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

Orla Doyle	Caitriona Logue	Seong Hyeok Moon
University College Dublin	University College Dublin	Univsersity of Chicago

Abstract

This paper examines the impact of an Irish early childhood intervention on child development, child health and parenting outcomes at six months. The theoretical argument for targeting investment towards young, disadvantaged children is presented and the hypothesis is tested using data from the *Preparing for Life* programme. *Preparing for Life* is a five-year home visiting programme which starts during pregnancy and ultimately aims to improve children's school readiness. The programme is being evaluated using a Randomised Control Trial. This paper examines the impact of the programme at birth and up to six months of age. Due to the small sample size, permutation testing methods are adopted instead of traditional t tests, and fifty outcome variable are analysed in a nonparametric framework. To account for the increased likelihood of false discoveries in a multiple hypothesis model, nine groups of variables are constructed and a stepdown procedure is applied which strongly controls the family-wise error rate. The results indicate that the programme has no significant treatment effect for eight of the outcome categories, six months after birth. For one category, parental stress, there is evidence of a statistically significant and favourable treatment effect. This paper is the first to adopt this framework for testing outcomes at six months of age and provides the first rigorous evaluation of an early childhood intervention in a non-U.S. setting.

Keywords: Early childhood intervention, Randomised control trial, Multiple hypotheses, Permutation testing, Human capital formation.

JEL Classification: C12 , C93, J13, J24.

*Caitriona Logue' re earch i upported by The Innovation Academy. We thank Paul Devereux and Colm Harmon for their advice and u eful comment . The evaluation of the Preparing for Life programme i funded by the North ide Partner hip through the Department of Children and Youth Affair and The Atlantic Philanthropie . We would like to thank all tho e who participated and upported thi re earch, e pecially the participating familie and community organi ation , the PFL intervention taff, the Expert Advi ory Committee, and the Scientific Advi ory Board, and finally, the UCD Geary In titute evaluation team. Ethical approval for thi tudy wa granted by the UCD Human Re earch Ethic Committee, the Rotunda Ho pital' ethic committee and the National Maternity Ho pital' ethic committee.

1 Introduction

Intervening early in life to improve the skills of 'at risk' children, and thus reduce their later burden on society, is becoming an increasingly popular argument in the economics literature. Research on U.S. early childhood interventions, such as the *Nurse Family Partnership* and the *Perry Preschool Program*, has shown high rates of return to government investment. Children who received these interventions grew up to be less reliant on welfare, less likely to be involved in crime and were more likely to be employed as adults (Currie, 2001; Karoly et al., 2005; Heckman et al., 2010). These results indicate that such programs can be beneficial for the participants themselves and for society as a whole.

Yet to date, early childhood interventions have received relatively little attention in Europe. Given the social, economic, and cultural differences, especially with respect to the social welfare system, it cannot be assumed that the findings from the American studies can be extended to European countries. In this paper, we conduct one of the first rigorous evaluations of an early childhood programme in a non-U.S. setting. Specifically, we examine the first wave of outcome data from the *Preparing for Life* (*PFL*) programme, which is currently being implemented in a disadvantage community in Dublin, Ireland.

Socioeconomic inequality opens up early in life and tends to persist and widen into adulthood (Roberts, 1997; Victora et al., 2003). Therefore, the largest returns to early investment are found when the intervention targets higher risk samples (this is often defined as teen mothers, mothers from a low SES background, or mothers with drug or alcohol addictions). Karoly et al. (2005) calculate a \$2.88 return for every dollar invested in the *Nurse Family Partnership*, a home visiting programme open to first-time mothers who were either unmarried or of low income. However, their estimate of the benefit-cost ratio increases to \$5.70 when the sample is restricted to participants who were *both* unmarried and classified as having a low income background. Cunha and Heckman (2007) present a model of human skill formation which suggests that there are two key processes at play which explain the effectiveness of early intervention: early skill formation facilitates the accumulation of more advanced skills (self-productivity), which in turn makes investment throughout the lifecycle more productive (dynamic complementarity). Thus, they argue, that intervening early in life to strengthen disadvantaged children's foundation of skills is the most cost-effective way to bridge the inequality gap.

In line with these findings, the *PFL* programme targets a disadvantaged community. Census data from 2006, before the Irish economic crisis, showed that over 60 percent of residents in this community lived in

social housing, the unemployment rate was three times the Irish national average, and just five percent of residents had received some form of third level education. Such socioeconomic disadvantage is also evident in the local schools, with children in the *PFL* catchment area lagging behind when they enter the classroom environment. Doyle and McNamara (2011) find that teachers rated their students below a Canadian norm with respect to their physical health as well as their cognitive and non-cognitive skills. *PFL* was developed by community representatives and local service providers to address this problem. The ultimate aim of the early intervention is to improve children's school readiness. If the programme is successful, we hypothesise that the self-productivity and dynamic complementarity processes (Cunha and Heckman, 2007) will allow the inequality gap to dissipate over time.

PFL is a 5-year home visiting programme which starts during pregnancy and works with families up until the child starts school. The intervention is being evaluated using a randomised control trial design which involves the comparison of a 'high treatment' and a 'low treatment' group. In order to understand the effectiveness of the programme at different life stages, data is collected from the participants when the child reaches various developmental milestones (6 months, 12 months, 18 months, 2 years, 3 years, 4 years, 5 years). In this paper, we examine child and parent outcomes at the 6 months stage. This will allow us to determine whether or not the treatment effect has manifested itself at this early stage. Doyle et al. (2009) review the biological argument for intervening in early childhood when the brain is at its most malleable and susceptible to sensitive periods of development. The authors highlight the need to rigorously evaluate programmes that begin in utero to understand whether the postulated high returns to antenatal investment hold in reality. The majority of studies examining the effectiveness of early childhood programmes examine child outcomes later than the six months stage (Aos et al., 2004) or indeed later in adulthood (Olds et al., 1997; Reynolds et al., 2002; Heckman et al., 2010). Many of the studies that do examine six month outcomes are evaluations of programmes that begin at birth rather than prenatally (Guyer et al., 2003; Johnston et al., 2004). The limited results on six month outcomes indicate that intervention programs have mixed results at this early stage in a child's life, and few significant effects have been found (Anisfeld et al., 2004; Jungmann et al., 2009; Koniak-Griffin et al., 2000; LeCroy and Krysik, 2011).

One of the major contributions this paper makes to the literature is the use of frontier econometric methods for impact evaluation. The sample size in the *PFL* evaluation is small with data available for 173 participants at the six months stage. Traditional econometric methods assume large samples that follow normal distributions. Applying these methods when these assumptions do not hold could lead to incorrect estimates of the true treatment effect. For this reason, we use a permutation testing strategy that was applied in Heckman et

al. (2010) for a similar evaluation problem. This method places no distributional assumptions on the data being examined. Another statistical challenge is that the analysis of *PFL* involves studying many different outcomes relating to child development, child health, and parenting styles. Conducting individual tests on multiple outcomes increases the probability of Type I errors. We adopt the stepdown procedure developed by Romano and Wolf (2005) and applied in Heckman et al. (2010), a multiple hypotheses testing method which conducts joint significance tests on a family of outcomes. Thus, we can refute any accusation of 'cherry picking' favourable results.

The *Perry Preschool Program* was a landmark experiment that began in the 1960s and the participants have been followed into their 40s (Anderson, 2008; Heckman et al., 2010). Building on the lessons that have been learnt from this and other programmes, we aim to provide the first rigorous evaluation of a *modern* intervention using small sample methods. The randomisation process in Perry was violated. *PFL*, on the other hand, adopted design strategies to ensure that the experiment was not compromised. The randomisation process was carefully implemented and tested using baseline data, questions were included in each data collection wave to measure any contamination or spillover effects, and the evaluation team have access to implementation data to facilitate in-depth analyses. Building on lessons that have been learnt from the pioneering experiments, our methodology provides researchers with the tools to conduct pilot studies at a feasible scale.

The paper proceeds as follows. Section 2 describes the theory underlying human capital production and how this relates to early childhood intervention. In Section 3, we discuss the background to the*PFL* programme, the experimental design, and the data used in our analysis. We provide an in-depth description of the econometric framework in Section 4. We then present the results from our analyses in Section 5. Finally, in Section 6, we discuss the policy implications of our results and the strengths of our methodology.

2 Theoretical Context

A major aim of early childhood interventions is to reduce the inequality gaps that exist in society. In order to understand how effective this strategy is, the economics of human capital is reviewed. A theoretical model of skill formation, based on Cunha and Heckman (2007), which best captures the process is then presented.

2.1 Human Capital and Economic Inequality

Over the past 40 years, the human capital literature has evolved substantially. Initially, it was a contentious subject, one economists shied away from for fear of reducing man to 'a mere component, to something akin to property' (Schultz, 1961). Becker (1964) popularised the topic by focusing on education and job training, and their subsequent effects on wages. Today, the term 'human capital' has shed its negative connotations and is widely used when discussing the quality of an economy's labour force (Becker, 1993). Indeed the concept is no longer confined to the market sector and has been used to explain the demand for health (Grossman, 1972), the marriage market (Benham, 1974), fertility choices (Becker, 1964), and social organisation (Coleman, 1988).

Human capital theory not only explains the returns to self-investment but also the process by which parents invest in their children. The intergenerational transmission, however, is a complex process. Let us consider education, one of the most important inputs in the human capital production process (Becker, 1964). Higher educated parents tend to have higher educated children, yet as Plug and Vijverberg (2003) point out, it is hard to disentangle the impact of nature and nurture in this relationship. Individuals with more ability are more likely to pursue education (Mincer, 1958), and ability is difficult to control for. It is difficult to determine how much of transmission is simply genetic (ability is the driving force) and how much represents a human capital investment (individuals invest in education to improve their children's outcome). Currie and Moretti (2003) find a positive, causal link between mothers' education levels and their children's birth outcomes: higher educated mothers produce higher birth weight babies, they were more likely to be married at birth and their smoking reduced during pregnancy. Each of these birth outcomes are understood to be important channels for child health, which in turn could impact upon educational success. Black et al. (2003), however, find little evidence of a causal effect of maternal education on children's schooling, which suggests that the relationship can be explained through a genetic link of innate characteristics.

