Can early intervention policies improve well-being? Evidence from a randomized controlled trial[♦]

Orla Doyle[a,b*], Liam Delaney[c,b], Christine O'Farrelly[b], Nick Fitzpatrick[b], Michael Daly[c]

[a] UCD School of Economics, University College Dublin, Belfield, Dublin 4, Ireland.
[b] UCD Geary Institute for Public Policy, University College Dublin, Belfield, Dublin 4, Ireland.
[c] Behavioural Science Centre, Stirling Management School, Stirling University, FK94LA, United Kingdom.

*Corresponding Author:
Orla Doyle,
UCD Geary Institute for Public Policy, University College Dublin, Dublin 4, Ireland.
E-mail: Orla.Doyle@ucd.ie; Phone: 00353 7164637

ABSTRACT
This study uses an experimental design to estimate the effect of a targeted policy intervention on global and experienced measures of maternal well-being. Participants are randomly assigned during pregnancy to an intensive parenting program or a control group. Well-being is assessed after approximately four years of program exposure. Global well-being is captured using measures of life satisfaction and parenting stress. Experienced well-being is captured using episodic reports of affect derived from the Day Reconstruction Method and a measure of mood yesterday. The intervention has no impact on global or negative measures, yet some individual treatment effects are observed on experienced measures of positive affect and mood yesterday, particularly during times spent without the target child. This may reflect a greater value being placed on non-parenting time, given the additional investment in parenting promoted by the intervention. These results suggest that early policy interventions may produce meaningful improvements in experienced well-being.

Keywords: Well-Being, Randomised Controlled Trial, Early Intervention.
JEL Classification: C12, C93, I01, I39, J13

[♦] This study was funded by the Irish Research Council through the Government of Ireland Collaborative Project Scheme. We would like to thank the Northside Partnership who fund the main evaluation of the Preparing for Life program through the Department of Children and Youth Affairs and The Atlantic Philanthropies. Funding support was also made available through a European Research Council (ERC) for the Advanced Investigator Award to James J. Heckman. We would like to thank all those who supported this research including the participating families, the PFL intervention staff, Judy Lovett as project coordinator, Catherine O'Melia for her assistance with data collection, and the UCD Geary Institute Early Childhood Research Team. The UCD Human Research Ethics Committee, the Rotunda Hospital Ethics Committee and the National Maternity Hospital Ethics Committee granted ethical approval for this study. Helpful comments from participants at "Measurement and Determinants of Well-being" workshop at the University of Stirling and the "Society for Research in Child Development, Developmental Methodology Meeting, San Diego" and UCD School of Economics seminar participants are gratefully acknowledged.

1. Introduction

Understanding the impact of targeted early intervention policies on the life-long development of children is an increasingly important focus of modern policymakers. One potential externality of such interventions is welfare improvements for parents, particularly for policies that target parenting and coping skills. Such benefits may yield value both directly, through their immediate impact on parental utility, and indirectly, through impacting improved child health and development. Understanding how to quantify these benefits is essential for providing a full account of the costs and benefits of early intervention policies.

The identification of the utility effects of many public policies is frequently hampered by non-experimental designs which cannot infer causality. Randomized controlled trials are widely considered the most robust means of determining impact (Craig et al., 2008), yet few experimental policy evaluations have incorporated comprehensive measures of utility into estimates of treatment effects. Global well-being measures are increasingly used as direct measure of utility and are based on retrospective assessments of evaluative (e.g. life satisfaction) and hedonic (e.g. happiness) well-being. Such global measures are often elicited as single-item questions asking respondents to rate their well-being generally or over several weeks. More recently, a set of papers have argued for a more disaggregated approach which measures experienced utility at the level of the day or even in real-time (e.g. Dolan and Kahneman, 2008; Kahneman et al., 2004). To date, few studies have utilized these utility flow measures to evaluate policies including targeted intervention programs.

