, even though

the null hypothesis is not true. A Type II Error is the probability of (falsely) concluding that there is no treatment effect, even when there is, and power is the probability of avoiding Type II errors. Traditionally, we aim for 80% power in experiments. This means that there is a 20% chance of making Type II error i.e. 20% of the time you will not be able to reject the null hypothesis of a zero treatment effect despite there being a significant effect.

One of the key factors in determining the power of the experiment is sample size. In general, for a given expected effect size, power is higher when the sample size is larger. Thus, as the sample size falls over time with attrition, the ability to detect statistically significant effects falls. Thus, assuming a power of 80%, with a sample size of 99 (53 = high treatment, 47 = low treatment), a power analysis was conducted to determine what size effects we are powered to detect e.g. how large does the difference in outcomes between the high and low treatment group need to be in order to identify a statistically significant result. This analysis showed that we have the power to detect effects of 0.50 standard deviations. For context, the effect size we found at Age 9 on children's cognitive skills was 0.55 standard deviations. Thus, if we observe a similar effect at Age 14 on cognitive skills, we will be able to conclude that the program still has a statistically significant effect on cognitive skills. However, for smaller effect sizes, which may be practically meaningful, we may not be able to detect statistically significant differences. In addition, for effects of 0.20SD, power is only 26%, which means there is a high likelihood of making a Type II error. Thus, when interpreting the results, we will refer to both statistically significant results (concerning the *p* value) and clinical significant results (concerning the effect size).

3.8 Statistical 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 outcomes 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).[6] The 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:

$$(3) \qquad MP(t) = P(t^* > t) + 0.5P(t^* = t)$$

where *P(.)* is the probability distribution, *t** is the randomly permuted *t*-statistic, and *t* is the observed *t*-statistic. Similar to other early childhood intervention 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% are used given the hypothesis that the high treatment will have a positive effect on children's outcomes.

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

In order to account for any potential bias due to differential attrition, 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 the Age 14 assessment. The measures which are the significant predictors of attrition in Figure 2 are included in the logistic models modelling the likelihood of participating in the Age 14

[6] As permutation testing does not depend on the asymptotic behaviour 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, therefore if the null hypothesis is true, taking random permutations of the treatment variable does not change the underlying distribution of 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. Mewhort 2005). Although this method is useful for dealing with non-normal data, it cannot be used to compensate for an under-powered study.

assessment. Separate models are estimated for the high and low treatment groups. 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.

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 all outcome measures are placed into a series of stepdown families each representing an underlying construct. 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 found.

The results are discussed using *p*-values to indicate statistically significant effects, where $p<0.1$ is considered statistically significant, and Cohen's d effect sizes, where a small effect is 0.2, a medium effect is 0.5, and a large effect is 0.8.

4 Results

4.1 Cognitive outcomes

Cognitive skills are measured using the *British Ability Scales III* (Elliot & Smith, 2011) which consists of six subscales: word definitions, verbal similarities, matrices, quantitative reasoning, recognition of designs, and pattern construction. These sub-scales yield an overall score reflecting general cognitive ability (General Conceptual Ability, GCA), as well as three cluster scores for Verbal Ability, Non-Verbal 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, spatial visualisation, and short-term visual memory. The *Nonverbal Reasoning* score assesses inductive reasoning. The *Verbal Ability* score assesses children's verbal reasoning, verbal knowledge, and expressive

language. Age-based *T* scores are calculated for each domain that are standardized to have a mean of 100 and a standard deviation of 15, as well as cutoff scores indicating whether the child scores above or below average for the GCA and cluster scores.

Executive function is measured using the *National Institutes of Health Toolbox for Assessment of Neurological and Behavioral Function Cognition Battery* (NIH Toolbox; Zelazo & Bauer, 2013). Executive functions are higher order meta-cognitive processes involved in concentration, reasoning, problem solving, and planning[7]. *The Flanker Task* was used to assess inhibitory control. Children were asked to indicate the left-right orientation of a centrally presented stimulus arrow surrounded by congruent or incongruent stimuli arrows. The *Dimensional Change Card Sort* task was used to assess attention flexibility. Children were asked to match test pictures to a target picture that varied along two dimensions, colour and shape. Finally, the *List Sorting* task was used to assess working memory. Children were presented with a series of stimuli (either food or animals) on screen and orally and asked to order the list of items from smallest to largest; and then presented with a series of stimuli and asked to recall the food items in size order followed by the animals in size order from smallest to largest. Age-corrected scores for each of the three NIH toolbox measures was then standardised and summed to create a composite indicator of executive functions.

Table 4 reports the Inverse Probability Weighted (IPW) adjusted means, standard deviations, and *p*-values that result from weighted individual and stepdown permutation tests, controlling for gender, alongside the effect size (as measured by the ratio of the treatment effect and the pooled standard deviation), for children's cognitive outcomes.[8] The results indicate that the *PFL* programme had a significant impact on children's skills. The treatment increased children's overall GCA score by 0.70 SD. The effect size is equivalent to the ones observed at both Age 4 and 9 assessments. The results also demonstrate that *PFL* had a significant impact on each dimension of cognitive ability including spatial ability (0.51 SD);

[7] Executive functions are comprised of three core abilities. 1. Inhibitory control which involves the ability to override impulse responses and ideally replace them with a more adaptive behaviour. For example, delaying eating a treat to receive a reward. 2. Attention flexibility which involves being able to deliberately focus and maintain attention or to divert attention to a new task if required. For example, blocking out distraction to complete a task or shifting attention from one task to another. And 3, working memory which involves the ability to retain and manipulate information over brief periods of time. Working memory is central to remembering instructions or rules or pieces of information that are necessary to solve a problem.

[8] The non-IPW adjusted results for the BAS outcomes are largely similar to the results reported here. The number of significant outcomes are the same. In terms of effect sizes, some are slightly larger in the non-IPW results and some are smaller. slightly more conservative with somewhat lower effect sizes. For the executive functioning outcomes, the number of significant results are the same in the IPW and non-IPW results, however the effect size is smaller and the level of significance higher in the non-IPW results.

non-verbal reasoning ability (0.53 SD); and also verbal ability (0.61 SD). In addition, all four composite scores survive adjustment for multiple hypotheses testing.

Children are classified as scoring above the norm if their score are above 110 points and below the norm if their scores are less than 90 points. Table 4 shows that high treatment children are more likely to score *above* the norm in terms of their overall cognitive ability and their spatial ability; with effect sizes of 0.55 to 0.46 SDs respectively. For example, 17% of children in the high treatment group score above the norm on their spatial skills compared to 4% in the low treatment group. Non-verbal and verbal ability are not statistically significant, although the effect sizes are of moderate size (~0.40 SD). In addition, the high treatment children are less likely to score *below* the norm across all three cognitive domains, as well as overall ability, results which are robust to multiple hypothesis adjustment. The effect sizes range from 0.35 to 0.86 of a standard deviation.

It is important to note that relatively few children, in either the high or low treatment group, score above the norm, while large proportions of children score below the norm. For example, only 7% of the high treatment group has above the norm GCA scores, while 53% have below the norm scores. The BAS III norms are based on a representative UK sample which includes children across all social groups. The scores identified here thus reflect the disadvantaged nature of the *PFL* cohort, where lower levels of cognitive ability are expected to be observed. Yet the counterfactual (low treatment group) reveals that without *PFL* intervention, a significantly greater proportion of the high treatment children would have scored below the norm, thus demonstrating the effectiveness of the program. Figure 3 shows that the distribution of GCA scores for the high treatment group is shifted to the right of the low treatment groups, indicating that the programme impacted children of all ability types – the program impacted both the average score, and the tails of the distribution.

The effects on the BAS scores at Age 14 are similar in magnitude to the BAS results measured at the end of the program (at approx. 51 months) and at Age 9, demonstrating the sustained impact of the program almost ten years after the treatment ended.

Figure 3 *Distribution of BAS GCA cognitive scores at age 14*

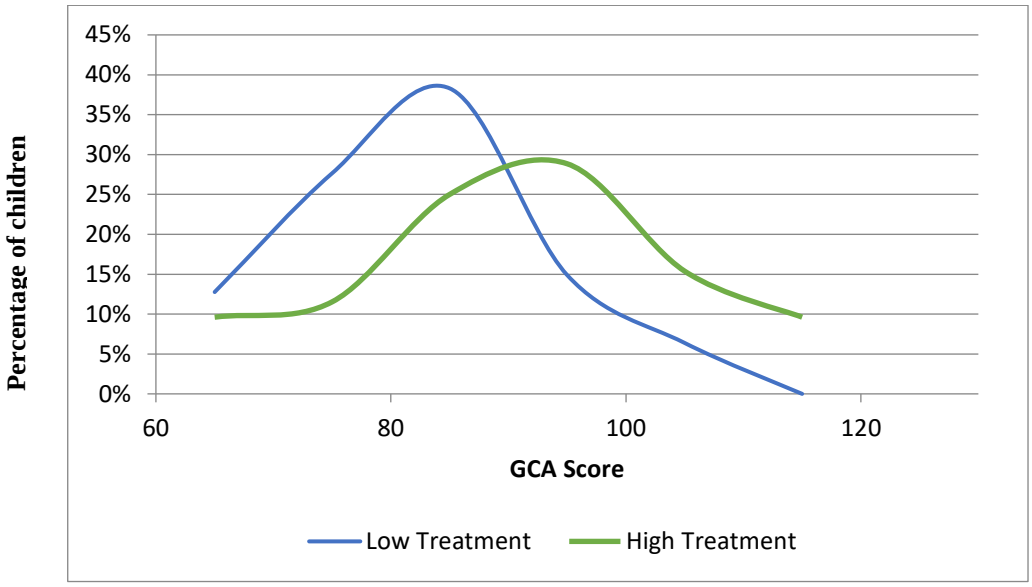


Table 4 also reports the results for executive functioning. At Age 9, the program impacted all three forms of executive functioning, however at Age 14, only impacts on working memory (the ability to retain and manipulate or use information over brief periods of time), are found. Children in the high treatment group score have significantly higher working memory skills than those in the low treatment group, with a large effect size of 0.52 SDs. This result also survived multiple hypothesis adjustment. There are no effects on inhibitory control, the ability to override impulse responses and attention flexibility, the ability to deliberately focus and maintain attention. Indeed, the low treatment group appear slightly better on these domains, however the effects are not significant and the effect sizes are small.

These results are somewhat in contrast to the results for executive functioning reported at Age 9, but they are more in line with the effects reported at the end of the program where a significant treatment effect was found for children's ability to control their attention, but not their ability to delay gratification. As the same tests were administered At age 9 and 14, these differences cannot be attributed to the use of different tests. It will be informative to measure the cohort's executive functioning skills later in adolescence to determine whether this Age 14 result is an artefact of the timing of data collection or a true fade out of effects on inhibitory control and attention flexibility.

Table 4 *Comparison of high and low treatment groups: Cognitive outcomes*

	N (HIGH/LOW)	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	p^1	p^2	*ES*
BAS Composite Scores						
General Conceptual Ability	99 (52/47)	85.41 (14.04)	76.76 (10.74)	**0.003**	**0.010**	0.70
Spatial Ability	99 (52/47)	94.09 (17.77)	86.28 (13.02)	**0.012**	**0.012**	0.51
Non-Verbal Ability	99 (52/47)	82.94 (12.36)	76.57 (11.62)	**0.015**	**0.023**	0.53
Verbal Ability	99 (52/47)	87.39 (12.82)	80.42 (9.89)	**0.016**	**0.023**	0.61
BAS Above the Norm %						
General Conceptual Ability	99 (52/47)	0.07 (0.26)	0.00 (0.00)	**0.090**	**0.091**	0.55
Spatial Ability	99 (52/47)	0.17 (0.38)	0.04 (0.19)	**0.042**	**0.044**	0.46
Non-Verbal Ability	99 (52/47)	0.06 (0.23)	0.00 (0.00)	0.153	0.153	0.48
Verbal Ability	99 (52/47)	0.04 (0.21)	0.00 (0.00)	0.158	0.158	0.42
BAS Below the Norm %						
General Conceptual Ability	99 (52/47)	0.53 (0.50)	0.88 (0.33)	**0.000**	**0.002**	0.86
Spatial Ability	99 (52/47)	0.38 (0.49)	0.55 (0.50)	**0.063**	**0.063**	0.35
Non-Verbal Ability	99 (52/47)	0.74 (0.44)	0.92 (0.28)	**0.045**	**0.062**	0.48
Verbal Ability	99 (52/47)	0.60 (0.49)	0.85 (0.37)	**0.020**	**0.039**	0.57
NIH Toolbox Executive Functioning						
Flanker Task - Inhibitory Control	98 (51/47)	95.10 (17.78)	98.27 (17.57)	0.790	0.790	-0.18
Dimensional Change Card Sort Task - Attention Flexibility	99 (52/47)	111.49 (20.52)	113.94 (19.84)	0.624	0.794	-0.12
List Sorting Task - Working Memory	99 (52/47)	102.44 (16.20)	94.68 (13.89)	**0.023**	**0.052**	0.52
Other						
Composite Executive Function Score	98 (51/47)	0.07 (0.75)	0.00 (0.78)	0.342	~	0.09

Notes: N' indicates the sample size. 'M' indicates the IPW-adjusted mean. '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. 'Effect size' is the ratio of the treatment effect to the pooled standard deviation.