Becker and Tomes (1979) produce a theoretical model which explains the family influence on a child via *both* mechanisms. Child outcomes can depend on *both* the degree of inheritability of parental endowments and a parent's propensity to invest in their child's human capital development. The model is constructed such that children's life outcomes are determined by the utility maximizing behaviour exhibited by their parents and a random shock in any given period. This framework goes some way in explaining why income distributions tend to be skewed. The authors show that even if all families are assumed to be the identical in the first period, an unequal distribution of random market luck will allow the lucky families to invest more

in their children, who in turn have more income to invest in their subsequent children, leading to skewed equality distribution in later generations.

2.2 Why Intervene Early in Life?

This leads us to examine the ways in which early childhood interventions can break the inequality cycle. As Currie (2001) explains, equity can be restored either by compensating for differences in final outcomes, or equalising initial endowments, or both. Early investment, she argues, avoids the moral hazard associated with compensating poorer individuals later in life, and is also likely to be more cost effective in the sense that 'an ounce of prevention is worth a pound of cure' (Henry De Bracton, 1240). It is commonly accepted that most parents are altruistic and receive utility from their children's well-being. However, income constraints prevent poorer, altruistic parents from making efficient investments (Becker and Murphy, 1998). As Cunha and Heckman (2007) highlight, the biggest market failure in the human capital production process is 'the inability of children to buy their parents', they cannot borrow against future life earnings to secure the family resources that will enhance their future (Heckman and Masterov, 2007).

Government intervention is needed not only to alleviate income constraints but also to tackle the issue of information failures. For example, Currie (2001) describes the empirical evidence which shows that parents struggle to evaluate their child care options. Early childhood interventions can overcome information failures by teaching parents how best to nurture their children. Research on the important aspects of childhood development has evolved considerably. Traditionally, the economics literature has focused solely on the importance of cognitive skills. However, it is now widely accepted that physical health, social skills, and emotional capabilities are just as important for children to develop into successful adults. For example, Hanushek and Woessmann (2008) examine the importance of developing cognitive capabilities at school, Smith (2009) highlights the link between child health and successful life outcomes, while Heckman and Masterov (2007) discuss the growing evidence on the importance of non-cognitive, as well as cognitive, skills. Indeed the Early Development Instrument (Janus et al., 2005), one of the most widely used measures of children's readiness to learn, defines five categories of school readiness which include *social competence*, *emotional maturity*, *physical health and well-being* as well as two cognitive categories, *language and cognition* and *communication skills and general knowledge*. This shows that child development is multifaceted and health and behavioural qualities are just as important as cognition.

2.3 Theoretical Framework

Cunha and Heckman (2007) present a technology of human skill formation which is widely cited in the literature using an overlapping generations model. In this framework, it is assumed that an agent lives for two periods, as a child in the first and as a parent in the second. The authors consider the following CES (constant elasticity of substitution) production function:

$$(1) \qquad h' = m_2(h; \ _1; [\ (I_1)^{\phi} + (1- \ )(I_2)^{\phi}]^{1/\phi})$$

Where h' represents an adult stock of skills, h represents parental characteristics, $_1$ is the endowments that a person is born with in period 1, is the skill multiplier ($0 \leq \ \leq 1$), represents the degree of complementarity between early and late investment ($\ \leq 1$), and I_1 and I_2 represent investment during childhood and adulthood respectively.

It should be noted that the skill multiplier relates to the process by which investment in the first period produces skills which allows investment in the second period to be more productive, a combination of self-productivity and dynamic complementarity. Following a typical CES production function, when $\ \rightarrow -\infty$, early and late investment are perfect complements, while $\ = 1$ when early and late investment are perfect substitutes. The budget constraint facing parents can then be denoted as follows:

$$(2) \qquad c_1 + I_1 + \frac{c_2 + I_2}{(1+r)} + \frac{b'}{(1+r)^2} = wh + \frac{wh}{(1+r)} + b$$

Where w represents wages, r is the interest rate, c_1 and c_2 is the household consumption in periods 1 and 2 respectively, and b is the bequest that an individual receives when he becomes an adult (b' is then the bequest that the adult leaves to their child). There are two final factors that need to be considered before looking at the utility maximising equation. Individuals have a discount rate, , and a degree of altruism towards their child, . Now, let $u(.)$ denote the utility function. The parent uses the technology presented in equation (1) to maximise the following subject to the budget constraint (2):

$$(3) \qquad V(h; b; \ _1) = max\{u(c_1) + \ u(c_2) + \ ^2 \ E[V(h'; b'; \ _1')]\}$$

The optimal ratio between early and late investment can then be derived as follows:

$$(4) \qquad \frac{I_1}{I_2} = [\frac{}{(1- \ )(1+r)}]^{\frac{1}{1-\phi}}$$

As $\ \rightarrow -\infty$, the ratio approaches unity and, thus, the skill multiplier share parameter does not play a significant role. As $\ \rightarrow 1$ the skill multiplier becomes more important suggesting that the bulk of investment should be made early if the skill multiplier is high. (Cunha and Heckman, 2007).

The empirical evidence suggests that a high skill multiplier effect plays a real role in the development process. Coneus and Pfeiffer (2007) find that an increase in birth weight leads to an increase in children's cognitive and non-cognitive skills. An increase in non-cognitive skills, in turn, has a positive effect on children's verbal skills. Black et al. (2007) also find that birth weight has a positive impact on adult outcomes. It should be noted that birth weight is a proxy for child health which is considered to be one attribute of human capital. Parents can invest in their children's development in utero by eating nutritiously and not engaging in risking health behaviour such as smoking and alcohol consumption all of which will impact on birth weight and future health.

The literature also presents evidence of complementarity between early and late investment. Pfeiffer and Reuß (2008) show that a Cobb-Douglas function can be assumed while Caucutt and Lochner (2011) find the elasticity of substitution between early and late investment to be 0.37. Similarly, Cunha et al. (2010) find evidence of complementarity for both cognitive and non-cognitive skills. Using a one stage model of childhood they find the elasticity of substitution is below one across various specifications. Combining this evidence with equation 4 above indicates that it is not easy to compensate for a weak foundation through later life investment. The level of malleability decreases as children get older and, therefore, it is importance to intervene early.

2.4 How to Intervene

Karoly et al. (2005) analyse the empirical evidence from twenty well-known early childhood interventions which can be categorised as either centre-based, home-based, or parental education programmes. The programme we examine in this paper, *PFL*, is classified as a home visiting programme with an element of parental education.

While centre-based programmes are considered to be child focused, the home visitation method targets two-generations (Karoly et al., 2005). Brooks-Gunn et al. (2000) explain that the family-focused approach has become increasing popular due to a strong belief that parental outcomes serve a mediating role in child development. However, as Shonkoff and Phillips (2000) point out, changing parenting behaviour is extremely difficult to do. Fifteen of the 20 interventions examined by Karoly et al. (2005) measure outcomes for mothers and only three of these have displayed a significant impact on parental outcomes and the effect sizes were small. Kahn and Moore (2010) synthesize the findings from 66 home visiting programmes that

were evaluated using randomised control trials. The authors indicate that there is weak evidence that one programme was effective at reducing substance abuse by parents, while seven interventions had a positive and significant effect on parenting practices. Karoly et al. (2005) state that even a small effect can have a large long-run impact on child development. Therefore, it is important to examine how *PFL* impacts upon parenting behaviour, caregivers' smoking and alcohol consumption during pregnancy as a proxy for for prenatal investment, as well as child development and child health. As Currie (2009) summarises, researchers need to open 'the black box' of the family in order to understand the impact of a child's background on their long run success.

This paper focuses on parental and child outcomes when the children were 6 months old. To date, few home visiting evaluations have focused on outcomes at this early stage of development and those that have show quite mixed results. Testing whether the returns to investment during pregnancy and early infanthood can be identified as early as six months will aid understanding of the mediating factors for successful child outcomes.

3 Preparing for Life: Programme Design and Impact Data

PFL has been designed as a randomised control trial evaluation. However, ethical considerations prevented the division of participants into a traditional treatment group and a pure control group. Thus, the evaluation is based on a dosage experiment whereby all participating families receive some low level of treatment and the group randomised into the high treatment group receive a higher intensity treatment. All participating families are provided with developmental toys, facilitated access to preschool, and public health information. The participants in the low treatment group have access to an information officer who informs them about services in the area and upcoming *PFL* events, while participants in the high treatment group receive the additional provisions of a home visiting mentoring service and group parent training. The home visiting mentors, from various professional backgrounds, act as advisors to the participating mothers. They have been trained to support and educate parents about child development using Tip Sheets. Each family has an assigned mentor who visits the home up to once a week for between 30 minutes and two hours. While the *PFL* manual prescribed weekly visits, the majority of families opt for fortnightly visits while some engage monthly. It should be noted that the group parenting programme does not begin until the child is two years of age. In this paper, we examine child outcomes at the six months stage, thus a comparison of the high and low treatment groups will allow us to extract the sole impact of the home visiting component.

The U.S. *Nurse Family Partnership* programme bears the closest resemblance to the *PFL* initiative. Both start during pregnancy and provide participants with home visits from a trained professional who focuses on improving birth outcomes (by discussing nutrition, alcohol, tobacco and illegal drug consumption) and later child outcomes (via semi-structured discussions about child development). *The Nurse Family Partnership* programme works with families up until the child reaches two years of age, while *PFL* continues to work with children up until age five. *PFL* is also more intensive in that participants are provided with additional provisions to supplement the home visiting element (group parent training, social events, baby massage classes).