In this paper, we report findings from a study designed to evaluate the well-being effects of an early intervention program on a sample of mothers in a disadvantaged area in Ireland. Our paper adds to the literature by exploiting a randomized controlled trial in which participants are assigned to either an intensive five-year parenting program or a control group that receives low level supports common to both groups. This study is the first to examine the impact of a policy intervention on measures of experienced and global well-being using an experimental design. This distinction between experienced and global well-being has been described by Kahneman as reflecting the difference between "living life" and "thinking about life" (Kahneman and Riis, 2005). In this study, global well-being is captured using measures of life satisfaction and a standardised measure of parenting stress. Experienced well-being is captured using daily reports of average, positive, and negative affect derived from the Day Reconstruction Method (DRM) and a measure of mood yesterday.

The DRM also allows us to measure well-being during times spent with and without the target child. This is particularly relevant given the ambiguity of the effect of children on parental well-being, an issue that is complicated by selection into parenthood (see Deaton and Stone, 2013; Deaton and Stone, 2014). Thus, the ability to measure well-being at multiple points of the day may help to improve understanding about the causal relationship between children and well-being. Time use data also allows us to determine whether any identified treatment effects are driven by differences in the daily activities of the participants. Utilizing the methodology of Heckman et al. (2010), we employ permutation testing to address issues relating to the small sample size and, as a robustness test, we apply a stepdown procedure to mitigate the likelihood of accepting a false positive due to multiple hypothesis testing.

Our results identified an individual treatment effect on experienced positive affect across episodes of the study day, yet only for time spent without the target child. The treatment group have similar levels of positive affect during episodes with and without their target child, while the control group experience a fall in positive affect during episodes without their child. Similarly, we find an individual and stepdown treatment effect on an experienced measure of mood yesterday, yet not for time spent with child(ren). Consistent with the early intervention literature, the program has no impact on negative aspects of well-being, including both experienced negative affect and a global standardized measure of parenting stress. In addition, while higher proportions of the treatment group report being satisfied with their lives compared to the control group, these differences did not reach significance. We also identify no differences in time use across the study day concerning the amount of time or types of activities mothers engage in with the target child.

The paper is structured as follows. Section 2 outlines the conceptual issues involved in measuring well-being and their relevance for the evaluation of early intervention programs. In Section 3 we provide details of the early intervention under investigation and the well-being measures employed. Section 4 outlines our empirical model and statistical methods. Section 5 presents the results, and Section 6 concludes.

2. Background and literature

2.1 Well-Being and evaluation of public policy

The use of well-being measures in public policy has been widely debated in recent years (OECD, 2013). One driver of this debate is concern that purely financial measures of utility, such as employment and consumption, do not adequately capture utility, particularly in the presence of various types of bounded rationality (e.g. hyperbolic discounting, loss aversion) and externalities (e.g. Beshears et al., 2008). Scholars from a wide range of disciplines have called for global well-being measures to be directly incorporated into the development of national progress indicators (e.g. Diener and Seligman, 2004; Forgeard et al., 2011; Stiglitz et al., 2009).

There has also been a growing interest in using well-being measures to evaluate public goods and the effects of specific policies (Dolan et al., 2011; Frey and Stutzer, 2002; Gruber and Mullainathan, 2005; Luechinger, 2009). One issue with this approach is the identification of causal effects, and in particular, the specific impact of the public good being tested. For example, individuals may sort into regions that provide higher levels of the public good or may be driven to choose higher levels of the good based on unobservable characteristics correlated with either well-being or the determinants of well-being. One approach is to develop instrumental variables estimates or exploit fine-grained exogenous variation in the provision of the good (e.g. Levinson, 2012). However, these methods may not be possible for all public goods and require restrictive assumptions. Thus, for public policies with unknown effects, it has become increasingly common to pilot test provision of the good using random assignment (Duflo et al., 2008).