4.2 Socio-emotional outcomes

Socio-emotional skills were measured using a range of different instruments including the Brief Problems Monitor (Achenbach, McConaughy, Ivanovaa, and Rescorla, 2011) and the Strengths and Difficulties Questionnaire (Goodman, 1997), both of which were used at the Age 9 assessment. The *Brief Problems Monitor* (BPM) yields scores across three subscales: *internalizing* (α = 0.83), *externalizing* (α = 0.77), and *attention* (α = 0.81) problems. The scores for each of the three problems subscales were summed to create a *Total Problems* (α = 0.87) score. Scores were then converted to standard scores based on the child's age and gender, and binary indicators of concerning problem behaviour were created. Higher scores are indicative of more behavioural problems.

The *Strengths and Difficulties Questionnaire* (SDQ; Goodman, 1997) is a 25-item questionnaire assessing behaviours, emotions, and relationships. The instrument yields scores across five subdomains: *conduct problems* (α = 0.71), *emotional symptoms* (α =0.73), *hyperactivity* (α = 0.79), *peer problems* (α = 0.62), and *pro-social behaviour* (α =0.63). The five items for each subscale were summed giving a total score of 0 to 10 for each subscale (α = 0.81). Cutoff scores were also created to indicate scores that were of clinical concern. In all cases, apart from the prosocial behaviour continuous outcome, higher scores are indicative of more problems.

Three new measures were also included in the Age 14 Follow-up. First, the *Short Mood and Feelings Questionnaire* (Angold and Costello, 1987) which was also used in the *Growing Up in Ireland Age 13 study* and the *Millennium Cohort Study Age 14* study, measures cognitive, affective, and behavioral-related symptoms of depression during the last two weeks. It includes 13 items which were used to create a summative score (α = 0.93), with higher values indicating more negative feelings, as well as a cutoff score (>=12) indicating that the participant is at risk of depression. Second, the *Rosenberg Self-Esteen Scale* (Rosenberg, 1965), includes 10 items which were summed to create a continuous score (α = 0.89), whereby higher values are indicative of higher levels of self-esteem. This measure was also administered to the *PFL* parents at baseline. Third, a single item life satisfaction question, which asks participants "*Here is a picture of a ladder. The top of the ladder "10" is the best possible life for you and the bottom "0" is the worst possible life for you. In general, where on the ladder do you feel you stand at the moment*?". Higher scores are indicative of

greater life satisfaction. The question was also used in the *Health Behaviour of School Age Children* (HBSC) survey.

Table 5 reports the IPW-adjusted means, standard deviations, and *p*-values that result from weighted individual and stepdown permutation tests, controlling for gender, alongside the effect size, for socio-emotional outcomes.[9] Consistent with Age 9, the results indicate that the *PFL* programme did not have a significant impact on children's socio-emotional development at Age 14. In sum, only one of the measures is statistically significant and survives adjustment for multiple hypothesis testing. Children in the high treatment group have less attention problems as measured by the *Brief Problems Monitor* scale compared to children in the low treatment group. The results are significant for both the continuous and cutoff scores, with moderate-large effect sizes of 0.61 SDs and 0.46 SDs respectively. For example, 63% of low treatment children are classified as having significant attention problems, comparted to 41% of high treatment children. For the other outcomes, the results are mainly in the right direction, e.g., high treatment children have better socio-emotional skills, the results are not statistically significant and the effect sizes are very small which is indicative of a true null effects rather than an underpowered sample.

[9] The non-IPW adjusted results are largely similar. There are few differences in the number (or level) of statistically significant results between the IPW and non-IPW adjusted models and the effect sizes are largely similar. There are two exceptions, the SDQ hyperactivity score reaches statistical significant at the 10% level and the BPM attention cutoff score does not reach significance in the non-IPW results.

Table 5 *Comparison of high and low treatment groups: Socio-emotional outcomes*

	N (HIGH/LOW)	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	p^1	p^2	*ES*
Brief Problem Monitor Scores						
BPM Internalising problems	98 (51/47)	58.80 (7.60)	58.56 (7.45)	0.655	0.655	-0.03
BPM Externalising problems	98 (51/47)	55.14 (6.12)	55.55 (6.56)	0.392	0.625	0.06
BPM Attention problems	98 (51/47)	60.75 (7.67)	65.31 (7.26)	**0.007**	**0.015**	0.61
Brief Problem Monitor Cutoff Scores %						
BPM Internalising problems	98 (51/47)	0.29 (0.46)	0.23 (0.43)	0.776	0.776	-0.13
BPM Externalising problems	98 (51/47)	0.09 (0.29)	0.12 (0.33)	0.280	0.528	0.11
BPM Attention problems	98 (51/47)	0.41 (0.50)	0.63 (0.49)	**0.033**	**0.067**	0.46
SDQ Scores						
SDQ Conduct Problems	98 (51/47)	2.04 (1.98)	2.56 (2.03)	0.108	0.421	0.26
SDQ Emotional Problems	98 (51/47)	4.16 (2.71)	4.35 (2.73)	0.525	0.702	0.07
SDQ Hyperactivity	98 (51/47)	5.57 (3.09)	6.20 (2.53)	0.196	0.472	0.22
SDQ Peer Problems	98 (51/47)	2.09 (2.03)	2.05 (1.83)	0.548	0.548	-0.02
SDQ Prosocial behaviour (+)	98 (51/47)	8.23 (1.57)	8.18 (1.71)	0.286	0.600	0.03
SDQ Cutoff Scores						
SDQ Conduct Problems %	98 (51/47)	0.10 (0.30)	0.08 (0.27)	0.675	0.675	-0.08
SDQ Emotional Problems %	98 (51/47)	0.31 (0.47)	0.34 (0.48)	0.517	0.828	0.08
SDQ Hyperactivity %	98 (51/47)	0.39 (0.49)	0.49 (0.51)	0.212	0.640	0.21
SDQ Peer Problems %	98 (51/47)	0.21 (0.41)	0.24 (0.43)	0.443	0.739	0.08
SDQ Prosocial behaviour %	98 (51/47)	0.06 (0.23)	0.07 (0.25)	0.343	0.633	0.04
Other Socio-emotional outcomes						
Short Mood and Feelings Questionnaire	98 (51/47)	6.95 (6.65)	7.84 (6.56)	0.413	0.508	0.13
Rosenberg Self-esteem Scale	98 (51/47)	18.76 (5.47)	17.60 (5.86)	0.277	0.455	0.21
Life Satisfaction (1-10)	98 (51/47)	7.48 (2.16)	7.30 (2.06)	0.435	0.435	0.08
Non Stepdown Outcomes						
BPM Total problems standardised score	98 (51/47)	59.62 (7.52)	61.56 (7.60)	0.147	~	0.26
BPM Total problems cutoff %	98 (51/47)	0.37 (0.49)	0.44 (0.50)	0.263	~	0.14
SDQ Total score	98 (51/47)	13.86 (7.04)	15.16 (7.25)	0.279	~	0.18
SDQ Total cutoff %	98 (51/47)	0.36 (0.49)	0.38 (0.49)	0.489	~	0.04
SMFQ Cutoff %	98 (51/47)	0.19 (0.40)	0.27 (0.45)	0.282	~	0.20

Notes: N' indicates the sample size. 'M' indicates the IPW-adjusted mean. '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. 'Effect size' is the ratio of the treatment effect to the pooled standard deviation.

Over one-third of children are classified as having high or very high socio-emotional problems in both the high and low treatment group. As shown in Table 6, this compares with only 6% in the nationally representative Growing Up in Ireland (GUI) cohort assessed at Age 13. In all cases the *PFL* cohort report much higher conduct problems, emotional problems, hyperactivity, peer problems and lower prosocial behaviour than the GUI cohort, A caveat to these results, is that the *PFL* measure is based on self-report while the GUI measure is based on parent report. However, there is one mental health measure that is common and self-reported across both studies – the *Short Mood and Feelings Questionnaire* (SMFQ). As shown in Table 6, the *PFL* cohort have considerably higher scores than the GUI cohort, indicating poorer mental health. In addition, over 40% of the *PFL* cohort reach the cutoff indicative of depression, compared to only 16% in the GUI cohort. Thus the *PFL* cohort have significantly poor socio-emotional skills compared to the average 13 year old in Ireland.

It is also possible to compare the average life satisfaction of the *PFL* cohort to a national representative sample of 15 year olds who took part in Health Behaviour of School Age Children (HBSC) survey and used the same instrument. The average life satisfaction of girls and boys in the *PFL* cohort is 6.8 and 7.6 respectively, compared to 5.9 and 6.7 among the HBSC sample and 7.8 and 8.3 among the GUI sample. Thus, *PFL* report lower life satisfaction compared to the GUI cohort but higher life satisfaction compared to HBSC cohort. Across all samples, boys report higher life satisfaction than girls.

Table 6 *Comparison pf PFL Cohort at Age 14 & Growing up in Ireland Cohort at Age 13*

	PFL_{HIGH}	PFL_{LOW}	*GUI*
SDQ Scores			
SDQ Conduct Problems	2.04	2.56	0.93
	(1.98)	(2.03)	(1.30)
SDQ Emotional Problems	4.16	4.35	2.24
	(2.71)	(2.73)	(2.29)
SDQ Hyperactivity	5.57	6.20	2.61
	(3.09)	(2.53)	(2.42)
SDQ Peer Problems	2.09	2.05	1.25
	(2.03)	(1.83)	(1.58)
SDQ Prosocial behaviour (+)	8.23	8.18	8.74
	(1.57)	(1.71)	(1.59)
SDQ Total score	13.86	15.16	6.98
	(7.04)	(7.25)	(5.42)
SDQ Total cutoff %	36%	38%	6.4%
Other Socio-emotional outcomes			
Short Mood and Feelings Questionnaire	6.95	7.84	3.86
	(6.65)	(6.56)	
SMFQ Cutoff % > 8	40.38%	44.68%	15.9%
N	51	47	6650

4.3 Health outcomes

A number of different measures were used to assess health and health behaviours at Age 14. Many of the instruments were used in other cohort studies including the GUI study and the HBSC study in Ireland and the MCS in the UK.