3.1 Recruitment

Recruitment into the *PFL* programme took place between January 2008 and September 2010. All pregnant women residing in the *PFL* catchment area were eligible to take part, regardless of income or family background. Eligible candidates were identified using hospital records and community referral. After consenting to take part in the programme, yet before engaging with the implementation team, the participant was assigned their level of treatment using an unconditional randomisation procedure. The randomisation programme used a random number generator so that each participant had an equal chance of being allocated to the high or low treatment group. A baseline survey was administered to 205 (low = 101; high = 104) participants before they began the programme. Analysis of this data reveals that the randomisation process was successful (Doyle et al., 2010). One hundred and twenty-three variables were analysed using permutation testing and no significant differences were found between the high treatment and low treatment groups for 119 (97 percent) of them.

3.2 Data

3.2.1 Description of Participants

Table 1 provides the descriptive statistics of the sample at baseline. It should be noted that the sample size is smaller than the total number of participants recruited as there was some attrition which is discussed in the results section. Mothers were 26 years old on average when they joined the programme, they were 21 weeks pregnant, approximately 82 percent had a partner, almost half were first time mothers, 33 percent had a low level of education[1], and the average level of cognitive resources was 81.54 measured using the *Wechsler*

[1]Participant are cla ified a having a low level of education if they indicated that their highe t level of education wa the Junior Certificate or lower. The Junior Certificate i an Iri h tate exam which i completed at 15 year of age.

Abbreviated Scale of Intelligence (*WASI*, Wechsler, 1999). Approximately 70 percent of mothers indicated they had at least one physical illness previously. A high proportion indicated that they had a mental health condition (26 percent). With respect to substance use, 50 percent of participants indicated that they had smoked during pregnancy, 26 percent stated that they had drank alcohol at some stage during pregnancy, and 16 percent of respondents indicated that they had used an illegal drug at least once in their lives.

Participants' average level of maternal attachment was measured using the *Vulnerable Attachment Style Questionnaire* (*VASQ*; Bifulco et al., 2003). A score above 15 indicates vulnerability for depressive disorders. The sample mean was above this cutoff (M = 17.96). Approaches to parenting prior to the intervention were measured using the *Adult Adolescent Parenting Inventory 2* (*AAPI-2*; Bavolek and Keene, 1999). This scale indicates a parent's tendency towards abuse and neglect. The average raw score (M= 118.51) indicates there was a moderate to small risk of these types of behaviours in our sample, based on standardised scores for the U.S. population.

Table 1: Baseline Descriptive Statistics for Participants Examined At Six Months

Variables	**N**	**Mean**	**SD**
Mother’ age	171	25.68	5.89
Week in pregnancy at programme entry	171	21.47	7.33
Partnered	171	0.82	0.39
Married	171	0.17	0.37
Living with parent( )	171	0.50	0.50
Fir t time mother	171	0.49	0.50
Low education	171	0.33	0.47
Mother employed	171	0.41	0.49
Save regularly	171	0.52	0.50
Social hou ing	171	0.55	0.50
‡Cognitive re ource (WASI)	169	81.54	13.10
Phy ical Health Condition	171	0.70	0.46
Mental Health Condition	171	0.26	0.44
Smoking during pregnancy	171	0.50	0.50
Drinking during pregnancy	171	0.25	0.43
Drug ever u ed	171	0.16	0.37
Vulnerable attachment (VASQ)	171	17.97	3.90
Po itive parenting attitude (AAPI)	171	118.51	13.30

‡ Mea ured at 3 month po tpartum

3.2.2 Six Month Outcome Measures

The analysis in the paper is based on child and parent outcomes reported at six months. These include the following measures.

Child Development

The child's level of development at six months of age is measured using three standardised measures: the *Ages and Stage Questionnaire* (*ASQ*; Squires et al., 1999), the *Ages and Stages Questionnaire: Social-Emotional* (*ASQ:SE*; Squires et al., 2003) and a difficult temperament assessment based on the *Infant Characteristics Questionnaire* (Bates et al., 1979).

Ages and Stage Questionnaire Using the *ASQ* questionnaire, that specifically relates to child development at six months of age, mothers were asked to indicate whether their children could perform thirty different tasks. There are three possible answers to each question: yes (10 points), sometimes (5 points), and no (0 points). These items are divided into five subdomains entitled *communication*, *gross motor*, *fine motor*, *problem solving*, and *personal-social* each with possible scores ranging from zero to 60. Higher scores are indicative of more advanced child development. This self-assessment measure has been shown to be highly correlated with other direct assessments of child development (Squires et al., 1999).

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

Difficult Temperament The mothers were asked seven questions which are based on the *Infant Characteristics Questionnaire* (Bates et al., 1979). The questions relate to whether the child gets upset easily or demonstrates fussy behaviour. Each item is scored on a six point scale resulting in a total possible

range of one to 42, where higher scores are indicative of a more difficult temperament.

Child Health

To examine the impact of the *PFL* programme on child health outcomes, we analyse three main areas. *Infant Health* relates to birth weight, breathing problems, and health problems that require hospital or GP care. *Mother's Health Decisions for her Infant* relates to parental behaviour which directly impacts upon the child's health. For example, her decision to get her child vaccinated, her feeding choices, and her knowledge of her baby's weight (a proxy for the mother's interest in her child's health). Finally, we examine *Sleep Routines*. Sleep is an extremely important aspect of a baby's life. This is evident when one considers that newborn babies sleep for 16 hours per day. Sleep impacts upon child development and parental failure to establish a bedtime routine is cited a major cause of sleep problems in children (Jaffa et al., 1993).

Parenting

In order to ascertain whether *PFL* has an impact on maternal parenting behaviour we examine four psychometric measures: *Parental Locus of Control* (*PLOC*; Campis et al., 1986), *Condon Maternal Attachment Scale* (*CMAS*; Condon and Corkindale, 1998), *Parenting Stress Index* (*PSI*; Abidin, 1995) and*Parental Cognition and Conduct Toward the Infant Scale* (*PACTOIS*, Boivin et al., 2005).

Parental Locus of Control This measure contains 20 items which are divided into 5 subdomains. *Parental Efficacy* relates to the mother's belief in their parental ability, *parental responsibility* indicates whether the mother believes they are responsible for their child's behaviour, *child control of parent's life* relates to the mother's belief that her life is dominated by her child, *parental belief in fate* measures the extent to which the mother believes that luck determines the child-parent relationship, and *parental control of child's behaviour* indicates whether the child believes they can control their child. The possible scores for each domain range from four to 20 with higher scores indicating a more external locus of control, which is considered a negative outcome.

Condon Maternal Attachment Scale This measure relates to the mother's feelings for her child, the pleasure she gets from her child, and her tolerance in the maternal role. There are 19 items on this scale which are organised into three categories. *Quality of attachment*, *pleasure in interaction*, and *absence of hostility*. Each is measured on a scale of one to five with higher scores indicating a more favourable outcome.

Parenting Stress Index The short version of this questionnaire, consisting of 36 items, was administered to participants. There are three subdomains. *Difficult child* relates to the presence of behavioural characteristics in a child which are difficult for a parent to manage, *parenting distress* relates to personal factors that affect parenting stress such as the presence of depression or a lack of social support, and *parent-child dysfunctional interactions* relates to the parent's perception that their baby is a negative element of their life. Possible scores for each domain range from 12 to 60 with higher scores indicating more maternal stress.

Parental Cognition and Conduct Towards the Infant Scale The *PACTOIS* scale measures the participants' view of their parental role. The questionnaire consists of 26 items which are divided into five subdomains: *parental self-efficacy*, *perceived parental impact*, *parental hostile-reactive behaviour*, *parental overprotection*, and *parental warmth*. Scores for each domain range from one to 11. Higher scores in the parental hostile-reactive behaviours and parental overprotection domains indicate a more negative outcome, while higher scores in the three remaining domains are indicative of a more positive parental role.

Substance Use During Pregnancy

We examine three binary measures which indicate whether or not the participant smoked tobacco, consumed alcohol, or took an illegal drug during pregnancy.

4 Econometric Framework

4.1 Statistical Issues in Small Samples

The evaluation of the *PFL* programme has been designed as a randomised control trial which allows us to extract the causal impact of the intervention by comparing the outcomes of the low and high treatment groups. However, using conventional methods to test for significant differences between these groups is not viable given the small sample size. Traditional t tests rely on the assumption that data follow a normal distribution or, if the sample is large, the central limit theorem ensures that the t statistic is normally distributed. In a small sample such as ours, the distribution is unlikely to be known. In addition, many of the variables included in the analysis are binary indicators. While the Mann-Whitney test is a non-parametric alternative, it is generally accepted that this test lacks statistical power (e.g., Siegel, 1957; Moir, 1998), a

problem which is exasperated in small samples. Fisher (1935) developed a permutation testing method which demonstrated attractive statistical qualities, yet when it was originally developed it was not considered viable due to the computation intensity required (Bradley, 1968). Technological advances have since facilitated the application of these methods in empirical work. Kennedy (1995) documents their introduction to the field of econometrics in the 1990s.

Permutation methods do not depend on distributional assumptions and Green (1977) discusses various studies which have found comparable levels of power for t tests and the permutation testing equivalent, asymptotically and for samples that are normally distributed. RCTs are often considered too costly and practically difficult to implement on a scale that allows for a sufficiently large sample size on which classical testing procedures can be applied. Permutation tests can be applied to small samples and thus they facilitate feasibly sized pilot studies. While two recent studies examining American early childhood interventions have used this approach (Anderson, 2008; Heckman et al., 2010), it has yet to be extensively explored in the policy evaluation literature.