2.2 Maternal welfare and early intervention programs

Regarding policies which specifically focus on boosting children's skills, recent studies using random assignment have examined the potential for targeted early intervention programs to have long-lasting effects on the emotional, social, health, and economic development of children (Campbell et al., 2014; Heckman et al., 2010; Gertler et al., 2014). Less work, however, has examined the effect of these interventions on the welfare of parents. While early intervention programs may have an impact on the economic well-being of parents, such effects are complex. For example, the impact of a program on employment and consumption may be ambiguous if substitution effects occur which result in a change in priorities due to the intervention. A program may lead to reduced parental employment due to a conscious decision by parents to spend more time with their children. Consequently, measuring a parent's welfare directly may prove more informative regarding the utility effects of early intervention.

Home visiting programs (HVPs), which are a common form of early intervention that work directly with parents, may particularly have an impact on parental utility. The prevailing pattern, based on meta-analytic findings, suggests that the effects of HVPs are concentrated on parenting behaviors, attitudes, and skills (Filene et al., 2013; Sweet and Appelbaum, 2004). There is also evidence, albeit less consistent, for improvements in parental life course outcomes (e.g. employment self-sufficiency, and reliance on public assistance, Filene et al., 2013; Sweet and Appelbaum, 2004).

Less is known about the impact of HVPs on psychological well-being, and the direction of this effect is ambiguous. On the one hand, HVPs may improve well-being if the supports delivered by the home visitor foster a therapeutic alliance which acts as a pathway for promoting utility (see Ammerman et al., 2010). Alternatively, drawing on the family investment theory (Becker, 1991), HVPs may have deleterious effects on well-being if the intervention promotes substantial parental investment in the child. This may come at a cost of increased parental time, effort, and emotional outlays in the short-run, with the expectation that such investments would increase parental utility in the long run.

Research examining the relationship between early intervention and psychological well-being has focused predominantly on global negative measures. A substantial literature has illustrated the harmful effects of stress and depression on parent functioning and the subsequent consequences for child well-being (e.g., Crnic and Low, 2002; Murray et al., 1996). Depression, in particular, affects a considerable proportion of mothers enrolled in HVPs due to elevated risk conferred by their disadvantaged status. Ammerman and colleagues' (2010) systematic review found that HVPs are not sufficiently powerful, in and of themselves, to substantially mitigate depression, as measured by standardized self-report instruments. Equally, HVPs tend not to be effective in reducing parent-reported levels of stress (Sweet and Appelbaum, 2004).

Comparatively fewer studies have examined the impact of HVPs on positive aspects of well-being such as self-efficacy and self-esteem. Theories of self-efficacy, which link people's beliefs about their capabilities to their subsequent motivation, behavior, and well-being (Bandura, 1977), are central to many HVPs. Parents' perceptions of their self-efficacy may influence their choices and the degree to which they invest in their own health and the development and care of their children (Olds, 2006). Studies that have examined positive aspects of well-being are inconclusive, and have yet to be subject to systematic review. While some HVPs have demonstrated positive treatment effects in this domain (e.g. Kitzman et al., 1997), no effects are observed in others (e.g., Mitchell-Herzfeld et al., 2005). Collectively, this evidence suggests that it may be easier for HVPs to alter parenting behaviors than emotional states (Brooks-Gunn and Markman, 2005).

2.3 Global versus experienced measures of well-being

A critical issue for evaluations of public policies, including targeted early intervention programs, is the question of how well-being should be measured. The possibility that experienced measures of well-being may have different determinants to global measures has been addressed in a number of studies. Knabe et al. (2010) have argued that the negative effects of unemployment may depend on whether self-reported life satisfaction measures or diurnal measures are used. Kahneman and Deaton (2010) also find that estimates of the well-being effect of income differ substantially by whether income is measured generally or as a feeling about the previous day.

A large body of literature has emerged on the use of global retrospective measures of well-being, such as evaluations of life or domain satisfaction and accounts of happiness. These measures have the strong advantage of providing information regarding the person's appraisal of their circumstances and their feelings about them; however considerable debate exists regarding their consistency. Kahneman and others have documented how immediate mood and context can bias retrospective evaluations, and have argued that the act of thinking about such quantities may focus individuals on aspects of their life that are not crucial to their actual well-being (Kahneman and Krueger, 2006). Furthermore, retrospective happiness accounts or remembered utility tend not to accurately represent experience, as such accounts are overly influenced by intense or recent experiences and the duration of experiences is typically neglected (Kahneman et al., 2004). Finally, alongside systematic recall biases, people may simply fail to accurately recall their well-being over extended periods of several days or weeks, introducing greater error into well-being estimates.