- **Self-assessed health** – Assessed using a single item "*Would you say your health in general is...Excellent, Very good, Good, Fair, Poor".* A binary variable was created where 0 = fair/poor and 1 = excellent/very good/good.
- **Diet** – Assessed using 4 items asking "*How often do you eat breakfast/fruit/veg/fast food over a week"*. Four binary variables were created where 0 = sometimes and 1 = never.
- **Puberty** – Assessed using the *Pubertal Development Scale* (Petersen, Crockett, Richards, and Boxer, 1988). The instrument is based on 3 common items for girls and boys (growth spurts, skin changes, and body hair), and 2 additional questions for boys (voice changes, facial hair) and girls (breast development, menstruation). Response options on all items (apart from menstruation) are not yet started (1 point); barely started (2 points); definitely started (3 points); seems complete (4 points); I don't know (missing). On the menstruation item yes = 4 points; no = 1 point. The point values were averaged to create a Pubertal Development Scale (PDS) score, whereby higher values indicates the participant is more developed.
- **Waist-to-Height (WTH) ratio** – Assessed by measuring the participant's height and waist circumference (by the fieldworker). The waist-to-height (WTH) ratio was calculated by waist size (cm) divided by height (cm). Higher values are associated with more health problems. A binary risk score was also created where 1 = WTH score > 0.5[10] (moderate/high risk) and 0 if < 0.5 (low risk) of obesity.
- **Substance use** – Assessed using 4 items asking whether the participant ever smoked cigarettes, vaped, drank alcohol, smoked cannabis, or took illegal drugs [0 = never; 1 = ever].

Table 7 report the IPW-adjusted means, standard deviations, and *p*-values that result from weighted individual and stepdown permutation tests, controlling for gender, alongside the effect size, for health outcomes. The results indicate that the *PFL* program had few effects on

[10] A number of number of studies state that a WHT ratio cutoff value of 0.5 is a suitable marker for screening of central obesity in children and adolescents of all genders (Browning et al., 2010).

participants' health at Age 14. Four of the individual measures are statistically significant in the individual tests, however none survive adjustment for multiple hypothesis correction. In addition, in the non-IPW weighted results, only two of the measures are significant. The two measures are related – waist circumference and the waist-to-height ratio (which is derived from the waist measure). Overall, the high treatment group has a lower WTH ratio compared to the low treatment group with a moderate effect size of 0.41 SDs. In addition, there is almost a 6cm difference in the waist circumference of the high and low treatment groups. Although the result is not significant in the more conservative test, it is suggestive that the program has had a long term impact in reducing the waist size of the high treatment group. Note, that this result is consistent with findings from the Age 4 assessment whereby the high treatment group were less likely to be obese/overweight, however at Age 9, no such effect was found.

For the Age 14 assessment a decision was made to measure waist circumference (to derive WTH ratio) instead of weight (to derive BMI) for two reasons. First, for young adolescents, measuring waist circumference is less sensitive than measuring body weight using a weighting scales. Second, evidence suggests that the WTH ratio is a better measure of central obesity which is a risk factor for cardiometabolic disease in both adults and children (Eslami *et al.*, 2023). Within the sample, 38% of the high treatment group and 44% of the low treatment group are classified as having a high WTH ratio, indicating that a large proportion of the cohort are at risk of obesity. The average WTH ratio is 0.47 and 0.50 for the high and low treatment group respectively. This compares with an average of 0.40 found in a representative sample of adolescences (aged 13-18) from the *Irish National Nutrition Survey* in 2020 (Moore Heslin *et al.*, 2023).

Table 7 *Comparison of high and low treatment groups: Health outcomes*

	N (HIGH/LOW)	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	p^1	p^2	*ES*
Health outcomes						
Self-rated health %	99 (52/47)	0.79 (0.41)	0.77 (0.42)	0.517	0.764	0.04
Never eats breakfast %	99 (52/47)	0.14 (0.35)	0.25 (0.44)	**0.098**	0.440	0.29
Never eats fruit %	99 (52/47)	0.19 (0.40)	0.25 (0.44)	0.303	0.740	0.13
Never eats vegetables %	99 (52/47)	0.19 (0.40)	0.23 (0.42)	0.250	0.758	0.09
Eats fast food up to 6 times per week %	99 (52/47)	0.19 (0.39)	0.19 (0.40)	0.486	0.486	0.01
Puberty Development Scale Score	76 (38/38)	3.10 (0.42)	3.07 (0.51)	**0.097**	0.758	0.06
Waist-to-height ratio	92 (47/45)	0.47 (0.06)	0.50 (0.08)	**0.053**	0.271	0.41
Substance use %						
Ever drank alcohol	93 (47/46)	0.23 (0.43)	0.25 (0.44)	0.547	0.671	0.04
Ever smoked cigarettes	95 (49/46)	0.05 (0.22)	0.06 (0.25)	0.398	0.713	0.06
Ever vaped	98 (51/47)	0.33 (0.47)	0.23 (0.43)	0.845	0.845	-0.21
Ever tried cannabis	97 (51/46)	0.05 (0.22)	0.10 (0.31)	0.140	0.432	0.22
Ever took illegal drugs	96 (50/46)	0.00 (0.00)	0.02 (0.13)	0.106	0.514	0.25
Non-stepdown measures						
Waist measurement (cms)	92 (47/45)	78.82 (11.32)	84.60 (14.17)	**0.024**	~	0.45
Height (cms)	97 (51/46)	166.63 (8.04)	167.96 (7.45)	0.938	~	-0.17
High WTH ratio %	92 (47/45)	0.38 (0.49)	0.44 (0.50)	0.405	~	0.12
Age at first drink	25 (13/12)	13.19 (1.03)	13.04 (1.19)	0.460	~	0.14

Notes: N' indicates the sample size. 'M' indicates the IPW-adjusted mean. '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. 'Effect size' is the ratio of the treatment effect to the pooled standard deviation.

Regarding the other health outcomes, in most cases, the high treatment group has better health than the low treatment group, but none of these differences are statistically significant in either the individual or stepdown tests. Among the cohort, 79% and 77% of the high and low treatment groups respectively report their health to be good or very good. This is comparable to the 72% of parents in the GUI sample who report their Age 13 children's health to be 'very healthy, with no problems' (GUI, 2023).

Some of the health measures included in the Age 4 assessment are also present in the Health Behaviour of School-Aged Children (HBSC) Age 15 survey. As shown in Table 8, *PFL* boys and girls are less likely to eat breakfast, fruit, and vegetables every day compared to the HBSC cohort, illustrating that the *PFL* cohort have a poorer diet than the average Irish

teenager. In particular, 31% of the *PFL* cohort eat breakfast every day, compared with 54% in the HBSC cohort and 74% in the GUI cohort. However, they exhibit somewhat better behaviour regarding substance use. The *PFL* cohort are less likely to have smoked cigarettes and drank alcohol compared to the HBSC cohort, however rates of vaping and cannabis use are similar in the *PFL* and HBSC cohorts. Approximately 20% of boys and 30% of girls have vaped in both cohorts, and 8% of both cohorts have used cannabis. In general, cigarette use is low, and about 40% of the samples have ever drank alcohol. Interestingly, girls engage in substance use more frequently than boys in both samples.

Table 8 *Comparison of PFL Cohort at Age 14 & HBSC Cohort at Age 15*

	PFL Boys	*HBSC Boys*	*PFL Girls*	*HBSC Girls*
Eats breakfast daily	43%	62%	24%	46%
Eats fruit daily	8%	39%	14%	42%
Eats vegetables daily	18%	42%	15%	44%
Ever smoked cigarettes	5%	12%	9%	15%
Ever vaped	23%	22%	32%	31%
Ever drank alcohol	14%	39%	39%	45%
Ever used cannabis	8%	8%	8%	8%

Note: HBCS data is taken from Health Behaviour in School-aged Children study (2023), Data browser (findings from the 2021/22 international HBSC survey): https://data-browser.hbsc.org.

Within, the *PFL* cohort, the average puberty development score is 2.67 for boys and 3.37 for girls (with a maximum score of 4). This indicates that boys in the *PFL* cohort are at an earlier stage of pubertal development than girls at Age 14, which is the norm. As the Puberty Development Scale was also included in the Millennium Cohort Study, it is possible to compare the development of the *PFL* cohort and the MCS cohort as the average age in both cohorts at the time of interview was 14.3 years. Table 9 shows the proportion of both groups who have started or completed various stages of puberty. In almost all cases, the *PFL* cohort are further along in their pubertal development than the MCS cohort. For example, 86% of the *PFL* cohort have started their growth spurt compared to 65% in the MCS cohort. In addition, the average age to begin menstruation was 11.4 years in the *PFL* cohort compared to 12.1 years in the MCS cohort. There is some evidence that lower levels of socio-economic status is associated with an earlier onset of puberty and earlier age of menarche in developed countries (e.g., Arim et al., 2007; Deardorff et al., 2014; James-Todd et al., 2010; Sun et al., 2017). Thus, the results reported here align with this literature as the MCS is a nationally representative cohort, while the *PFL* cohort reside in disadvantaged communities.

The mechanisms through which low SES may predict earlier pubertal onset are still unknown. However, one possible explanation is that early exposure to stress (initiated by low SES) may impact the epigenome and the regulation of hormones.

Table 9 *Comparison of PFL Cohort and MCS Cohort at Age 14*

% started/completed	**PFL**	**MCS**
Growth spurt	86%	65%
Body hair	84%	85%
Skin changes	77%	66%
Voice change (males)	71%	63%
Facial hair (males)	29%	37%
Breast growth (females)	92%	79%
Menstruation began (females) (% yes)	96%	91%
Age menstruation began (females)	11.4 yrs	12.1 yrs
N	99	11,000
Average age at time of interview	14.3yrs	14.3yrs

4.4 Educational Engagement & Time-Use outcomes

Several measures were used to assess educational engagement and time use at Age 14. As above, many of these instruments were used in GUI and MSC studies.

- **School liking** – Assessed using a single item *"How do you feel about school in general?"* (on a scale of 1-5). A binary variable was created where 0 = Dislike school and 1= Like school.
- **School engagement** – Assessed using the 4-item *Classroom Climate Measure* (Rowe, Kim, Baker, Kamphaus, & Horne, 2010) which included items such as "*I look forward to going to school*" (on a scale of 1-5). A standardised summative score (α = 0.83) of the 4 items was created, whereby higher values are equal to more positive school engagement.
- **School belonging** – Assessed using the 9-item *School Belonging Scale* (used in the German Socio-Economic Panel) which includes items such as *"I feel like an outsider in school*" (on a scale of 1-4). A standardised factor score was created whereby higher values are equal to a greater sense of belonging (α = 0.80).
- **School absences** – Assessed using a single item *"In the last 12 months, how often did you miss school without your parents' permission? (most days…never)".* A binary variable was created where 0 = Ever missed school and 1= Never missed school.
- **Expectations** – Assessed using a single item "*How likely do you think it is that you will go to third level education e.g. university/college?* Participants were asked to

select a number from 0 to 100 using a slider. Higher scores indicate a higher expectations that the participant will attend university.

- **Time Use** – Assessed using three items asking *"On a normal weekday during term time, how many hours do you spend doing homework / using the internet / on social networking or messaging sites or Apps?".* There were 8 response options ranging from none to more than 7 hours. Continuous measures, whereby higher values mean more time spent on the activity, were used. In addition, binary variables were created whereby 1 = None and 0 = Any for homework, and 1 > 7hrs and 0 = < 7hrs for the internet and social media. Thus, in all cases, higher values correspond to more negative outcomes.

Table 10 report the IPW-adjusted means, standard deviations, and *p*-values that result from weighted individual and stepdown permutation tests, controlling for gender, alongside the effect size, for education engagement and time use outcomes.[11] It shows that the program had little impact on educational engagement and no impact on time use. For most outcomes, the high treatment group have better scores than the low treatment group, but only one outcome is statistically significant in the individual test, and almost reaches significance in the stepdown test (and reaches significance in the non-IPW results). In particular, 66% of the high treatment group state that they intend to go to university in the future compared to 51% in the low treatment group, with an effect size of 0.51 SDs.

There are some items which are common or similar in the *PFL* Age 14 assessment and the GUI Age 13 assessment and the MCS Age 14 assessment. For example, 66% and 51% of the high and low treatment group respectively intend to go to university. This compared to 76% in the GUI cohort and 70% in the MCS cohort. If we compare the MCS and *PFL* cohorts, there is no statistically significant difference in the expectation of the high treatment group and the MCS cohort to go to university (p=0.331), however the low

[11] Similar results are found in the non-IPW results.

treatment group is significantly less state that they expect to go to university (p=0.000). This suggests that the program raised the aspirations of the high treatment group to the national average in the UK. In addition, 55% of the *PFL* cohort report doing less than 30 minutes of homework on a normal weekday. This compares to 12% in the GUI cohort and 7.9% in the MCS cohort. Also, 16% and 15% of the high and low treatment group respectively spend more than 7 hours a day on social networking sites, compared to 9.56% in MCS. Finally, 47% of the *PFL* cohort have missed school without permission in the last 12 months, compared to only 2.5% in the GUI cohort and 9.26% in the MCS cohort. Thus, the *PFL* cohort have lower levels of engagement with education compared to the national cohorts.