4.2 Permutation Testing

A permutation test is a method whereby an observed test statistic is compared to a distribution of test statistics constructed using multiple, random permutations of the data. The only restriction that is placed on the data is that of exchangeability under the null hypothesis. In other words, if there are no significant differences between the high and low treatment groups, it should be possible to exchange the treatment labels without affecting this result (see Good, 2005). In practice, our permutation testing procedure starts by recording the observed t statistic that results from testing the equivalence of means for the low and high treatment groups. The treatment labels are then shuffled multiple times (1,000 replications are permuted using Monte Carlo simulations in our analyses) and a separate t test is performed using each new permutation of the data. The p-value is computed by examining the proportion of permutations that have a t statistic greater than or equal to the observed statistic in the original sample. If the proportion is small ($p < 0.1$), we can deduce that the original statistic is not simply driven by chance and reject the null hypothesis. It should be noted that we report right-sided, mid-p-value in our results. Upton (1992) explains that the mid-p-value is more suitable when dealing with discrete data. The mid-p-value, MP, is calculated as follows:

$$MP(T) = P(T* > T) + 0{:}5P(T* = T)$$

where P is the probability distribution, T* is the randomly permuted t statistic and T is the observed t statistic. This leads to a less conservative estimate than the tradition p-value. We use right-sided testing in order to test whether the high treatment is having a positive effect on child and parenting outcomes.

4.3 The Issue of Multiple Hypotheses

The evaluation of this programme involves examining a large number of outcomes. Specifically, we conduct permutation tests for 50 individual outcomes. This increases the likelihood of committing a Type I error (rejecting a null hypothesis when it is in fact true) and studies of RCTs have been criticised for overstating treatment effects as a result of this multiplicity effect (Pocock et al., 1987). To address this problem, methods have been developed for jointly testing the null hypotheses which control the Family-Wise Error Rate (FWER, the probability of rejecting at least one true null hypothesis) at a pre-determined level, . The complement of the joint null hypothesis is that at least one single hypothesis is false. Testing for this involves adjusting the p values associated with individual tests to account for the effect of multiple outcomes.

The Bonferroni method, which is very common in the literature, is a simple example of this whereby the required p value for statistical significance is adjusted by dividing it by the number of hypotheses being tested. This method, however, results in a loss of power and does not take account of the interdependence across the outcomes. Holm (1979) develops a stepdown procedure (the Bonferroni-Holm method) which is less conservative as a result of simply ordering the hypotheses in terms of their significance level, and sequentially reducing the family size denominator. Thus, as Westfall and Wolfinger (1997) explain, this procedure readjusts the required p-value to account for situations where some hypotheses are in fact false. Further improvements in power can also be made if the correlation structure among the variables being jointly tested is considered. Westfall and Young (1993) develop a method which relies on the assumption of 'subset pivotality' for strong FWER control. This condition, which was coined by the authors, requires that the multivariate distribution of p values for a subset of hypotheses is the same regardless of whether the excluded hypotheses are true or false. Romano and Wolf (2005) improve this method with a weaker assumption, that of monotonicity with respect to the critical values. This ensures that the stepdown procedure is ordered so that the most significant individual hypothesis in a family of outcomes is the first to be tested. This method is applied in Heckman et al. (2010) and we employ the same approach for our multi-hypothesis analyses. The procedure is described in the next section.

4.4 The Stepdown Procedure

The stepdown procedure involves firstly calculating a test statistic for each null hypothesis in a family of outcomes. We use the t statistic to test the equivalence of means for the low and high treatment group in the original sample. Using the permutation testing method, described above, 1,000 additional t statistics are generated by randomly exchanging the treatment labels; this constructs the appropriate t distribution under the null hypothesis. The stepdown procedure starts by extracting the largest observed t statistic and comparing it with the distribution of maximum statistics for the joint hypotheses. If the probability of observing this statistic by chance is high ($p \geq 0.1$) we accept the joint hypothesis and the procedure stops. In other words, we would deduce that the high treatment does not have a statistically significant impact on the family of outcomes being tested and there is no need to drill down further to examine subsets of the hypotheses. On the other hand, if the probably of observing this t statistic is low ($p < 0.1$) we reject the joint null hypothesis. We can then proceed by excluding the most significant hypothesis and testing the subset of hypotheses that remain for joint significance. This process of dropping the most significant hypothesis continues until the resulting subset of hypotheses is accepted, or only one hypothesis remains. Stepping down through the hypotheses in this manner allows us to isolate the hypotheses which lead to a rejection of the null.

One limitation of the stepdown procedure relates to the grouping of variables into suitable families for joint hypothesis testing. As Westfall and Young (1993) explain, there is a trade off between family size and the ability to detect an effect. If too many variables are included in a family, the p-values adjustments become too severe. Therefore, including all fifty of the *PFL* outcome variables in one single category is not feasible (although we test it for robustness). Grouping variables is a somewhat arbitrary process. One possible approach is to use factor analysis and choose families based on the interdependencies found in the data. For this study, a theoretical approach was adopted and variables were grouped according to hypothesised interdependence structures. This resulted in four families relating to the two primary outcomes (child development and child health) and five families relating to the two secondary outcomes (parenting and parental alcohol, smoking, and substance use during pregnancy). Standardised psychometric measures provided a logical grouping; subdomain measures on each psychometric scale were always grouped together. Other measures were grouped thematically. The results of our application of the Romano and Wolf (2005) method to these nine families of outcomes are contained in the next section.

5 Results

5.1 Six Month Analysis

We divide our results into four main categories: child development, child health, parenting and substance use during pregnancy. We present the mean outcome scores by group, the p-values that result from individual permutation testing, the adjusted p-values calculated using the stepdown procedure, and Cohen's d effect sizes. It is important to note that in order to execute the stepdown method, it was necessary to ensure that a higher score on all measures was indicative of a favourable outcome. Therefore, any variables that measure the tendency towards a negative outcome were reverse scored before applying the stepdown method.

The outcomes within a category are ordered from most significant to least significant based on the individual right-sided permutation results. This helps to interpret the stepdown p-values that are presented. Each adjusted p-value represents the likelihood of rejecting the joint null hypothesis for the variable in the corresponding row and each variable below it. Thus, as we step down through the hypotheses, the most significant variables are dropped for the joint test.

5.1.1 Child Development

Table 2 presents the results for the child development outcomes. First, the results from the individual tests indicate that none of p-values are statistically significant at the 10 percent level. It logically follows that the stepdown procedure fails to reject the null hypothesis of no treatment effect on child development outcomes at the six months stage. In addition, the effect sizes for each outcome are small. The mean scores indicate that the children in the high treatment group are scoring more favourably on five out of seven measures relating to cognitive development and behavioural aspects. The *ASQ Gross Motor* score and the *ASQ Communication* score are close to the $p < 0.1$ cutoff, yet we cannot claim that the differences are precisely determined.

Table 2: Comparison of High and Low Treatment Outcomes: Child Development

Outcome	N (NHIGH N LOW)	M HIGH (SD)	M LOW (SD)	p-values Individual Test[1]	p-values Stepdown (Adj.)[2]	Effect Size (d)
ASQ Scores & Difficult Temperament						
ASQ Gross Motor Score	17	40.78	8.50	0.124	0.490	0.18
	(8 /90)	(11.9)	(12.99)			
ASQ Communication Score	17	5 .07	51.78	0.154	0.511	0.16
	(8 /90)	(7.84)	(8.49)			
‡Difficult Temperament	17	11.70	12.21	0.278	0.712	0.09
	(8 /90)	(5.71)	(5.50)			
ASQ Personal Social Score	17	46.69	45.94	0. 61	0.755	0.06
	(8 /90)	(12.10)	(1 .57)			
‡ASQ Social-Emotional Score	17	14.76	15.17	0. 85	0.746	0.0
	(8 /90)	(10.68)	(1 .75)			
ASQ Fine Motor Score	17	50.78	51. 9	0.671	0.854	0.06
	(8 /90)	(9.48)	(10.17)			
ASQ Problem Solving Score	17	51.87	52.56	0.710	0.710	0.07
	(8 /90)	(9. 9)	(9.92)			

Notes: 'N' indicates the sample size. 'M' indicates the mean. 'SD' indicates the standard deviation.
[1] one-tailed (right-sided) p-value from an individual permutation test with 1000 replications. [2] one-tailed (right-sided) p-value from a stepdown permutation test with 1000 replications. ** Significant at the 5 percent level * Significant at the 10 percent level. ‡ Represents a negative outcome and was reverse scored for the stepdown analyses. The sample sizes reported are those used in the individual tests and may differ from the sample size used in the stepdown procedure. For the stepdown procedure, any observations that have missing data for measures within the stepdown family are dropped. The variables are reported in the order in which they are dropped from the stepdown procedure.

5.1.2 Child Health

To examine child health, we divide the outcomes into three families of variables. Table 3 displays twenty-three individual permutation test p-values and the results from the three separate stepdown procedures. With respect to *Child Health*, no significant differences were found between the low and high treatment group. The treatment groups are similar in child health outcomes, including birth weight and the effect sizes are very small.

With respect to *Mothers' Health Decisions for Her Infant*, we find three significant differences at the five percent level using the individual permutation tests. Specifically, parents in the high treatment group were more likely to feed their baby appropriate food, their children were more likely to have received the recommended immunizations, and their children were more likely to have a suitable feeding routine. Medium effects are found on the three significant outcomes. However, based on the stepdown results, we fail to reject the joint hypothesis of no treatment effect on all variables in the maternal decision making stepdown category.

However, the first adjusted p-value is close to the 10 percent cutoff.

In regards to the final family of variables relating to *Sleep Routines*, we find no significant treatment effect and the stepdown procedure rejects the joint null hypothesis at the first step. The direction of the effects are mixed. For example, the high treatment participants were more likely to prepare their babies for sleep in the recommended manner (this implies putting the baby to bed while they are still awake). This variable almost reaches statistical significance at the 10 percent level. On the other hand, the babies in the low treatment group were more likely to sleep in an appropriate location (alone in their own cot and not in the mother's bed) and this difference was precisely determined when a left-sided test was employed. Overall, two variables were in the hypothesised direction, while the remaining six are not.