Dolan and Kahneman argue that experienced utility is a more reliable measure of an individual’s well-being, in that it directly captures emotional experiences in real time as opposed to being filtered through cognitive biases associated with evaluating and remembering one’s overall state (Dolan and Kahneman, 2008). The experience sampling approach captures flows of utility by collecting information on individuals’ self-reported emotional responses to their daily experiences in real time at specific points during a day using electronic devices as prompts (Stone and Shiffman, 1994). It has been widely applied in clinical psychology and psychiatry studies (e.g. Bowen et al., 2013; Bylsma et al., 2011; Henquet et al., 2010; Palmier Claus et al., 2012; Peeters et al., 2006; Thompson et al., 2012).

Kahneman et al. (2004) proposed the use of the DRM as an alternative means of recording diurnal fluctuations in experienced well-being in a less burdensome manner than the experienced sampling approach. The DRM is completed in a single session during which respondents divide the previous day into discrete activities or episodes which are then rated across several positive and negative emotional/affective states. Compared with experience sampling, the DRM has the advantage of eliciting events over an entire day without interfering with the day's activities or placing administrative or respondent burden associated with carrying equipment to record events. The DRM has been used in a variety of settings, including measuring time use and emotional well-being among the unemployed (Knabe et al., 2010; Krueger and Mueller, 2012), examining individuals with optimal mental health (Catalino and Fredrickson, 2011), and studying women during the transition to motherhood (Hoffenaar et al., 2010).

Another important distinction when measuring well-being using the DRM, concerns positive and negative affect. Positive affect includes feelings of happiness, calm, focus, and control, whereas negative affect includes feelings of stress, anxiety, anger, and impatience. Positive and negative affect have been shown to represent different dimensions of well-being with distinct correlates. For example, negative affect is traditionally associated with health issues, whereas positive affect is associated with social engagement (see Crawford and Henry 2004; Tellegen et al 1999; Watson, Clark and Tellegen, 1988). An advantage of the DRM is its ability to elicit ratings of a series of episodes on dimensions of both positive and negative affect.

One potential concern when using the DRM is that respondents may not accurately recall emotions experienced the previous day. Several studies have examined this issue by comparing DRM ratings with ratings provided in real time using experienced sampling methods, and all find a reasonably high degree of convergence (Bylsma et al., 2011; Dockray et al., 2010; Kahneman et al., 2004; Kim et al., 2013; Miret et al., 2012)[1]. Furthermore, Daly et al., (2010) find a positive correlation between DRM measures of negative affect and fluctuations in heart rate, an objective indicator of psychological stress. Thus, there is a substantial degree of concordance among different studies demonstrating that the DRM provides a reliable means of measuring flows of emotional states (see Diener and Tay 2014 for a review of DRM research).

Although the DRM is arguably less burdensome than experienced sampling, it nonetheless requires considerable participant effort (Atz, 2013). Consequently, interest has developed in less intensive measures of experienced well-being that are still robust to cognitive biases which affect global measures. One proposed approach is a measure of mood

[1] For example, Dockray et al. (2010) observes between-persons correlations between experience sampling and DRM measures ranging from 0.58 to 0.90.

yesterday. This requires respondents to provide an overall appraisal of a given emotional state across the course of the previous day, and thus may be a more practical alternative than the DRM. Although these measures have recently been incorporated in some large scale social surveys, such as those conducted by the Gallup Organization and the UK Office of National Statistics, evidence is still needed to endorse their value as a viable proxy for more intensive measures of experienced affect (Stone and Mackie, 2013).