Table 10 *Comparison of high and low treatment groups: Educational Engagement Outcomes & Time Use*

	N (HIGH/LOW)	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	p^1	p^2	*ES*
Educational outcomes						
Likes school %	99 (52/47)	0.76 (0.43)	0.68 (0.47)	0.301	0.610	0.18
School engagement score (std)	99 (52/47)	0.14 (0.92)	0.13 (1.14)	0.188	0.468	0.27
School belonging score (std)	99 (52/47)	0.14 (0.96)	0.00 (0.93)	0.280	0.280	0.15
School absence (% never)	99 (52/47)	0.57 (0.50)	0.48 (0.50)	0.304	0.474	0.17
Intention of going to university/college %	99 (52/47)	66.34 (28.55)	50.73 (32.58)	**0.012**	0.100	0.51
Time use continuous measures						
Time spent on homework per day (1-8)	99 (52/47)	2.43 (1.37)	2.17 (1.24)	0.173	0.392	0.20
Time spent on social media per day (1-8)	99 (52/47)	5.15 (2.21)	5.65 (1.72)	0.239	0.393	0.26
Time spent on internet (1-8)	99 (52/47)	6.27 (1.97)	6.55 (1.55)	0.321	0.321	0.16
Time use binary outcomes						
No homework %	99 (52/47)	0.33 (0.47)	0.42 (0.50)	0.192	0.538	0.19
>7hrs social media %	99 (52/47)	0.16 (0.37)	0.15 (0.37)	0.604	0.604	-0.01
>7hrs internet %	99 (52/47)	0.34 (0.48)	0.37 (0.49)	0.487	0.649	0.05

Notes: N' indicates the sample size. 'M' indicates the IPW-adjusted mean. '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. 'Effect size' is the ratio of the treatment effect to the pooled standard deviation.

4.5 Antisocial Beliefs and Attitudes outcomes

Antisocial beliefs are measured using the *Antisocial Beliefs and Attitudes Scale* (Butler et al., 2015) which has 28 items measured on a 4-point scale from 'Strongly disagree' to 'Strongly agree'. The instruments contains 3 sub-domains 1) Rule non-compliance (α = 0.73) e.g., "*I don't like having to obey all the rules at home and in school"*, 2) Peer conflict (α = 0.72) e.g., "*It's ok to walk away from a fight",* and 3) Aggression (α = 0.72) e.g., "*It's ok to hit my mother as long as I don't hurt her"*, and a total score (α = 82). Summative scores are created whereby higher scores indicate greater acceptance of antisocial behaviors.

Table 11 report the IPW-adjusted means, standard deviations, and *p*-values that result from weighted individual and stepdown permutation tests, controlling for gender, alongside the effect size, for the participants' antisocial beliefs and attitudes.[12] It shows that there are no treatment effects on antisocial beliefs and attitudes at Age 14. In some cases, the high treatment group report higher (i.e., worse) scores, but these are not statistically significant and the effect sizes are small. Thus, overall the program has no impact on antisocial beliefs.

Table 11 *Comparison of high and low treatment groups: Antisocial Beliefs*

Antisocial Beliefs and Attitudes Scale	N (HIGH/LOW)	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	p^1	p^2	*ES*
Rule non-compliance	99 (52/47)	12.25 (3.66)	12.98 (4.31)	0.260	0.483	0.18
Peer conflict	99 (52/47)	9.87 (3.72)	9.21 (4.52)	0.702	0.702	-0.16
Aggression	99 (52/47)	4.70 (3.06)	4.44 (3.53)	0.661	0.838	-0.08
Non stepdown measure						
Total score	99 (52/47)	26.82 (7.62)	26.63 (9.53)	0.549	0.929	-0.02

Notes: N' indicates the sample size. 'M' indicates the IPW-adjusted mean. '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. 'Effect size' is the ratio of the treatment effect to the pooled standard deviation.

4.6 Parent-Child Relationship outcomes

Two instruments were used to assess the child's perception of their relationship with their parents. First, the *Inventory of Parents and Peer Attachment-Revised* (IPPA; Gullone & Robinson, 2005). This instrument includes 50 items measured on a 5-point scale from 'Always true' to 'Never true', 25 pertaining to the participant's relationship with their mother

[12] In the non-IPW results, there is a significant treatment effect on the Rule Non-compliance sub-domain, with the high treatment group reporting better compliance than the low treatment group.

(or the person acting as their mother), and 25 pertaining to their father (or the person acting as their father). The instrument contains 3 sub-domains 1) Trust ($\alpha_M = 0.89$; $\alpha_F = 0.93$) e.g., "*My mother/father respects my feelings*", 2) Communication ($\alpha_M = 0.89$; $\alpha_F = 0.91$) e.g., "*My mother/father can tell when I'm upset about something*", and 3) Alienation ($\alpha_M = 0.81$; $\alpha_F = 0.86$) e.g., "*I don't get much attention from my mother/father*", and a Total Score for each parent. For the Trust and Communication sub-domains, higher scores represent more positive outcomes, while higher scores on the Alienation sub-domain represents more negative outcomes. The total score is the sum of the Trust and Communication sub-domains, minus the Alienation sub-domain.

Second, parent-child relationship is assessed using 10 items (5 for mothers, and 5 for fathers) used in the German Socio-economic Panel (SEOP) survey. Each item asks the participant how often 1) *they turn to their mother/father when worried about something*, 2) *their mother/father encourages or helps them when something is important*, 3) *their mother/father orders them around*, 4) *their mother/father tells them it's important to do well in school and to study a lot*, and 5) *they argue with their mother and father*. Each item is assessed on a 5-point scale from "Very often" to "Never". Binary variables are created whereby 1 = Often/Very often and 0 = Less often, seldom, never.

Table 12 report the IPW-adjusted means, standard deviations, and *p*-values that result from weighted individual and stepdown permutation tests, controlling for gender, alongside the effect size, for parent-child relationship outcomes.[13] It shows that the program had an impact on some dimensions of the parent-child relationship. For the IPPA instrument, the high treatment group reports significantly better communication with their mothers and trust with their fathers, with effect sizes of 0.36 and 0.45 SDs respectively. Neither result, however, survive adjustment for multiple testing. The total IPPA score for mothers is also statistically significant with an effect size of 0.37 SDs. For the SOEP items, the high treatment group report a significantly higher likelihood of turning to their mother when they are worried, with an effect size of 0.70 SDs. They also report that their mothers and fathers are more likely to encourage or help them with something important, with effect sizes of 0.48, and 0.48 SDs respectively. Importantly, the result for '*turning to mother when worried'* survived multiple hypothesis adjustment. In particular, 75% of the high treatment group state that they often or very often turn to their mothers when worried compared to 42% in the low

[13] In the non-IPW results, the results are largely similar to the IPW-adjusted results, however the IPPA Communication with mothers results is not statistically significant in the non-IPW results.

treatment group. Interestingly, there is one treatment effect in a non-hypothesized direction. The low treatment group is more likely to report that their mother tells them it is important to study (79% in the low treatment group vs 63% in the high treatment group). It is possible that the parents of high treatment children do not need to remind their children to study as the high treatment children are likely to study without the need for reminders. This is consistent with the finding that they spend more time on homework as shown in Table 9 (although this result is not statistically significant).

Table 12 *Comparison of high and low treatment groups: Parenting Relationship outcomes*

	N (HIGH/LOW)	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	p^1	p^2	*ES*
IPPA						
Communication Mother	98 (51/47)	35.93 (7.64)	33.12 (7.91)	**0.058**	0.244	0.36
Trust Mother	98 (51/47)	37.94 (5.58)	38.40 (7.15)	0.690	0.690	-0.07
Alienation Mother	98 (51/47)	12.59 (5.05)	13.77 (4.86)	0.211	0.410	0.24
Communication Father	98 (51/47)	31.17 (9.11)	28.49 (10.78)	0.142	0.392	0.27
Trust Father	98 (51/47)	37.10 (7.64)	32.85 (11.09)	**0.029**	0.150	0.45
Alienation Father	98 (51/47)	12.51 (5.53)	13.76 (6.26)	0.313	0.429	0.21
SOEP						
Turn to when worried (Mother) %	98 (51/47)	0.75 (0.44)	0.42 (0.50)	**0.001**	**0.027**	0.70
Encourages/helps you with something important (Mother) %	98 (51/47)	0.90 (0.30)	0.72 (0.45)	**0.030**	0.242	0.48
Tells you it's important to study Mother %	97 (50/47)	0.63 (0.49)	0.79 (0.41)	0.972^	0.972	-0.36
Argues with mother %	96 (50/46)	0.22 (0.42)	0.15 (0.36)	0.199	0.729	0.19
Orders you around (Mother) %	96 (51/45)	0.29 (0.46)	0.37 (0.49)	0.810	0.979	-0.17
Turn to when worried (Father) %	88 (49/39)	0.40 (0.49)	0.39 (0.50)	0.518	0.942	0.00
Encourages/helps you with something important (Father) %	**87 (48/39)**	0.81 (0.40)	0.60 (0.50)	**0.030**	0.268	0.48
Tells you it's important to study (Father) %	85 (47/38)	0.56 (0.50)	0.69 (0.47)	0.885	0.984	-0.27
Argues with father %	85 (48/37)	0.25 (0.44)	0.13 (0.34)	0.138	0.583	0.29
Orders you around (Father) %	84 (48/36)	0.17 (0.38)	0.30 (0.47)	0.817	0.979	-0.31
Non stepdown measures						
IPPA Total Score Mothers	98 (51/47)	55.76 (20.06)	47.58 (23.71)	**0.077**	0.527	0.37
IPPA Total Score Fathers	98 (51/47)	61.28 (16.63)	57.75 (18.21)	0.228	0.894	0.20

Notes: N' indicates the sample size. 'M' indicates the IPW-adjusted mean. '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. 'Effect size' is the ratio of the treatment effect to the pooled standard deviation. ^ result in non-hypothesised direction.

4.7 Time and Risk Preference outcomes

Time and risk preferences were assessed in two ways, First, using two self-assessed single item questions which were also asked in the MCS assessment. Second, using task-based elicitation across a series of games, which are the standard means of measuring preferences in an experimental setting.

For the self-assessed items, time preferences were measured by asking "*On a scale of 0-10, where 0 is never and 10 is always, how patient would you say you are?*". Higher levels of patience is associated with lower time preferences i.e., the participant places a higher value on the future and a lower value on the present. While low levels of patience is associated with higher time preferences i.e., the participant places a higher value on the present and a lower value on the future. In gender, low time preferences are associated with more positive/healthy behaviors and high time preferences are associated with more negative/unhealthy behaviors (e.g., smoking, drinking, not studying etc). Higher scores on the measure presented in Table 13 is indicative of lower time preferences (i.e., more patience).

Risk preferences were measured by asking "*On a scale of 0-10, where 0 is never and 10 is always, how willing to take risks would you say you are?".* Depending on the domain, a person who is willing to take more risks may have more negative outcomes (e.g., driving fast) or more positive outcomes (e.g., financial investment). Higher scores on the measure presented in Table 13 is indicative of being less risky.

Table 13 report the IPW-adjusted means, standard deviations, and *p*-values that result from weighted individual and stepdown permutation tests, controlling for gender, alongside the effect size, for time and risk preferences.[14] It shows that the high treatment group have lower time preferences (i.e., they place a higher value on the future than the present) compared to the low treatment group. While the result is not statistically significant, the effect size of 0.31 SD is sizeable, which suggests that the study could be underpowered to detect a significant effect on this item. In terms of risk preferences, the low treatment group are less risky than the high treatment group, but again, the result is not statistically significant and the effect size is very small (-0.14 SD), suggesting that the program had no impact on risk preferences.