Table 3: Comparison of High and Low Treatment Outcomes: Child Health

Outcome	N (N_{HIGH} N_{LOW})	M_{HIGH} (SD)	M_{LOW} (SD)	p-values		Effect Size (d)
				Individual Test[1]	Stepdown (Adj.)[2]	
Child Health						
‡Age (in days) left hospital	17 (8 /90)	.2 (7.0)	.16 (.72)	0.564	0.965	0.01
Birth weight (grams)	170 (80/90)	19 (589)	8 (61)	0.587	0.948	0.0
Good health since birth	17 (8 /90)	0.9 (0.26)	0.9 (0.25)	0.576	0.905	0.02
‡Stayed in hospital during first 6 months	17 (8 /90)	0.10 (0. 0)	0.09 (0.29)	0.591	0.891	0.0
‡No. of health problems taken to medical centre	17 (8 /90)	1. 7 (1.62)	1.28 (1.09)	0.669	0.827	0.07
‡Problem breathing	17 (8 /90)	0.22 (0.41)	0.14 (0. 5)	0.910	0.910	0.19
Mothers' Health Decisions for her Infant						
Baby eats appropriate food	17 (8 /90)	0.87 (0. 4)	0.77 (0.4)	0.01 **	0.129	0.26
Necessary immunizations at 4 months	172 (82/90)	0.96 (0.19)	0.88 (0.)	0.029**	0.149	0 . 2
Appropriate frequency of eating	17 (8 /90)	0.77 (0.42)	0.6 (0.48)	0.02 **	0.1 5	0. 0
‡Leave baby to cry	17 (8 /90)	0.41 (0.49)	0.46 (0.50)	0. 0	0.885	0.09
Necessary immunizations at 6 months	172 (82/90)	0. 5 (0.48)	0. 1 (0.47)	0. 7	0.877	0.09
Mother breastfed as a baby	171 (81/90)	0.15 (0. 6)	0.12 (0.)	0.4	0.846	0.08
‡Baby's crying a problem	17 (8 /90)	0.12 (0.)	0.11 (0. 2)	0.414	0.797	0.0
Attempted breastfeeding	17 (8 /90)	0.24 (0.4)	0.22 (0.42)	0.482	0.74	0.04
Knows baby's weight	17 (8 /90)	0.41 (0.49)	0.48 (0.5)	0.807	0.807	0.14

Table 3 (continued)

Outcome	N (N_{HIGH} N_{LOW})	M_{HIGH} (SD)	M_{LOW} (SD)	p-values Individual Test[1]	 Stepdown (Adj.)[2]	Effect Size (d)
Sleep Routines						
Appropriate sleep preparation	17 (8 /90)	0.48 (0.5)	0. 9 (0.49)	0.114	0.575	0.19
‡Time to sleep (>15 mins)	172 (82/90)	0.29 (0.46)	0. (0.47)	0. 55	0.878	0.09
‡Baby awakening a problem	17 (8 /90)	0.24 (0.4)	0.2 (0.4)	0.515	0.9	0.02
Sleeps more than 8 hrs per night	171 (8 /88)	0.76 (0.4)	0.78 (0.41)	0.609	0.958	0.06
Sleeps undisturbed through the night	17 (8 /90)	0.75 (0.44)	0.77 (0.4)	0.607	0.952	0.05
‡Difficulty falling asleep	17 (8 /90)	0.45 (0.5)	0. 8 (0.49)	0.841	0.994	0.14
Sleeps undisturbed by months	17 (8 /90)	0. 6 (0.48)	0.46 (0.5)	0.914	0.988	0.19
Appropriateness of sleeping location	17 (8 /90)	0.9 (0.)	0.99 (0.11)	0.998	0.998	0. 9

Notes: 'N' indicates the sample size. 'M' indicates the mean. 'SD' indicates the standard deviation.
[1] one-tailed (right-sided) p-value from an individual permutation test with 1000 replications. [2] one-tailed (right-sided) p-value from a stepdown permutation test with 1000 replications. ** Significant at the 5 percent level * Significant at the 10 percent level. ‡ Represents a negative outcome and was reverse scored for the stepdown analyses. The sample sizes reported are those used in the individual tests and may differ from the sample size used in the stepdown procedure. For the stepdown procedure, any observations that have missing data for measures within the stepdown family are dropped. The variables are reported in the order in which they are dropped from the stepdown procedure.

5.1.3 Parenting

The parenting outcomes are divided into four families and the results from the analyses are presented in Table 4. The high treatment group had more favourable scores on all *Parental Locus of Control* subdomains on average. The mean differences, however, are not statistically significant and the effects sizes are quite small. On the *Condon Maternal Attachment Scale* subdomains, the mean differences are very small and none are precisely determined.

There is strong evidence, however, of a significant treatment effect on the *Parental Stress Index* domain. The results from the stepdown procedure indicate that we can reject the join null hypothesis of no impact on all subscales at the 10 percent level. Note that the most significant single hypothesis in this domain relates to the *parent-child dysfunctional interactions* subscale. In the second stage of the stepdown procedure, this measure is discarded and the two remaining measures are tested for joint significant. The adjusted p-value

indicates that we fail to reject the joint null hypotheses for this subset of measures. Therefore, when the joint probability distribution is considered, it is the large effect on the *Parent-Child Dysfunctional Interactions* measure which is driving the small p-value in the first step.

The analyses of the *Parental Cognition and Conduct Towards the Infant* subscales using individual permutation tests finds that the high treatment group is scoring significantly higher with respect to the *Baby Comparison Score* and the *Parental Hostile-Reactive Behaviour* subdomains at the 5 percent and 10 percent level respectively. However, when we adjust these p-values to account for multiple outcomes, we fail to reject the null hypotheses of no effect on the entire category. The effects, although not significant, are in the hypothesised direction for four out of the six measures.

Table 4: Comparison of High and Low Treatment Outcomes: Parenting

Outcome	N (N_{HIGH} N_{LOW})	M_{HIGH} (SD)	M_{LOW} (SD)	p-values: Individual Test[1]	p-values: Stepdown (Adj.)[2]	Effect Size (d)
Parental Locus of Control (PLOC)						
‡Parental Control of Child's Behaviour	17 (8 /90)	6.92 (2.82)	7.22 (2.64)	0.27	0.71	0.11
‡Child Control of Parent's Life	17 (8 /90)	8.4 (. 6)	8.74 (.11)	0.281	0.705	0.10
‡Parental Responsibility	17 (8 /90)	12.57 (.18)	12.86 (.02)	0.25	0.608	0.09
‡Parental Belief in Fate	17 (8 /90)	9.7 (.65)	9.97 (. 2)	0.	0.502	0.08
‡Parental Efficacy	17 (8 /90)	6.65 (2.4)	6.76 (2.4)	0. 99	0. 99	0.04
Maternal Attachment (CMAS)						
Quality of Attachment	17 (8 /90)	4.69 (0. 0)	4.68 (0. 7)	0.459	0.772	0.0
Pleasure in Interaction	17 (8 /90)	4. (0. 8)	4. 4 (0.4)	0.578	0.805	0.02
Absence of Hostility	17 (8 /90)	4. 9 (0.5)	4.41 (0.5)	0.61	0.61	0.04
Parenting Stress Index (PSI)						
‡Parent-Child Dysfunctional Interaction	17 (8 /90)	16.94 (4.81)	18.4 (5.71)	0.041**	0.082*	0.28
‡Difficult Child	17 (8 /90)	19.45 (5.00)	20.19 (5.50)	0.174	0.277	0.14
‡Parental Distress	17 (8 /90)	26.02 (7.98)	25.71 (7.47)	0.60	0.60	0.04

Table 4 (continues)

				p-values		
Outcome	**N (N_{HIGH} N_{LOW})**	**M_{HIGH} (SD)**	**M_{LOW} (SD)**	**Individual Test[1]**	**Stepdown (Adj.)[2]**	**Effect Size (d)**
Parental Cognition and Conduct Toward the Infant Scale (PACTOIS)						
Baby Comparison Score	17 (8 /90)	7.52 (1.92)	7.04 (1.90)	0.047**	0.228	0.26
‡Parental Hostile Reactive Behaviour	17 (8 /90)	0.8 (1.1)	1.04 (1.21)	0.077*	0. 5	0.20
Parental Self-Efficacy	17 (8 /90)	8.8 (1.11)	8.67 (1.24)	0.255	0.665	0.10
Parental Impact	17 (8 /90)	7.25 (2.00)	7.07 (2.2)	0. 04	0.664	0.08
‡Parental Over-Protection	17 (8 /90)	6.18 (2.19)	6.14 (1.99)	0.5 5	0.8 5	0.02
Parental Warmth	17 (8 /90)	9.18 (1.17)	9.24 (1.27)	0.649	0.649	0.06

Notes: 'N' indicates the sample size. 'M' indicates the mean. 'SD' indicates the standard deviation.
[1] one-tailed (right-sided) p-value from an individual permutation test with 1000 replications. [2] one-tailed (right-sided) p-value from a stepdown permutation test with 1000 replications. ** Significant at the 5 percent level * Significant at the 10 percent level. ‡ Represents a negative outcome and was reverse scored for the stepdown analyses. The sample sizes reported are those used in the individual tests and may differ from the sample size used in the stepdown procedure. For the stepdown procedure, any observations that have missing data for measures within the stepdown family are dropped. The variables are reported in the order in which they are dropped from the stepdown procedure.

5.1.4 Substance Use

Finally, the results from our analyses of substance use during pregnancy are presented in Table 5. No participant in the high treatment group took an illegal drug during pregnancy, while two participants in the low treatment group did. Due to the small numbers, it is not possible to conduct a meaningful analysis of this measure. There is no evidence of a statistically significant treatment effect on participants' smoking and alcohol consumption during pregnancy. It logically follows that the stepdown procedure fails to find an overall impact of the programme on this category of outcomes.