3. Experimental treatment and methods

3.1 Experimental set-up

Participants were randomly assigned during pregnancy to an intervention group receiving the *Preparing for Life* (*PFL*) HVP (*PFL* and The Northside Partnership, 2008) and the Triple P Positive Parenting Program (Sanders et al., 2003), or a control group. The treatment aims to improve the health and development of children by intervening during pregnancy and working with families until the children start school at age 4/5. Home visiting is a widely used form of early intervention which provides parents with direct instruction on parenting practices, as well as information, social support, and access to other community services (Howard and Brooks-Gunn, 2009). The program was developed in response to evidence that children from the catchment area were lagging behind their peers in terms of cognitive and non-cognitive skills at school entry (Doyle et al., 2012). *PFL* is a manualized program which is grounded in the theories of human attachment (Bowlby, 1969), socio-ecological development (Bronfenbrenner, 1979), and social-learning (Bandura, 1977). The trial is registered with controlled-trials.com (ISRCTN04631728).

3.1.1 Treatment

The intervention prescribes twice monthly home visits, lasting approximately one hour, delivered by mentors from a cross-section of professional backgrounds including education, social care, and youth studies. Mentors received extensive training prior to program implementation and monthly supervision thereafter. Each family is assigned the same mentor over the course of the treatment where possible. The home visits are tailored based on the age of the child and the needs of the family and are guided by a set of Tip Sheets which present best-practice information on pregnancy, parenting, and child health and development.

This study refers to the impact of the treatment on maternal well-being and includes participants who were engaged with the program for at least two and a half years. The program is anticipated to have an impact on maternal well-being due to the nature of the mentor-mother relationship and the supports provided. Specifically, the mentors support mothers by building a strong relationship with them and helping them to improve their parenting and problem solving skills using role modelling, coaching, discussion, encouragement, and feedback. In addition, a number of Tip Sheets delivered between pregnancy and the child's second birthday focus on maternal personal and social well-being, including the mother's relationship with the father, social support, support services available in the community, self care, exercise, and postnatal depression. For example, one Tip Sheet provides information on the prevalence and symptoms of postnatal depression, while the Tip Sheet on relationships and quality time recommends that mothers talk to their partner every day and schedule time to be together. A further Tip Sheet on self-care suggests that mothers reward themselves by relaxing and doing something that makes them feel good.

The treatment group are invited to participate in an additional parenting course (Triple P Positive Parenting Program; Sanders et al., 2003) when their children are between 2 and 3 years old. Triple P promotes healthy parenting practices and positive parent-child attachment. Meta-analysis of Triple P has demonstrated positive effects for parents regarding parenting practices, and for children regarding social, emotional, and behavioral outcomes (Sanders et al., 2014). The majority of participants who availed of Triple P took part in Group Triple P which consists of five 2-hour group discussion sessions and three individual phone calls facilitated by the mentors.

3.1.2 Common supports

While the HVP and the Triple P program is the treatment under investigation, both the treatment and control groups receive common supports including developmental materials and book packs. Both groups are also encouraged to attend public health workshops on stress management and healthy eating which are already available to the wider community, however relatively few members of either group attend these sessions. The control group also has access to a support worker who can help them avail of community services if needed, while this function is provided by the mentors for the treatment group. Further information on the program and the design of the evaluation has been published elsewhere (Doyle, 2013).

3.2 Participants

The original RCT study enrolled pregnant women from a suburban community in Dublin, Ireland, which had above national average rates of unemployment, early school leavers, lone parent households, and public housing. All pregnant women from this community, regardless of parity, were eligible for voluntary participation in the program. Recruitment took place between 2008 and 2010 through two maternity hospitals or self-referral in the community. In total, 233 participants were recruited and an unconditional probability randomization procedure assigned 115 participants to the treatment group and 118 to the control group. A computerised randomisation program was used, with no stratification or block techniques.