[14] The non-IPW results are largely similar to the IPW-adjusted results,

It is possible to compare the average time and risk preferences of the *PFL* cohort to the MCS cohort. Within the MCS cohort, the average time preference score was 5.69 (on a 0-10 scale), compared to an average of 5.40 in the high treatment group and 4.66 in the low treatment group. This result is as expected as the MCS is a nationally representative sample of children, thus we would expect them to exhibit higher levels of patience than children living in a disadvantaged community. It is interesting that the high treatment group is closer to the average MCS score than the low treatment group. Indeed, there is no statistically significant difference in scores between the MCS cohort and the high treatment group (p = 0.374), however there is a significant difference in the scores of the MCS cohort and the low treatment group (p = 0.004). This suggests that the program played a role bringing the time preferences of the high treatment group into line with the national (UK) average.

A similar analysis can be conducted for the risk preference measure. Within the MCS cohort, the average risk preference score was 4.90 (on a 0-10 scale) whereby higher values correspond to being less risky. This compares to an average of 5.14 in the high treatment group and 5.38 in the low treatment group. Thus, in both cases, the *PFL* cohort are less likely to take risks than the MCS cohort, but the differences are not statistically significant (MCS v High treatment: p = 0.446; MCS v Low treatment: p = 0.333).

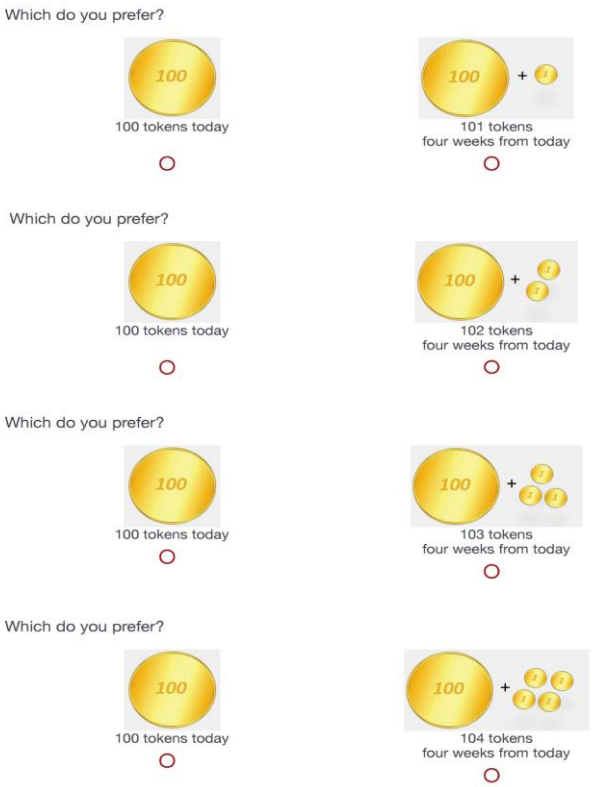
For the task-based elicitation methods, participants were asked to complete 3 games assessing time preferences (Coller and Williams, 1999) and 2 games assessing risk preferences (Holt & Laury, 2002) on the IPad.

To measure time preferences, participants were asked to make a choice between receiving 100 tokens today or X tokens 4 weeks from today (Game 1 – 9 choices), or 100 tokens today or X tokens 8 weeks from today (Game 2 – 10 choices), or 100 tokens 4 weeks from today or X tokens 8 weeks from today (Game 3 – 9 choices). In each game, the number of tokens increased sequentially from 101 to 125 in Game 1, from 101 to 150 in Game 2, and from 101 to 125 in Game 3. An example is provided below. Each game is analysed separately. Note, these choices were hypothetical i.e., participants did not receive any monetary payment based on their choice.

Responses to these games were used to create two measures used in the outcome analysis. The first assesses the proportion of times the participant choose the 'later' option over the 'sooner' option e.g., if they choose the ‘later’ option 3 times and the ‘sooner’ option 6 times, the receive a score of 67% (6/9). If they always choose the ‘sooner’ option, they

receive a score of 0%, and if they always choose the ‘later’ option, they receive a score of 100%. Thus, higher values correspond to lower time preferences e.g., the participant is more patient. The second measure assesses the first point at which the participant switches from the ‘later’ option to the ‘sooner’ option, e.g., if they choose the ‘sooner’ option for their first seven choices, and switch to the ‘later’ option for their eight choice, they receive a score of 8. If they always choose the ‘sooner’ option they receive a score of 10 and if they always choose the ‘later’ option, they receive a score of 1. If their first choice is the ‘later’ option, they receive a score of 1. For this measure, higher scores correspond to higher time preferences e.g., the participant is less patient.

Table 13 shows that there are no statistically significant differences across the high and low treatment groups on any of the time preference measures assessed. In addition, the effect sizes are very small, suggesting that the program had no impact on time preferences. For example, the high treatment group chose the 'later' option over the 'sooner' option 31% of the time, compared to 32% of the time for the low treatment group.



Example of task-based elicitation of time preferences

To measure risk preferences, participants were asked to make a choice between two options: receiving X tokens 'for certain' or having their reward determined by a coin flip where there is a 50% chance of winning Y tokens and a 50% change of winning Z tokens. For example, would you prefer option A "*to receive 70 tokens for sure*" or option B "*to have a 50% chance of winning 10 tokens or a 50% chance of winning 100 tokens*". Participants were informed that each token represents 10 cents, and that they should treat each decision as if it were real money, however, they did not receive a real monetary payout. In the first game, participants made 11 choices whereby the 'risky' choice did not change (50% chance of winning 10 tokens and 50% chance of winning 100 tokens), but the value of the 'certain' choice descended by 5 tokens (from 70 to 20 tokens) over the course of the 11 questions. For the second game, participants made 11 choices whereby the 'risky' choice did not change (50% chance of winning 20 tokens and 50% chance of winning 200 tokens), but the value of the 'certain' choice ascended by 5 tokens (from 40 to 140 tokens) over the course of the 11 questions. An example is provided below.

Responses to these games were used to elicit the two measures used in the outcome analysis. The first assesses the proportion of times the participant chose the 'safe' option over the 'risky' option e.g., if they chose the risky option 5 times and the safe option 6 times, their risk preference score is 55% (6/11). If they always chose the safe option, their score is 100%. If they always chose the risky option, their score is 0%. Thus, higher values correspond to less risky choices. The second measure assesses the first point at which the participant switches from choosing the safe option to the risky option e.g., if they chose the safe option for their first nine choices, and then choose the risky option for their tenth choice, they receive a score of 10. If they always choose the safe option, they receive a score of 12, and if they always choose the risky option, they receive a score of 1. If their first choice was the risky option, they receive a score of 1. Again, higher values correspond to less risky choices. Each game is analysed separately.

Table 13 shows that the high treatment have consistently lower scores than the low treatment group i.e., they are more risky loving and less risk averse. In these analyses, we are testing the hypothesis that the high treatment group is less risky than the low treatment group (in a one-tailed test), and we find no evidence that this hypothesis is supported. Indeed, the high *p*-values on Game 1 (for both measures), indicates that the high treatment group is significantly more risky than the low treatment group. For example, the high treatment group chooses the safe option over the risky option 35% of the time, while the low treatment group

chooses the safe option 45% of the time. Thus, the program may have led to less risk aversion. As discussed above, the willingness to take risks may have positive or negative outcomes depending on the domain and the level of risk considered. There is some research which shows that individuals with higher levels of cognitive skills are less risk averse than individuals with lower levels of cognitive skills. For example, Andreoni et al. (2000) finds that children and adolescents with higher cognitive skills (especially math skills) are more willing to take risks, possibly as it impacts their ability to process information on probability. As the high treatment group have significantly higher cognitive skills than the low treatment group, this may explain this finding on risk preferences. However, there is no correlation between the cognitive scores and risk preferences of the *PFL* cohort (correlation coefficient =0.04).

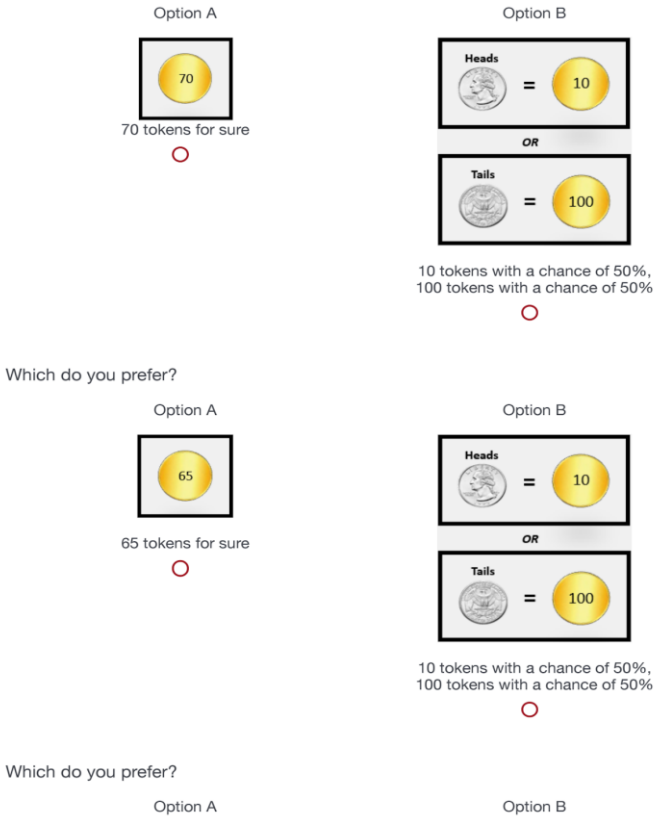


Table 13 *Comparison of high and low treatment groups: Time & Risk Preferences*

	N (HIGH/LOW)	M_{HIGH} (*SD*)	M_{LOW} (*SD*)	p^1	p^2	*ES*
Self-assessment						
Time preferences (higher = more patient)	96 (52/44)	5.40 (2.31)	4.66 (2.50)	0.149	0.223	0.31
Risk preferences (higher= less risky)	99 (52/47)	5.14 (1.56)	5.38 (1.87)	0.566	0.566	-0.14
Time preference games: % Later choice (more patient)						
Game 1	98 (51/47)	0.31 (0.32)	0.32 (0.27)	0.582	0.582	-0.01
Game 2	98 (51/47)	0.29 (0.26)	0.27 (0.25)	0.426	0.645	0.09
Game 3	98 (51/47)	0.33 (0.32)	0.31 (0.26)	0.498	0.631	0.07
Time preference games: First switch (less patient)						
Game 1	98 (51/47)	6.89 (3.19)	6.50 (2.98)	0.738	0.738	-0.13
Game 2	98 (51/47)	7.85 (2.86)	7.76 (3.16)	0.579	0.708	-0.03
Game 3	98 (51/47)	6.55 (3.38)	6.72 (2.87)	0.504	0.656	0.06
Risk preference games: % Safe choice (less risky)						
Game 1	98 (51/47)	0.35 (0.25)	0.45 (0.28)	0.932^	0.932^	-0.36
Game 2	98 (51/47)	0.53 (0.26)	0.60 (0.27)	0.794	0.900	-0.25
Risk preference games: First switch (less risky)						
Game1	98 (51/47)	3.48 (2.62)	4.47 (3.34)	0.926^	0.926^	-0.33
Game1	98 (51/47)	1.39 (1.41)	1.55 (1.93)	0.691	0.872	-0.09

Notes: N' indicates the sample size. 'M' indicates the IPW-adjusted mean. '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. 'Effect size' is the ratio of the treatment effect to the pooled standard deviation. ^ indicates the result is significant in the non-hypothesized direction.

5 Conclusions

The aim of the Age 14 Follow-up study was to examine whether the large and significant impacts of *PFL* found at the end of the program and at age nine were sustained. Prior evidence on the long-term impact of home visiting program into adolescence is inconclusive, as very few studies continue to track children beyond the lifetime of the intervention, and of the few NFP studies, that do, they fail to find effects. In contrast, this study finds that *PFL* continues to have a sizeable impact on children's cognitive skills approximately ten years after the children have finished the program. There is no evidence of

cognitive fade-out, with effect sizes of 0.70 a standard deviation on overall cognitive ability, and significant effects on some dimensions of executive functioning and health.