Table 5: Comparison of High and Low Treatment Outcomes: Substance Use

Outcome	N (N_{HIGH} N_{LOW})	M_{HIGH} (SD)	M_{LOW} (SD)	p-values: Individual Test[1]	p-values: Stepdown (Adj.)[2]	Effect Size (d)
Substance Use During Pregnancy						
‡Illegal drugs during pregnancy	172 (82/90)	0 (0.00)	0.02 (0.15)	0.1 9	0.259	0.21
‡Smoked during pregnancy	172 (82/90)	0.49 (0.50)	0.51 (0.50)	0. 94	0.609	0.05
‡Drank alcohol during pregnancy	172 (82/90)	0. 4 (0.48)	0.28 (0.45)	0.847	0.847	0.14

Notes: 'N' indicates the sample size. 'M' indicates the mean. 'SD' indicates the standard deviation.
[1] one-tailed (right-sided) p-value from an individual permutation test with 1000 replications. [2] one-tailed (right-sided) p-value from a stepdown permutation test with 1000 replications. ** Significant at the 5 percent level * Significant at the 10 percent level. ‡ Represents a negative outcome and was reverse scored for the stepdown analyses. The sample sizes reported are those used in the individual tests and may differ from the sample size used in the stepdown procedure. For the stepdown procedure, any observations that have missing data for measures within the stepdown family are dropped. The variables are reported in the order in which they are dropped from the stepdown procedure.

5.2 Robustness and Sensitivity

5.2.1 Attrition Analyses

The data analysed in this study was collected for 173 (low = 90; high = 83) participants indicating that 16 percent of the original sample did not complete a six month interview (low = 11 percent; high =20 percent). It should be noted that not all of these participants have officially dropped out; some participants have disengaged with the programme and some interviews were missed during the appropriate six month window. The official drop-out rate was 9 percent (low =6 percent, high = 13 percent).

To examine the impact of attrition on six month outcomes, the participants who officially dropped out and the disengaged/non-response participants have been grouped together for comparison with the sample that completed a six month interview. It is important to determine whether attrition is happening at random or the result of a systematic process. Twenty-one key baseline measures relating to maternal characteristics were examined and permutation tests were applied to determine whether there were statistically significant differences between the participants who completed a six month interview and those who did not. The results for the high treatment group are presented in Table A1. Only one significant difference was found indicating that mothers who did not complete a six month interview (in the high treatment group) were less likely to be employed. The same analysis was conducted for the low treatment group. The results are presented in Table A2. Statistically significant differences were found on four measures. Specifically, the participants

who did not complete a six month interview in the low treatment group were younger, more likely to be first time mothers, had lower cognitive resources and had lower scores on the *Knowledge of Infant Development measure.*

Four statistically significant differences between the attrition and non-attrition sample in the low treatment group is cause for concern. To examine whether the high treatment group and low treatment group were no longer

intervirwho5432(57(re)661(and)661365:9areer)661o(irtmen)28eand

5.2.2 Robustness

For robustness, we also test individual hypotheses using traditional t tests and permutation tests using 10,000 replications. The results are quite similar regardless of the technique used. Using 10,000 replications the same number of statistically significant differences were found when carrying out individual tests. Also, using traditional t tests, although the p-values varied slightly from the permutation testing approach, the results were by and large similar with all five statistically significant differences from permutation testing being identified. The traditional t test result for drug use during pregnancy indicates that the difference is statistically significant but the percentage of respondents who answered 'yes' to this question is too small for reliable estimation.

The stepdown procedure was also conducted using 10,000 replications and again, the joint null hypothesis could only be rejected for one group of variables. Specifically, the *Parenting Stress Index* was the only family that remained statistically significant after adjusting the p-values for strong FWER control and the *Parent-Child Dysfunctional Interactions* category was, again, found to be driving the joint rejection. Finally, all outcomes were treated as one family including all 50 variables and a joint hypothesis test was applied using the stepdown method. Not surprisingly, the joint hypothesis was rejected in the first step. Westfall and Young (1993) recommend against including too many variables in one family as it reduces the statistical power. This defends our construction of outcome groups using a theoretical approach.

6 Summary and Conclusion

6.1 Early Childhood Intervention Effects Six Months After Birth

This paper investigates the effectiveness of human capital investment during the in utero stage and up until six months of age. We examine two primary outcome categories (child development and child health) and a secondary outcome relating to parental measures. The *PFL* programme is a family focused intervention that targets disadvantaged parents in order to observe a positive mediating effect on their children. We find that at six months postpartum, few significant treatment effects are identified between the low and high treatment group. The only outcome category that survives multiple hypothesis testing relates to parenting (*Parenting Stress Index*). Few studies to date have examined the impact of home visiting programmes on a comprehensive range of outcomes at, or before, six months of age. Those that have typically report results which are mostly in line with those for *PFL*.[2]

[2] The Home Vi iting Evidence of Effectivene (HomVEE) organi ation provide the main ource of information on rigorou evaluation of home vi iting program .

With respect to child development, we find that the stepdown procedure fails to reject the joint hypothesis of no effect at the first step. While two of the measures (*ASQ Gross Motor Score* and *ASQ Communication Score*) are close to reaching significance in a one-sided test with reported p-values of 0.124 and 0.154 respectively. The effect sizes are small, but close to the 0.2 cutoff for a medium effect size. The literature on child development outcomes at the 6 month stage indicates similar results. Whipple and Whyte (2010) find no effect of the *Healthy Families America* programme on *ASQ* scores at six month of age. They do, however, find significant differences between the control and treatment groups at eight months. Similarly, Anisfeld et al. (2004) find no impact of *Healthy Families America* on cognitive development at 6 months of age, and Jungmann et al. (2009) find that the German *Pro Kind* programme does not have an effect on cognitive functioning at the same age.

In addition, we do not identify a precisely determined treatment effect with respect to the non-cognitive development measures (*Difficult Temperament*, *ASQ Personal Social Score*, *ASQ Social-Emotional Score*), yet the mean differences indicate that the high treatment group is performing better on average in this regard. Other studies have found socioemotional characteristics to be somewhat more responsive at an early age. Olds et al. (2002) find that the *Nurse Family Partnership* programme was effective at reducing emotional vulnerability in response to fear stimuli, and Jungmann et al. (2009) find that *Pro Kind* reduces the symptoms of a difficult temperament at six months of age.

The stepdown procedure failed to reject the three joint null hypotheses relating to child health. One outcome of particular interest was birth weight; we find no evidence of a treatment effect for this measure. Many comparable programmes have also failed to find a significant effect on birth weight (Stabile and Graham, 2000; Koniak-Griffin et al., 2000; Kitzman et al., 1997; Jungmann et al., 2009). However, some studies have found favourable effects. Lee et al. (2009) examine a much larger sample (~500) and find that participating in *Healthy Families America* reduces the likelihood of babies having low birth weight. One concern is that the birth weight measure indicates that mothers in the high treatment group gave birth to lower birth weight babies on average, although the difference is not statistically significant. It should be noted that this is a self-reported measure and, therefore, could be susceptile to measurement error. Further analysis of the data revealed that one participant in the low treatment group reported giving birth to a baby weighing 5273 grams (much higher than the national average of 3470 grams), while one participant in the high treatment group reported giving birth to a baby weighing 1588 grams (the *World Health Organisation* classifies babies weighing less than 2500 grams as low birth weight). If these two participants are excluded, the high treatment

group had a higher average birth weight. The difference, however, is not statistically significant. It should be noted that participants joined *PFL* when they were 22 weeks pregnant on average (half way into a normal term preganancy). Therefore, the treatment may have started too late in pregnancy to significantly impact upon factors that affect birthweight such as maternal nutrition, alcohol consumption and smoking behaviour. This is discussed further below.

Koniak-Griffin et al. (2000) report that the Early Intervention Programme was effective at reducing the duration of birth related hospitalisation. This programme, however, targeted teenage women aged 14-19, which represent a higher risk group than the participants in *PFL*. This may explain why we observe no significant differences between the high and low treatment group with respect to the number of days the baby was kept in hospital after birth.

The results from our analysis indicate that the stepdown p-values for one category, *mother's health decision for infant*, were close to the 10 percent cutoff for the first three joint hypothesis tests. This reflects the three significant findings that resulted from individual permutation tests in this category. Specifically, high treatment participants were statistically more likely to feed their baby appropriate foods, at the appropriate frequency, and high treatment children were more likely to have received the appropriate immunisations. This indicates a tendency towards some significant effects that have been reported in the literature. Guyer et al. (2003) find that the *American Healthy Steps* programme significantly increased the likelihood that children had received appropriate immunisations.

With respect to parenting, we find that one family of variables, measuring parental stress, survives joint hypothesis adjustment. Stepping down through the variables in this family, we find that the *Parent-Child Dysfunction Interactions* subdomain is driving the rejection. This measure relates to the respondent's perception that their child is a negative element of their life. Specifically, whether the parent feels rejected/abused by their child, or feels that the child does not meet their expectations. The *PFL* programme is successful at reducing these symptoms in participants. For the three other outcome families relating to parental locus of control, maternal attachment and perceptions about the parental role, there is no evidence of an overall treatment effect. To the best of our knowledge, we examine parenting instruments which have not been studied at the six month stage in the home visiting literature before. Using other parenting instruments, the literature reveals no observable effect of home visitation at this early stage (Anisfeld et al., 2004; Johnston et al., 2004; LeCroy and Krysik, 2011). Similarly, with respect to caregivers' smoking behaviour and alcohol consumption during pregnancy, the results from this study indicate no significant treatment effects. This

is in line with Johnston et al. (2004) who indicate that the *Healthy Steps* programme has no effect on a participating household's smoking or alcohol consumption at three months postpartum. Similarly, Jungmann et al. (2009) measured smoking at the 36th week of pregnancy and found that the *Pro Kind* programme did not exhibit an effect. LeCroy and Krysik (2011) also find no impact of the *Healthy Families America* programme on alcohol consumption at the six month stage.

The literature on home visiting programmes indicates that during the first six months of a baby's life, the effects of early childhood interventions are not easily observed. In this regard, *PFL* is no different. Many of these programmes do show significant favourable effects at later ages. *PFL* has been designed to target parents in order to have a positive dissemination effect on their children. Consider that the average *PFL* participant began engaging with the programme half way into their pregnancy (22 weeks) and had received 14 home visits on average at this point. It is likely that this small window did not allow much time for the participants to adopt the strategies advised by their mentors as the bond between mentor and participant was still being formed. Barlow et al. (2005) highlight that weak effects at the start of an intervention, as a result of low-intensity delivery, can still serve as the prelude to more engagement as trust levels increase. In fact, the majority of studies that calculate high returns to early childhood investment are based on analyses conducted when the participating children have reached adulthood (Olds et al., 1997; Heckman et al., 2010). As discussed in Section 2, the theory on human skill formation points to a skill multiplier effect (Cunha and Heckman, 2007). The majority of measures we examine indicate a treatment effect which is in a favourable direction but not precisely determined. These small differences could potentially result in a large return over time. However, it is important that later waves of outcome data are examined in the same manner to understand the true effectiveness, or indeed ineffectiveness, of the programme. The next wave of data collection, measuring child and parent outcomes at 12 months of age, will be complete in April 2012.

6.2 Benefits of our Methodology

A major advantage of our study is that we use new testing methods that are particularly suited to experimental evaluations which involve multiple outcomes and small samples. An analysis of this data using tradition statistical methods such as t tests could be criticised for a lack of statistical power. Indeed, no other study examining six month outcomes has used this methodology. Some studies have the advantage of larger samples (Olds et al., 2002; Kitzman et al., 1997; Lee et al., 2009) and others acknowledge the issue of small samples yet do not adapt their statistical approach (LeCroy and Krysik, 2011; Jungmann et al., 2009). All of the studies which examine child outcomes at six months of age are confined to the disciplines of psychology

and public health. Moving the topic into the economic sphere has facilitated the application of advanced quantitative models. Heckman et al. (2010) and Anderson (2008) employ permutation testing methods and account for the multiplicity effect when examining American programmes that began in 1960s and 1970s. This study is the first to adopt this strategy when testing a modern programme that examines early child outcomes. It is also unique in that it examines a home visiting intervention in a European setting where the governance structure and cultural setting could result in divergent effects.

It is important to consider the multiplicity effect. A naive evaluation strategy would be to examine each outcome measure individually and subsequently calculate the proportion of measures for which a significant difference has been found. In our analyses, this would indicate a significant effect for 12 percent of measures (6/50) and could therefore be deemed as an overall significant effect. A less rigorous approach would be to focus on the significant findings and would equate to data mining. Our hypotheses have been formed by theory and the p-values have been adjusted to account for the increased likelihood of a Type I error in a multiple hypotheses setting. We find that only one category of outcomes, the *Parenting Stress Index*, rejects the joint null hypothesis of no treatment effect. Looking at this category in more detail, we can determine that that one single subdomain measure (*Parent-Child Dysfunctional Interactions*) is driving the rejection. This more rigorous method paints quite a different picture to the naive approach of examining all outcomes separately. We recommend that other researchers examining small scale experimental pilot studies adopt the same approach.

6.3 Future Work

In the *PFL* programme all of the participants, both high and low treatment, received some provisions. Therefore, our analysis does not follow the standard RCT format of a treatment and pure control group comparison. It is possible that both groups are benefiting from their common provisions and our analyses could be failing to identify this lower treatment effect. An additional control group was recruited from a separate community and baseline and six month interviews were administered to these participants. This group did not receive any treatment yet did receive a weekly shopping voucher as an incentive for completing each survey. Ideally, we would like to compare this control group with the entire *PFL* cohort in order to extract the combined impact of the low and high treatment and to compare the effects of each dosage level. However, an analysis of the baseline data reveals that this group differed significantly from the low and high treatment group on 25 percent of measures analysed. The differences were mainly found with respect to socioeconomic measures, indicating that the comparison community represents a relatively higher SES group.

In order to compare these participants to the low and high treatment groups, it is necessary to condition on certain variables. This is a challenging task in small samples using permutation testing methods and is outside the scope of this paper. In forthcoming work, we will explore methods for dealing with non-random treatment assignment in order to generate counterfactuals for all *PFL* participants.

In this paper we examine average treatment effects. In future studies, we will examine whether the treatment effect depends on treatment intensity. Participants in the high treatment group vary with respect to the number of home visits they receive and the total duration of these visits. By identifying the individual characteristics which are associated with higher levels of engagement it will be possible to match participants with appropriate control group counterparts. This will allow us to determine the optimal level of programme intensity and will inform public policy on how to improve life outcomes for disadvantaged children.

Appendix

Table A1: Comparison of Attritors and Non-Attritors: High Treatment Group

	Attrition			Non-Attrtion			Individual
	N	**M**	**(SD)**	**N**	**M**	**(SD)**	**p-value**
Week in pregnancy at programme entry	22	20.86	8.04	82	21.78	7.83	0.635
Mother' age	22	24.68	6.23	82	25.67	5.76	0.501
Partnered	22	0.68	0.48	82	0.80	0.40	0.402
Married	22	0.09	0.29	82	0.16	0.37	0.528
Living with parent( )	22	0.64	0.49	82	0.55	0.50	0.470
Fir t time mother	22	0.59	0.50	82	0.52	0.50	0.639
Low education	22	0.50	0.51	82	0.29	0.46	0.139
Mother employed	22	0.14	0.35	82	0.43	0.50	0.007**
Save regularly	22	0.36	0.49	82	0.50	0.50	0.330
Social hou ing	21	0.62	0.50	82	0.54	0.50	0.595
‡Cognitive Re ource (WASI)	9	80.67	8.49	81	82.49	13.01	0.596
Phy ical Health Condition	22	0.73	0.46	82	0.76	0.43	1.000
Mental Health Condition	22	0.32	0.48	82	0.27	0.45	0.771
Smoking during pregnancy	22	0.50	0.51	82	0.51	0.50	1.000
Drinking during pregnancy	22	0.18	0.39	82	0.27	0.45	0.428
Drug ever u ed	22	0.05	0.21	82	0.16	0.37	0.171
Vulnerable attachment (VASQ)	22	18.95	3.77	82	18.05	3.76	0.332
Po itive parenting attitude (AAPI)	22	117.91	16.82	82	120.13	12.92	0.541
Self efficacy (Pearlin)	22	2.79	0.60	82	2.93	0.49	0.344
Self e teem (Ro enberg)	22	11.91	2.89	82	13.06	2.60	0.105
Knowledge of infant development (KIDI)	22	71.69	9.39	82	72.40	7.10	0.757

‡ Mea ured at 3 month po tpartum. ** Significant at the 5 percent level. p-value were obtained u ing permutation ba ed, two- ided t-te t with 1,000 replication .

Table A2: Comparison of Attritors and Non-Attritors: Low Treatment Group

	Attrition			Non-Attrtion			Individual
	N	M	(SD)	N	M	(SD)	p-value
Week in pregnancy at programme entry	12	22.50	7.70	89	21.18	6.87	0.579
Mother' age	12	22.42	4.94	89	25.69	6.04	0.053*
Partnered	12	0.92	0.29	89	0.83	0.38	0.500
Married	12	0.25	0.45	89	0.17	0.38	0.685
Living with parent( )	12	0.58	0.51	89	0.45	0.50	0.551
Fir t time mother	12	0.75	0.45	89	0.46	0.50	0.078*
Low education	12	0.67	0.49	89	0.36	0.48	0.127
Mother employed	12	0.33	0.49	89	0.40	0.49	0.763
Save regularly	12	0.42	0.51	89	0.53	0.50	0.559
Social hou ing	12	0.50	0.52	89	0.56	0.50	0.764
‡Cognitive re ource (WASI)	3	72.67	3.79	88	80.66	13.20	0.090*
Phy ical Health Condition	12	0.50	0.52	89	0.64	0.48	0.521
Mental Health Condition	12	0.08	0.29	89	0.26	0.44	0.201
Smoking during pregnancy	12	0.33	0.49	89	0.49	0.50	0.357
Drinking during pregnancy	12	0.42	0.51	89	0.25	0.43	0.306
Drug ever u ed	12	0.08	0.29	89	0.16	0.37	0.69
Vulnerable attachment (VASQ)	12	17.33	3.63	89	17.89	4.04	0.664
Po itive parenting attitude (AAPI)	11	115.36	15.91	89	117.02	13.55	0.741
Self efficacy (Pearlin)	12	2.96	0.38	89	3.03	0.53	0.595
Self e teem (Ro enberg)	12	13.00	2.13	89	12.75	2.95	0.736
Knowledge of infant development (KIDI)	12	64.64	5.21	89	70.51	8.29	0.004**

‡ Mea ured at 3 month po tpartum. ** Significant at the 5 percent level. p-value were obtained u ing permutation ba ed, two- ided t-te t with 1,000 replication .

References

Abidin, R.R., *Parenting Stress Index (PSI)* Psychological Assessment Resources, Inc. 1995.

Anderson, M. L., "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 2008, *103* (484), 1481–1495.

Anisfeld, E., J. Sandy, and N.B. Guterman, "Best Beginnings: A Randomized Controlled Trial of Paraprofessional Home Visiting Program," Technical Report, Columbia University School of Social Work, New York 2004. Report to the Smith Richardson Foundation and New York State Office of Children and Family Services.

Aos, S., R. Lieb, J. Mayfield, M. Miller, and A. Pennucci, *Benefits and Costs of Prevention and Early Intervention Programs for Youth*, Olympia: Washington State Institute for Public Policy, 2004.

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

Bates, J.E., C.B. Freeland, and M.L. Lounsbury, "Measurement of Infant Difficulties," *Child Development*, 1979.

Bavolek, S. J. and R.G. Keene, *Adult-adolescent parenting inventory – AAPI–2: administration and development handbook* Family Development Resources, Incr 1999.

Becker, G.S., *Human Capital*, New York: National Bureau of Economic Research, 1964.