Overall, the IQ scores of the *PFL* children are above that of their parents (i.e., the Flynn effect), however the correlation between the high treatment children and their mothers is small and not statistically significant at either age five (r[15] = 0.07, p = 0.562) or age nine (r = 0.18, p = 0.148) and significant at age 14 (r = 0.34, p = *0.015*), compared to the large and significant correlation between the low treatment children and their mothers at age five (r = 0.31, p = 0.018), age nine (r = 0.57, p = 0.001) and age 14. (r = 0.54, p = 0.001).[16] Thus the program appears to be effective in reducing the intergenerational transmission of IQ scores, however, the correlation between parents and children's IQ is growing over time.

The program impacted all dimensions of cognitive skill including spatial ability, non-verbal ability, and verbal ability, in addition to reducing the proportion of children scoring below the standardized norm. The magnitude of the cognitive effects at age 14 (0.54 – 70 SD) are similar to those observed at the end of the program (0.56 - 0.77 SD) and at age nine (0.39 - 0.76 SD). An additional analysis shows that controlling for age five cognitive scores, slightly reduces the size of the age 14 treatment effects, however the impact of the program is still statistically significant.[17] This suggests that *PFL* is continuing to have an impact on children's development beyond the lifetime of the program.

This provides evidence in support of the skill formation model (Cunha and Heckman, 2007) which posits that developing children's skills early in life helps them to develop more advanced skills later in life (a process called self-productivity), and this raises the effectiveness of later investments, such as investments in schooling (a process called dynamic complementarity). If this process continues, and the high treatment group continue to utilize their higher cognitive skills, this is likely to translate into improved outcomes throughout the life cycle.

The size of the cognitive effects are substantially larger than those found in much of the existing literature. For example, the meta-analyses discussed earlier in the paper find effect sizes of less than 0.30 for cognitive outcomes (e.g. Layzer *et al.* 2001; Sweet and Appelbaum, 2004; Miller *et al.* 2011; Filene *et al.* 2013; Rayce *et al.* 2017). The effects are

[15] r is Pearson's correlation coefficient.

[16] Maternal IQ was measured using the Weschler Abbreviated Scale of Intelligence which assesses cognitive ability across four subscales: vocabulary, similarities of constructs, block design, and matrix reasoning. From this, standardised measures of verbal ability, perceptual reasoning, and a full-scale measure of cognitive functioning, standardised 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.

[17] Available upon request.

also larger than the German home visiting program, *Pro Kind*, which finds average effect sizes for cognition of 0.20 - 0.30 SD for girls only at age 2 (Sandner and Jungmann, 2017). The effect sizes are also larger than a recent re-analysis of the Nurse Family Partnership Memphis trial which reported effects of 0.13 - 0.27 SD for cognitive skills at age six (Heckman *et al.* 2017). The *PFL* effects are more similar in magnitude to those found in studies of low and middle income countries. The Jamaica study, which is based on weekly home visits for two years starting between nine and 24 months, identified no significant effects on IQ at ages seven to eight , however the cognitive effects re-emerged at the 11, 17, and 22 year follow-ups with effect sizes ranging from 0.40 to 0.60 (Grantham-McGregor and Smith, 2016).

In addition to cognitive skills, another key aspects of children's development is their executive functioning skills, especially self-regulation, as independent of IQ, self-regulation has been shown to predict later academic performance, health, and finances in adulthood (Blair and Raver 2012; Liew 2012). However, children from disadvantaged backgrounds typically have poorer self-regulation (Evans and Rosenbaum 2008). Thus improving children's early skills in these domains could yield cascading benefits into adulthood (Diamond and Lee 2011). At age nine we found that the program had a large and substantive impact on all dimensions of the children's executive functioning skills with effect sizes ranging from 0.56 – 0.65 of a standard deviation. However, at age 14, we only find effects on working memory (0.52 SD), which means that the high treatment group are better able to retain, manipulate and use information over brief periods of time. There were no effects on their ability to override their automatic impulses (inhibitory control) or maintain and focus their attention (attention flexibility). In addition, the effect sizes on these measures were small and in the non-hypothesized direction, which suggests that the study is not under-powered to detect these effects; rather the effects observed at age nine have dissipated. There is no clear explanation as to why this occurred. It is possible that hormonal fluctuations and an overactive limbic system associated with the teenage period may impede their ability to delay gratification and focus (Arain et al. 2013). Perhaps relatedly, the program had no impact on the participant's attitudes towards anti-social behavior.

Similarly, the program had no impact on various dimensions of educational engagement including liking school, a sense of belonging at school, and school absences. Given the finding that the program raised cognitive scores, it is somewhat surprising that this did not translate into higher levels of school engagement or more positive feelings about

school among the high treatment group. However, it is important to note that the majority of the *PFL* cohort feel positive about school, with 76% and 68% in the high and low treatment group respectively like school, yet over 50% report missing school (without permission) in the last 12 months. Related to this, the results show that the program had no impact on changing time use patterns related to homework or internet/social media use. Over one-third of students report not doing any homework during the week, which is substantially higher than the 12% reported in the GUI cohort. The cohorts are somewhat more similar regarding social media use, with about 15% of the PFL cohort and 10% of the MCS cohort spending more than 7 hours a day on social networking sites.

While there are no significant differences across the groups about current school engagement, there is a statistically significant difference regarding future educational engagement with a significantly higher proportion of the high treatment group reporting that they intend to attend university (66% v 51%). In addition, while lower than the national average found in the GUI cohort (76%). This suggests that the program may be effective in improving the educational aspirations of the students.

The results also indicate that the significant effects observed for children's socio-emotional development at age four are no longer present at age 14, which is largely in line with the age nine findings. At earlier time points we found that the program was effective in reducing the proportion of children within the clinical range of behavioral problems, however, few effects were identified for continuous scores of children's socio-emotional development. These earlier measures were based on parent reports only. At age nine and 14 we assessed children's socio-emotional skills using child reports. At age 14, only two of the 22 sub-domains considered were statistically significant (and remained significant in the stepdown tests – the Brief Problem Monitors Attention Problems continuous score and the cutoff score. In particular, 41% of the high treatment group are classified as having clinically significant attention problems, compared to 63% in the low treatment group. Thus, the program was effective in reducing the incidence of attention problems, which means the high treatment group are less likely to have problems with concentration or sitting still. However, none of the other measures assessing different dimensions of socio-emotional skills, depression, self-esteem, or life satisfaction were impacted. In addition, a comparison of the *PFL* cohort to other nationally representative samples shows that the *PFL* cohort have poorer socio-emotional skills overall. In particular, >40% of the *PFL* cohort reach the cutoff for depressive symptoms, compared to only 16% in the national sample. This is consistent with

findings that families from low socioeconomic status communities face more mental health challenges than those from other communities (Kirkbride et al. 2024).

In-line with studies of other home visiting programs, there is little evidence that the program continues to have an impact on children's health. While few studies examine the long-term impact of home visiting programs on health, those that do typically find little evidence of effects on children's physical or mental health (Dumont *et al.* 2010; Kitzman *et al.* 2010; Minkowitz *et al.* 2007; Olds *et al.* 2004; Olds *et al.* 2007). At age 14, the *PFL* program had no impact on children's general health, health behaviours regarding substance use, and dietary intake. The only significant finding with a moderate effect size is on the waist-to-height ratio, which is primarily driven by a reduction in the participant's waist size (there is an almost 6cm difference between the high and low treatment group). As discussed earlier, WTH was used instead of BMI as there is evidence that it is a more reliable measure of obesity. That said, at previous waves (age 4 and 9), we assesses BMI. While no effects were found at age nine, there were significant differences at age four. In particular, a lower proportion of high treatment children (26%) were categorised as overweight or obese compared to the low treatment group (41%). The re-emergence of an effect at age 14 may be a result of using a different instrument. However, as identified at previous waves, a significant proportion of the *PFL* cohort are at risk of obesity (>40%). This compares with 21% in the GUI cohort.[18] This again speaks to evidence that children from lower socioeconomic backgrounds have an increased risk of obesity (Cronin et al., 2022). This is consistent with the finding that the *PFL* cohort experienced an earlier onset of puberty than the MCS cohort (who are of a similar age) in the UK. Earlier onset has been associated with poorer outcomes in adolescence such as more emotional and behavioral problems, as well as later in life such as cardiometabolic diseases (Day et al., 2015; Mensah et al. 2013).

For the first-time, the study measured the quality of the parent-child relationship from the perspective of the child. While only one of the results survived adjustment for multiple hypothesis testing, there is some evidence that the program improves the child's relationship with both their mothers and fathers (or father figures). The high treatment group are more likely to turn to their mothers when they are worried and they state that their parents (mothers and fathers) are more likely to help them with something important. In addition, they report better communication with mothers and more trust with fathers. The effect sizes of 0.3-0.7 indicate that these are sizeable impacts. These results are in contrast to findings at previous

[18] Note, in GUI at age 13, BMI was measured using self-reported height and weight. In PFL, measured waist size was used instead of weight.

waves where we find no effects on parent-child relationships when using parent reports (e.g., on the Condon Maternal Attachment Scale, the Maternal Separation Anxiety Scale, or the Parental Acceptance-Rejection Questionnaire. However, we did identify significant effects during the trial on certain dimensions of parenting related to parental behavior. For example, Doyle *et al.* (2017a) identify significant treatment effects on parenting skills at six and 18 months in terms of improving the quality of the home environment, O'Sullivan *et al.* (2017) find positive treatment effects regarding improved nutrition at 24 months, and Doyle and *PFL* Evaluation Team (2015) report improved parenting behaviour regarding the use of appropriate disciplinary techniques and increased parental interactions. These practices, interactions, and activities are recognised as key means of stimulating children's development (Farah *et al.* 2008), however, they may have also positively impacted the child's perception of their parents in adolescence.

Also, for the first time, we measured the participant's time and risk preferences. Time preferences are measured by asking participants how patient they are and then conducting a series of games where the participant can choose between a sooner, but smaller payment, and a larger, but later payment. Time preferences are important as previous studies have found that high time preferences (e.g. less patience) is associated with more disciplinary referrals in school, higher school dropout, less saving, and poorer health behaviors in adolescence (Benjamin, Brown, Shapiro, 2013; Castillo et al., 2011; Castillo et al., 2019; Sutter et al., 2013b). In general, the literature finds that children from lower socioeconomic families make more impatient choices (Sutter, Zoller, and Glätzle-Rützler, 2019). Indeed, if we compare the time preferences of the *PFL* and MCS cohorts (using the single self-assessment question) we find that the PFL cohort as a whole are less patient, however, the high treatment groups exhibits higher levels of patients than the low treatment group. While the difference is not statistically significant, the effect size of 0.31 suggests that the study may be underpowered to detect the effect, and the difference between the high treatment group and the MCS cohort is not statistically significant. This suggests that the program may have changed the time preferences, indicating the malleability of time preferences to early intervention.

Risk preferences are measured using a single item question asking participants how risky they are, as well as through a series of games that ask the participants to decide between a safe amount of tokens and a lottery that pays either a higher or lower amount of tokens than the safe alternative. In general, higher levels of risk taking are associated with poorer educational outcome (e.g., Castillo et al., 2018). Evidence suggests that children from lower

socioeconomic status families are more likely to take risks (Sutter, Zoller, and Glätzle-Rützler, 2019). However, we find that the *PFL* cohort are less risky than the MCS, although the difference is not statistically significant. The findings regarding risk preferences are less straightforward. We hypothesized that the program would result in higher levels of risk aversion among the high treatment group (e.g. less risky), however we found the opposite. In two of the games, the high treatment group were significantly less likely to take the safe option over the risky option. However, given that the PFL cohort already exhibit lower risk preferences than a nationally representative sample, the program's impact on increasing the likelihood to taking a risk, may not necessarily be an issue. While excessive risk taking may have negative consequences, exhibiting moderate level of risk may be an optimal strategy.

In total, we found that 25 of the 105 outcomes tested (24%) reached statistical significant in the individual tests, and 7 of the 20 stepdown tests (35%) were significant. As we used a 10% cutoff level, this indicates that these findings are unlikely to be a result of random Type I errors. We find that the *PFL* program has a long-term effect on children's cognitive development, with large effect sizes of 0.70 SDs. Significant effects are also found on working memory, attention, and educational expectations, however there are relatively few effects on health or socio-emotional outcomes. There is some evidence that the program reduced children's waist-to-height ratio, and improved parental-child relationships. While 43% of the original sample recruited during pregnancy participated at the Age 14 Follow-up, the treatment groups are still balanced on all key baseline characteristics. This is one of the few experimental home visiting programs that tracks participants into adolescence and finds evidence that *PFL* continues to have a significant impact on important dimensions of children's skills ten years after the families have finished the programme.

Although 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), the larger effect sizes identified for the *PFL* program, particularly for the cognitive outcomes, may be attributed to its prenatal start, its longer program length, its multiple connected treatments, and its inclusive eligibility criteria. In particular, *PFL* both starts earlier and is longer in duration than most other home visiting programs. The *PFL* home visitors worked with participants for a substantial and critical period of their children's lives; therefore the positive and sizable treatment effects may be a result of the strength and quality of the visitor-parent relationship which was given an appropriate length of time to build and develop. This is consistent with the home visiting literature which finds that the

bond between parents and program staff is key for understanding program effects (Wesley, Buysse, and Tyndall 1997).

The larger effects may also be attributed to the extensive and diverse supports offered to the high treatment group. The *PFL* treatment included baby massage classes during infancy and the *Triple P* program from age two, yet the majority of the other standalone home visiting programs, such as Nurse Family Partnership and its European equivalents, do not provide such supports. Therefore, a multi-component approach, which offers supports in a variety of formats and settings may help to engage families who favour one form of treatment over another. The larger effect sizes may also be attributed to the nature of the sample. Compared to many other home visiting programs which include ethnically diverse samples, the *PFL* cohort is relatively homogenous, consisting mainly of ethnically-Irish born participants. This, coupled with the individual-level randomisation in a confined geographical space, reduces variability within the sample, and allows us to uncover treatment effects if indeed they exist.

To conclude, the sizable cognitive advantages generated by the *PFL* program are likely to have positive impacts on the participant's outcomes throughout life. Thus it is critical to continue to track the *PFL* cohort as they progress through secondary school and potentially, into higher level education. One concern as we move forward with the *PFL* evaluation is sample size. While a response rate of 43% was achieved at the Age 14 Follow-up, this figure is likely to reduce further as the participants become older and start attending university or leaving the family home. Therefore maintaining the *PFL* cohort should be a priority if we are to assess the long term impact of the program. This is particularly important given the magnitude of the cognitive effects, especially in comparison to other intervention programs, as *PFL* can provide a model for other communities aiming to reduce long-term socioeconomic inequalities.

References

Aboud, Frances E., and Aisha K. Yousafzai. 2015. "Global Health and Development in Early Childhood". *Annual Review of Psychology* 66: 433–457.

Achenbach, T.M., McConaughy, S.H., Ivanova, M.Y., & Rescorla, L.A. 2011. *Manual for the ASEBA Brief Problem Monitor (BPM). Burlington,* VT: ASEBA, 1-33.

Almond, Douglas, Janet Currie, and Valentina Duque. 2017. "Childhood Circumstances and Adult Outcomes: Act II". NBER Working Paper No. 23017.

Anderson, Marti J., and Pierre Legendre. 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.

Arain, M., Haque, M., Johal, L., Mathur, P., Nel, W., Rais, A., …Sharma, S. 2013. "Maturation of the adolescent brain". *Neuropsychiatric Disease and Treatment 9*: 449–461.

Arim, R.G., Shapka, J.D., Dahinten, V.S., and Willms, J.D. 2007. "Patterns and correlates of pubertal development in Canadian youth: effects of family context". *Can J Public Health* 98(2): 91–6.

Avellar, Sarah, Diane Paulsell, Emily Sama-Miller, Patricia Del Grosso, Lauren Akers, and Rebecca Kleinman. 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.

Avellar, Sarah A., and Lauren H. Supplee. 2013. "Effectiveness of Home Visiting in Improving Child Health and Reducing Child Maltreatment." *Pediatrics* 132 (Supplement 2): S90.

Bailey, Drew, Greg J. Duncan, Candice L. Odgers, and Winnie Yu. 2017. "Persistence and Fadeout in the Impacts of Child and Adolescent Interventions". *Journal of Research on Educational Effectiveness* 10(1): 7–39.

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

Barlow, Jane, Hanna Bergman, Hege Kornør, Yinghui Wei, and Cathy Bennett. 2016. "Group-based Parent Training Programmes for Improving Emotional and Behavioural Adjustment in Young Children". *Cochrane Database of Systematic Reviews* Issue 8: Art. No.: CD003680.

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

Benjamin, D.J Brown, S.A. and Shapiro, J.M. 2013. "Who is "Behavioral"? Cognitive ability and anomalous preferences". *J. Eur. Econ. Assoc.* 11(6): 1231-1255.

Bennett, Cathy, Angela Underdown, and Jane Barlow. 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, Sandra, Paul Devereux, and Kjell G. Salvanes. 2009. "Like Father, Like Son? A Note on the Intergenerational Transmission of IQ Scores." *Economics Letters* 105 (1): 138–140.

Blair, C., and Raver, C.C. 2012. "Child Development in the Context of Adversity: Experiential Canalization of Brain and Behavior". *American Psychologist* 67(4), 309-318.

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

Bradley, Robert H., Bettye M. Caldwell, Stephen L. Rock, Craig T. Ramey, Kathryn E. Barnard, Carol Gray, … Dale L. Johnson. 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, Robert H., and Robert F. Corwyn. 2002. "Socioeconomic Status and Child Development." *Annual Review of Psychology* 53 (1): 371-399.

Britto, Pia R., Angelica Ponguta, Chin Reyes, and Romilla Karnati. 2015. *A Systematic Review of Parenting Programmes for Young Children in Low- and Middle-Income Countries*. New York, NY: United Nations Children's Fund.

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

Browning, L., Hsieh, S., and Ashwell M. 2010. "A systematic review of waist-to-height ratio as a screening tool for the prediction of cardiovascular disease and diabetes: 0.5 could be a suitable global boundary value". *Nutr Res Rev.* 23: 247–69.

Bus, A.G., and van Ijzendoorn, M.H. 1999. "Phonological Awareness and Early Reading: A meta-analysis of Experimental Training Studies". *Journal of Educational Psychology* 91(3): 403–414.

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

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

Castillo, J.. Jordan, L. and Petrie, R. 2018. "Children's rationality, risk attitudes and field behavior". *Eur. Econ. Rev.*, 102: 62-81.

Castillo, M., Ferraro, P.J., Jordan, J.L., and Petrie, R. 2011. "The today and tomorrow of kids: time preferences and educational outcomes of children". J. *Public Econ.* 95(11): 1377-1385.

Castillo, Marco, Jordan, Jeffrey, L. and Petrie, Ragan, 2019. "Discount Rates of Children and High School Graduation". The Economic Journal, 129(619): 1153–1181.

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

Chazan-Cohen, R., Raikes, H. H., and Vogel, C. 2013. "Program Subgroups: Patterns of Impacts for Home-based, Center-based, and mixed-approach programs". *Monographs of the Society for Research in Child Development 78*(1), 93-109.

Clements, D.H., Sarama, J., Wolfe, C.B., and Spitler, M.E. 2013. "Longitudinal Evaluation of a Scale-up Model for Teaching Mathematics with Trajectories and Technologies: Persistence of Effects in the Third Year". *American Educational Research Journal* 50(4): 812–850.

Côté, Sylvana, Massimiliano Orri, Richard E. Tremblay, and Orla Doyle. 2018. "A Multi-Component Early Intervention Program and Trajectories of Behavior, Cognition, and Health". *Pediatrics* 141(5): e20173174.

Coy, D. and Doyle, O. 2024. "Should Early Health Investments Work: Evidence from an RCT of a Complex Early Childhood Intervention". *Journal of Human Capital* 18(1): 74-104.

Cronin, F.M., Hurley, S.M., Buckley, T. *et al.* 2022. "Mediators of socioeconomic differences in overweight and obesity among youth in Ireland and the UK (2011–2021): a systematic review". *BMC Public Health* 22: 1585.

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

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

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

Cunha, Flávio, James J. Heckman, Lance Lochner, and Dimitriy V. Masterov. 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, Flávio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78: 883–931.

Day, F.R., Elks, C.E., Murray, A., Ong, K.K., and Perry, J.R. 2015. "Puberty timing associated with diabetes, cardiovascular disease and also diverse health outcomes in men and women: the UK Biobank study". *Sci Rep*. 5:11208.

Deardorff, J., Abrams, B., Ekwaru, J.P., and Rehkopf, D.H. 2014. "Socioeconomic status and age at menarche: an examination of multiple indicators in an ethnically diverse cohort". *Ann Epidemiol*. 24(10): 727–33.

De Graaf, I., Speetjens, P., Smit, F., de Wolff, M., and Tavecchio, L. 2008. "Effectiveness of the Triple P Positive Parenting Program on Behavioral Problems in Children: A Meta-analysis". *Behavior Modification 32*(5): 714-735.

Del Boca, Daniela, Christopher Flinn, and Matthew Wiswall. 2014. "Household Choices and Child Development." *The Review of Economic Studies* 81 137–185.

Del Bono, Emilia, Marco Francesconi, Yvonne Kelly, and Amanda Sacker. 2016. "Early Maternal Time Investment and Early Child Outcomes." *Economic Journal* 126: 96-135.

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

Derryberry, D., and Reed, M.A. 2002. "Anxiety-related Attentional Biases and their Regulation by Attentional Control". *Journal of Abnormal Psychology 111*(2), 225.

Diamond, A., and Lee, Kathleen. 2011. "Interventions Shown to Aid Executive Function Development in Children 4 to 12 Years Old". *Science* 333(6045): 959-964.

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

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

Doyle, Orla. 2020. "The First 2000 Days and Children's Skills". *Journal of Political Economy* 128(6) 2067–2122.

Doyle, Orla. 2024. "Can Early Intervention have a Sustained Effect on Human Capital?" *Journal of Human Resources* 59(5): 1599-1636.

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

Doyle, Orla, Nick Fitzpatrick, Caroline Rawdon, and Judy Lovett. 2015. "Early Intervention and Child Health: Evidence from a Dublin-based Trial". *Economics and Human Biology* 19: 224-245.

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

Doyle, Orla, Colm Harmon, James J. Heckman, Caitriona Logue, and Seong Hyeok Moon. 2017a "Measuring Investment in Human Capital Formation: An Experimental Analysis of Early Life Outcomes." *Labour Economics* 45: 40-58.

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

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

Doyle, Orla, Edel McGlanaghy, Eylin Palamaro Munsell, and Fionnuala M. McAuliffe. 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, Orla, 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, Orla, 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.

Doyle, Orla, and *PFL* Evaluation Team. 2016. *Final Report: Did Preparing for Life Improve Children's School Readiness*. Report to Preparing for Life Programme. Atlantic Philanthropies & Department of Children and Youth Affairs.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. *Using Randomization in Development Economics Research: A Toolkit. Handbook of Development Economics*, Elsevier.

DuMont, K., Kirkland, K., Mitchell-Herzfeld, S., Ehrhard-Dietzel, S., Rodriguez, M. L., Lee, E., ... and Greene, R. 2010. "A Randomized Trial of Healthy Families New York (HFNY): Does Home Visiting Prevent Child Maltreatment". *Rensselaer, NY: New York State Office of Children & Family Services and Albany, NY: The University of Albany, State University of New York.*

Eble, Alex, Peter Boone, and Diana Elbourne. 2016. "On Minimizing the Risk of Bias in Randomized Controlled Trials in Economics". The *World Bank Economic Review* 0: 1-12.

Elliott, S.N., and Gresham, F.M. 2008. *Social Skills Improvement System: Rating scales Manual.*

Elliott, Colin D., Pauline Smith, and Kay McCulloch. 2011. *British Ability Scales III.* Windsor, UK: NFER-Nelson.