_ , *Human Capital*, 3rd ed., Chicago: The University of Chicago Press, 1993.

_ **and K.M. Murphy**, "The Family and the State," *Journal of Law and Economics*, 1998, *31* (1), 1–18.

_ **and N. Tomes**, "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility," *Journal of Political Economy*, 1979, *87* (6), 1153–1189.

Benham, L., "Benefits of Women's Education within Marriage.," *Journal of Political Economy*, 1974, *82* (2), S57–S71.

Bifulco, A, J. Mahon, J.-H. Kwon, P. M. Moran, and C. Jacobs, "The Vulnerable Attachment Style Questionnaire (VASQ): an interview-based measure of attachment styles that predict depressive disorder," *Psychological Medicine*, 2003, *33* (06), 1099–1110.

Black, S.E., P.J. Devereux, and K.G. Salvanes, "Why the Apple Doesn't Fall Far: Understanding Intergenerational Transmission of Human Capital," Working Paper 10066, National Bureau of Economic Research 2003.

_ **,** _ **, and** _ , "From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes," *Quarterly Journal of Economics*, 2007, *122* (1), 409–439.

Boivin, M., D. Pérusse, G. Dionne, V. Saysset, M. Zoccolillo, G.M. Tarabulsy, N. Tremblay, and R.E. Tremblay, "The genetic-environmental etiology of parents' perceptions and self-assessed behaviours toward their 5-month-old infants in a large twin and singleton sample," *Journal of Child Psychology and Psychiatry*, 2005, *46* (6), 612–630.

Bradley, J.V., *Distribution-Free Statistical Tests*, Englewood Cliffs, N.J.: Prentice Hall, Inc., 1968.

Brooks-Gunn, J., L. Berlin, and A. Fuligni, "Early Childhood Intervention Programs: What About the Family?," in J.P. Shonkoff and S.J. Meisels, eds., *Handbook of Early Childhood Intervention*, 3rd ed., New York: Cambridge University Press, 2000.

Campis, L.K., R.D. Lyman, and S Prentics-Dunn, "The Parental Locus of Control Scale: Development and validation," *Journal of Clinical Child Psychology*, 1986, *15*, 260–267.

Caucutt, E.M. and L.J. Lochner, "Early and Late Human Capital Investments, Borrowing Constraints, and the Family," 2011. Working paper. Retrieved from http://economics.uwo.ca/faculty/caucutt/papers/earlyandlate.pdf.

Coleman, J.S., "Social Capital in the Creation of Human Capital," *American Journal of Sociology*, 1988, *94*, S95–S120.

Condon, J.T. and C.J. Corkindale, "The assessment of parent-to-infant attachment: development of a self-report questionnaire instrument," *Journal of Reproductive and Infant Psychology*, 1998, *16*, 57–77.

Coneus, K. and F. Pfeiffer, "Self-Productivity in Early Childhood," 2007. ZEW - Centre for European Economic Research Discussion Paper No. 07-053; SOEPpaper, No. 39. Retrieved from SSRN: http://ssrn.com/abstract=1010302 or http://dx.doi.org/10.2139/ssrn.1010302.

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

_ , J.J. Heckman, and S.M. Schennach, "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 2010, *78* (3), 883–931.

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

_ , "Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development," *Journal of Economic Literature*, 2009, *47* (1), 87–122.

_ and E. Moretti, "Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings," *The Quarterly Journal of Economics*, 2003, *118* (4), 1495–1532.

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

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

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

Fisher, R.A., *The Design of Experiments*, Edinburgh: Oliver Boyd, 1935.

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

Green, B.F., "A Practical Interactive Program for Randomization Tests of Location," *The American Statistician*, 1977, *31* (1), 37–39.

Grossman, M., *The Demand for Health-A Theoretical and Empirical Investigation* 1972.

Guyer, B., M. Barth, D. Bishai, M. Caughy, B. Clark, C. Burkom, and C. Tang, "The Healthy Steps for Young Children Program National Evaluation," Technical Report, Johns Hopkins Bloomberg School of Public Health, Baltimore 2003.

Hanushek, E.A. and L. Woessmann, "The Role of Cognitive Skills in Economic Development," *Journal of Economic Literature*, 2008, *46* (3), 607–668.

Heckman, J.J. and D.V. Masterov, "The Productivity Argument for Investing in Young Children," *Applied Economic Perspectives and Policy*, 2007, *29* (3), 446–493.

_ , **S.H. Moon, R. Pinto, P.A. Savelyev, and A. Yavitz**, "Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program," *Quantitative Economics*, 2010, *1*, 1–46.

Holm, S., "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 1979, *6* (2), 65–70.

Jaffa, T., S. Scott, J.H. Hendriks, and C.M Shapiro, "Sleep Disorders in Children," *British Medical Journal*, 1993, *306*, 640–643.

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

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

Jungmann, T., Y. Ziert, V. Kurtz, and T. Brand, "Preventing adverse developmental outcomes and early onset conduct problems through prenatal and infancy home visitation: The German pilot project 'Pro Kind'," *European Journal of Developmental Science*, 2009, *3* (3), 292–298.

Kahn, J. and K.A Moore, "What works for home visiting programs: lessons from experimental evaluations of programs and interventions," 2010. Available at: `www.childtrends.org/Files/Child_Trends-2010_7_1_FS_WWHomeVisitpdf.pdf`. Accessed March 15, 2011.

Karoly, L. A., M. R. Kilburn, and J. S. Cannon, *Early Childhood Interventions: Proven Results, Future Promise, MG-341*, Santa Monica, CA: RAND Corporation, 2005.

Kennedy, P.E., "Randomization Tests in Econometrics," *Journal of Business & Economic Statistics*, 1995, *13* (1), 85–94.

Kitzman, H., D.L. Olds, C.R. Henderson, C. Hanks, R. Cole, R. Tatelbaum, K.M. McConnochie, K. Sidora, D.W. Luckey, D. Shaver, K. Engelhardt, D. James, and K. Barnard, "Effect of Prenatal and Infancy Home Visitation by Nurses on Pregnancy Outcomes, Childhood Injuries, and Repeated Childbearing," *JAMA: The Journal of the American Medical Association*, 1997, *278* (8), 644–652.

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

LeCroy, C.W. and J. Krysik, "Randomized trial of the healthy families Arizona home visiting program," *Children and Youth Services Review*, 2011, *33* (10), 1761–1766.

Lee, E., S.D. Mitchell-Herzfeld, A.A. Lowenfels, R. Greene, V. Dorabawila, and K.A. DuMont, "Reducing Low Birth Weight Through Home Visitation: A Randomized Controlled Trial," *American Journal of Preventive Medicine*, 2009, *36* (2), 154–160.

Mincer, J., "Investment in Human Capital and Personal Income Distribution," *Journal of Political Economy*, 1958, *66* (4), 281–302.

Moir, R., "A Monte Carlo Analysis of the Fisher Randomization Technique: Reviving Randomization for Experimental Economists," *Experimental Economics*, 1998, *1*, 87–100.

Olds, D.L., J. Eckenrode, C.R Henderson Jr, H. Kitzman, J. Powers, R. Cole, K. Sidora, P. Morris, L.M. Pettitt, and D. Luckey, "Long-term Effects of Home Visitation on Maternal Life Course and Child Abuse and Neglect," *Journal of the American Medical Association*, 1997, *278* (8), 637–643.

_ , J. Robinson, R. O'Brien, D.W. Luckey, L.M. Pettitt, C.R. Henderson, R.K. Ng, K.L. Sheff, J. Korfmacher, S. Hiatt, and A. Talmi, "Home Visiting by Paraprofessionals and by Nurses: A Randomized, Controlled Trial," *Pediatrics*, 2002, *110* (3), 486–496.

Pfeiffer, F. and K. Reuß, "Age-dependent skill formation and returns to education," *Labour Economics*, 2008, *15* (4), 631–646.

Plug, E. and W. Vijverberg, "Schooling, Family Background, and Adoption: Is It Nature or Is It Nurture?.," *Journal of Political Economy*, 2003, *111* (3), 611–641.

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

Reynolds, Arthur J., Judy A. Temple, Dylan L. Robertson, and Emily A. Mann, "Age 21 Cost-Benefit Analysis of the Title I Chicago Child-Parent Centers," *Educational Evaluation and Policy Analysis*, 2002, *24* (4), 267–303.

Roberts, H., "Children, Inequalities, and Health," *British Medical Journal*, 1997, *314*, 1122–1125.

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

Schultz, Theodore W., "Investment in Human Capital," *The American Economic Review*, 1961, *51* (1), 1–17.

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

Siegel, S., "Nonparametric Statistics," *The American Statistician*, 1957, *11* (3), 13–19.

Smith, James P, "The Impact of Childhood Health on Adult Labor Market Outcomes," *Review of Economics and Statistics*, 2009, *91* (3), 478–489.

Squires, J., D. Bricker, and E. Twombly, *The ASQ:SE user's guide* Brookes Publishing 2003.

Squires, J.K., L. Potter, and D. Bricker, *The Ages and Stages Questionnaire users guide* Brookes Publishing 1999.

Stabile, I. and M. Graham, "Florida Panhandle Healthy Start: A randomized trial of prenatal home visitation," Technical Report, FSU Center for Prevention and Early Intervention Policy, Tallahassee, FL 2000. Report to the Smith Richardson Foundation and New York State Office of Children and Family Services.

Upton, G.J.G., "Fisher's Exact Test," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 1992, *155* (3), 395–402.

Victora, C.G., A. Wagstaff, J.A. Schellenberg, D. Gwatkin, M. Claeson, and J.-P. Habicht, "Applying an equity lens to child health and mortality: more of the same is not enough," *The Lancet*, 2003, *362* (9379), 233–241.

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

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

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

Whipple, E.E. and J.D. Whyte, "Evaluation of a Healthy Families America (HFA) Programme: A Deeper Understanding," *British Journal of Social Work*, 2010, *40* (2), 407–425.