Eslami, M., Pourghazi, F., Khazdouz, M., Tian, J., Pourrostami, K., Esmaeili-Abdar, Z., Ejtahed, H.S., and Qorbani, M. 2023. "Optimal cut-off value of waist circumference-to-height ratio to predict central obesity in children and adolescents: A systematic review and meta-analysis of diagnostic studies". *Front Nutr*. 4(9): 985319.

Evans, Gary W., and Rosenbaum, Jennifer. 2008. "Self-regulation and the Income-Achievement Gap". *Early Childhood Research Quarterly*, *23*(4), 504-514.

Fantuzzo, J., Tighe, E., and Childs, S. 2000. "Family Involvement Questionnaire: A Multivariate Assessment of Family Participation in Early Childhood Education". *Journal of Educational Psychology* 92(2): 367-376.

Farah, Martha J., Laura Betancourt, David M. Shera, Jessica H. Savage, Joan M. Giannetta, Nancy L. Brodsky, ... and Hallam Hurt. 2008. "Environmental Stimulation, Parental Nurturance and Cognitive Development in Humans." *Developmental Science* 11 (5): 793-801.

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

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

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

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

Furlong, Mairead, Sinead McGilloway, Tracey Bywater, Judy Hutchings, Susan M. Smith, and Michael Donnelly. 2012. "Behavioural and Cognitive-behavioural Group-based Parenting Programmes for Early-onset Conduct Problems in Children Aged 3 to 12 Years". *Cochrane Database of Systematic Reviews* Issue 2: Art. No.: CD008225.

Gertler, Paul, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, Susan Walker, Susan M. Chang, and Sally Grantham-McGregor. 2014. "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica." *Science* 344 (6187): 998-1001.

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

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

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

Grantham-McGregor, Sally and Joanne A. Smith. 2016. "Extending The Jamaican Early Childhood Development Intervention." *Journal of Applied Research on Children: Informing Policy for Children at Risk*: 7 (2), Article 4.

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

Heckman, James. J., Margaret L. Holland, Kevin K. Makino, Rodrigo Pinto, and Maria Rosales-Rueda, 2017. "An Analysis of the Memphis Nurse-Family Partnership Program." NBER Working Paper No. 23610.

Heckman, James J. and Tim Kautz. 2014. "Fostering and Measuring Skills Interventions that Improve Character and Cognition." In *The GED Myth: Education, Achievement Tests, and the Role of Character in American Life*, edited by James J. Heckman, John E. Humphries, and Tim Kautz, 341-430. Chicago, IL: University of Chicago Press.

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

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

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

James-Todd, T., Tehranifar, P., Rich-Edwards, J., Titievsky, L., and Terry, M.B. 2010. "The impact of socioeconomic status across early life on age at menarche among a racially diverse population of girls". *Ann Epidemiol*. 20(11): 836–42.

Kitzman, H. J., Olds, D. L., Cole, R. E., Hanks, C. A., Anson, E. A., Arcoleo, K. J., ... and Holmberg, J. R. 2010. "Enduring Effects of Prenatal and Infancy Home Visiting by Nurses on Children: Follow-up of a Randomized Trial among Children at Age 12 years". *Archives of Pediatrics & Adolescent Medicine 164*(5): 412-418.

Kirkbride, J.B., Anglin, D.M, Colman, I., Dykxhoorn, J., Jones, P.B., Patalay, P., Pitman, A., Soneson, E., Steare, T., Wright, T., and Griffiths, S.L. 2024. "The social determinants of mental health and disorder: evidence, prevention and recommendations". *World Psychiatry* 23(1): 58-90.

Knudsen, Eric I., James J. Heckman, Judy L. Cameron, and Jack P. Shonkoff. 2006. "Economic, Neurobiological, and Behavioral Perspectives on Building America's Future Workforce." *Proceedings of the National Academy of Science USA* 103 (27): 10155–10162.

Layzer, Jean I., Barbara D. Goodson, Lawrence S. Bernstein L, and C. Price. 2001. *National Evaluation of Family Support Programs. Volume A: The Meta-analysis. Final Report.* Cambridge, Mass: Abt Associates Inc.

Liew, Jeffrey. 2012. "Effortful Control, Executive Functions, and Education: Bringing Self-Regulatory and Social-Emotional Competencies to the Table". *Child Development Perspectives* 6(2): 105-111.

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

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

Martins, Lurdes, and Paula Veiga. 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.

Mensah, F.K., Bayer, J.K., Wake, M., Carlin, J.B., Allen, N.B., and Patton, G.C. 2013. "Early puberty and childhood social and behavioral adjustment". *J Adolesc Health*. 53(1): 118–124.

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

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

Millar, D. 2018. *School Attendance Data from Primary and Post-Primary Schools, 2016/17, Analysis and Report to the Child and Family Agency*. Dublin: Educational Research Centre.

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

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

Mingebach, Tanja, Inge Kamp-Becker, Hanna Christiansen, and Linda Weber. 2018. "Meta-meta-analysis on the Effectiveness of Parent-based Interventions for the Treatment of Child Externalizing Behavior Problems". *PLoS One* 13(9): e0202855.

Moore Heslin, A., O'Donnell, A., Kehoe, L., Walton, J., Flynn, A., Kearney, J., and McNulty, B. 2020. "Adolescent overweight and obesity in Ireland-Trends and sociodemographic associations between 1990 and 2020". *Pediatr Obes*. 18(2) :e12988.

Morawska, A., Mitchell, A. E., Burgess, S., and Fraser, J. 201). "Effects of Triple P parenting intervention on child health outcomes for childhood asthma and eczema: Randomised controlled trial". *Behaviour Research and Therapy* 83, 35-44.

Neuenschwander, R., Röthlisberger, M., Cimeli, P., and Roebers, C.M. 2012. "How do Different Aspects of Self-regulation Predict Successful Adaptation to School?" *Journal of Experimental Child Psychology* 113(3): 353-371.

Nisbett, Richard E., Joshua Aronson, Clancy Blair, William Dickens, James Flynn, Diane F. Halpern, and Eric Turkheimer. 2012. "Intelligence: New Findings and Theoretical Developments." *American Psychologist* 67(2): 130–159.

O'Sullivan, Aifric, Nick Fitzpatrick, and Orla Doyle. 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, David. L. 2002. "Prenatal and Infancy Home Visiting by Nurses: From Randomized Trials Community Replication". *Prevention Science* 3: 153–172.

Olds, David L., Harriet Kitzman, Robert Cole, JoAnn Robinson, Kimberly Sidora, Dennis W. Luckey, Charles R. Henderson Jr, *et al.* 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., Kitzman, H., Hanks, C., Cole, R., Anson, E., Sidora-Arcoleo, K., ... and Stevenson, A. J. 2007. "Effects of nurse home visiting on maternal and child functioning: age-9 follow-up of a randomized trial". *Pediatrics 120*(4), e832-e845.

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

Olds, D. L., Holmberg, J. R., Donelan-McCall, N., Luckey, D. W., Knudtson, M. D., and Robinson, J. 2014. "Effects of home visits by paraprofessionals and by nurses on children: follow-up of a randomized trial at ages 6 and 9 years". *JAMA Pediatrics 168*(2), 114-121.

Peacock, Shelley, Stephanie Konrad, Erin Watson, Darren Nickel, and Nazeem Muhajarine. 2013. "Effectiveness of Home Visiting Programs on Child Outcomes: A Systematic Review." *BMC Public Health* 13 (1): 17.

Rayce, Signe B., Ida S. Rasmussen, Sihu K. Klest, Joshua Patras, and Maiken Pontoppidan. 2017. "Effects of Parenting Interventions for At-risk Parents with Infants: A Systematic Review and Meta-analyses". *BMJ Open* e015707.

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

Roggman Lori A., Gina A. Cook, Carla A. Peterson, and Helen H. Raikes. 2008. "Who Drops Out of Early Head Start Home Visiting Programs?" *Early Education And Development* 19 (4).

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

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

Sanders, Matthews R., James N. Kirby, Cassandra L Tellegen, and Jamin J. Day. 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, Matthews R., Carol Markie-Dadds, and Karen Turner. 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, Malte, and Tanja Jungmann. 2017. "Gender-specific Effects of Early Childhood Intervention: Evidence from a Randomized Controlled Trial." *Labour Economics* 45: 59-78.

Sarganas, G., Schaffrath Rosario, A., and Neuhauser, H. 2017. "Resting Heart Rate Percentiles and Associated Factors in Children and Adolescents". *The Journal of Pediatrics 187:* 174-181

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

Sidora-Arcoleo, K., Anson, E., Lorber, M., Cole, R., Olds, D., and Kitzman, H. 2010. "Differential Effects of a Nurse Home-visiting Intervention on Physically Aggressive Behavior in Children". *Journal of Pediatric Nursing 25*(1): 35-45.

Sitnick, S. L., Shaw, D. S., Gill, A., Dishion, T., Winter, C., Waller, R., ... and Wilson, M. 2015. "Parenting and the Family Check-Up: Changes in Observed Parent-Child Interaction Following Early Childhood Intervention". *Journal of Clinical Child & Adolescent Psychology 44*(6): 970-984.

Smith, J., Dishion, T., Shaw, D., and Wilson M. 2014. "Coercive Family Process and Early-Onset Conduct Problems from Age 2 to School Entry." *Development and Psychopathology 26:* 917-932.

Smith, T.M., Cobb, P., Farran, D.C., Cordray, D.S., and Munter, C. 2013. "Evaluating Math Recovery: Assessing the Causal Impact of a Diagnostic Tutoring Program on Student Achievement". *American Educational Research Journal* 50(2): 1–32.

Sun, Y., Mensah, F.K., Azzopardi, P., Patton, G.C., and Wake, M. 2017. "Childhood social disadvantage and pubertal timing: a national birth cohort from Australia". *Pediatrics* 139(6): e20164099.

Sutter, Matthias, Zoller, Claudia, and Glätzle-Rützler, Daniela. 2019. "Economic behavior of children and adolescents – A first survey of experimental economics results". European Economic Review 111: 98-121.

Sutter, M. Kocher, M.G. Glätzle-Rützler, D., and Trautmann, S.T. 2013. "Impatience and uncertainty: experimental decisions predict adolescents' field behavior". Am. Econ. Rev. 103(1): 510-531.

Sweet, Monica A., and Mark I. Appelbaum. 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.

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

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

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

Weaver, Ian C.G., Nadia Cervoni, Frances A. Champagne, Ana C. D'Alessio, Shakti Sharma, Jonathan R. Seckl, Sergiy Dymov, Moshe Szyf, and Michael J. Meaney.

2004. “Epigenetic Programming by Maternal Behavior” *Nature Neuroscience* 7: 847–854.

Wesley, Patricia W., Virginia Buysse, and Sabrina Tyndall. 1997. “Family and Professional Perspectives on Early Intervention: An Exploration Using Focus Groups.” *Topics in Early Childhood Special Education* 17 (4): 435-456.

Williams, James, Sheila Greene, Sinéad McNally, Aisling Murray, and Amanda Quail. 2010. *Growing up in Ireland National Longitudinal Study of Children. The Infants and their Families*. The Stationery Office, Ireland.

Wilson, P., Rush, R., Hussey, S., Puckering, C., Sim, F., Allely, C. S., ... and Gillberg, C. 2012. “How Evidence-based is an 'Evidence-based Parenting Program'? A PRISMA Systematic Review and Meta-analysis of Triple P”. *BMC Medicine* 10(1): 130.

Yap, Marie B.H., Amy J. Morgan, Kathryn Cairns, Anthony F. Jorm, Sarah E. Hetrick and Sally Merry. 2016. “Parents in Prevention: A Meta-analysis of Randomized Controlled Trials of Parenting Interventions to Prevent Internalizing Problems in Children from Birth to Age 18”. *Clinical Psychology Review* 50: 138–158.

Zelazo, P.D., and Bauer, P.J. 2013. “National Institutes of Health Toolbox—Cognitive Function Battery (NIH Toolbox CFB): Validation for Children Between 3 and 15 Years”. *Mons of the Society for Research in Child Development* 78(4).