Of the original 233 participants, 192 were eligible to participate in the present study of well-being as they had not voluntarily or involuntarily dropped out of the study at the time of data collection.[2] Appendix Figure A1 depicts the recruitment of participants in the original trial and the present study. Mothers were invited to take part in this study by telephone, and a flyer was sent to those who could not be reached. The study was described to participants as "A Day in the Life of a Parent", the goal of which was to collect information on parents' daily lives and to learn about the different emotions parents experience during a typical day. Of the 192 target participants, 102 (treatment = 46; control = 56) took part, 34 refused[3], 2 agreed but did not participate, and 54 could not be reached by telephone, text, or letter.[4] The participants were at various stages in the program when they participated in this study; the youngest child was 24.6 months and the oldest child was 62.5 months old.[5]

Participants who chose to take part did not differ from those who refused on 93% of the baseline characteristics collected during pregnancy (106/114).[6] Significant differences on 7% of measures indicated that mothers who chose to take part in this study were somewhat more disadvantaged than those who did not participate. For example, they reported consuming more alcohol, availing of a greater number of services, being more open [as per the Ten Item Personality Index (TIPI; Gosling et al., 2003)], having their activity impaired by illness, being in receipt of social welfare payments, and meeting the risk cutoff for lack of

[2] 32 participants (treatment = 17; control = 15) voluntarily dropped out of the study and a further 9 (treatment = 6; control = 3) involuntarily chose to drop out due to miscarriage, death, child death, or moved out of the catchment area at the time of data collection for the present study.

[3] The leading reason for refusal was lack of time, particularly amongst working participants.

[4] Of the 92 participants who did not participate in the present study, 83 completed a baseline interview, 70 completed a 6 month interview, 66 completed a 12 month interview, 57 completed an 18 month interview and 65 completed a 24 month interview.

[5] Length of time in the program is controlled for in all analysis.

[6] Two-tailed tests were conducted, p-values <0.10 were considered significant.

empathy towards their child's needs [as per the Adult Adolescent Parenting Inventory (AAPI; Bavolek and Keene, 2002)].

Appendix Table A1 presents descriptive statistics on the participating sample using baseline data disaggregated by treatment status. The treatment and control mothers were largely equivalent on the majority of demographic indicators. On average, mothers were between 25 and 26 years old and had one non-*PFL* child. Approximately half of participants were first time mothers, over 55% lived in public housing, and approximately 40% had not completed a second level education and identified themselves as being unemployed. However, a significantly higher proportion of treatment mothers had a boy as their *PFL* child (48%) than control mothers (31%). A detailed analysis of differences between the participating treatment and control groups on 114 baseline characteristics identified that the groups did not differ on 85% (99/116) of measures. Given the sample size, it is not possible to control for all variables upon which the two groups differ, therefore a multivariate logit model was estimated to determine the most relevant predictors of group membership.[7] Characteristics which emerged as significant predictors were then controlled for in the estimation of treatment effects.[8] In addition, we control for the infant's gender and the length of time exposed to the program at the time of data collection. Program duration differs for each participant as data collection was conducted over a one year period, and recruitment into the program took place over two and a half years.

3.3 Data collection

The study procedure was approved by the institution's human research ethics committee and maternity hospitals' respective ethics committees. The survey was piloted between November 2012 and January 2013 with a convenience sample of parents (n = 5), *PFL* program staff (n = 7), and *PFL* pilot families (n = 5). Data collection commenced in February 2013 and ended in November 2013 when the target sample was exhausted. Participants were visited in their homes or a community centre (based on the participants' preference) by a researcher, who was blind to treatment assignment, on two occasions over a three day period.[9] On the first day participants were given diaries and asked to record the next

[7] Three of the 15 variables were excluded as they either had too much missing data or were collinear with another control variable. Two control variables with minimal amounts of missing data were imputed so as to maintain sample size.

[8] The control set is composed of an emotional attachment score, a self-efficacy score, the number of neighbours known by the participant, whether or not the participant exercises at least three times per week, a community service use variable, and whether or not the participants' pregnancy was planned.

[9] The three day period never encompassed a weekend day.

day's activities (study day). On the third day the survey was completed. Participants were given a €20 (~$23) voucher as a thank you for their participation.

The survey consisted of: an adapted *Day Reconstruction Method* (DRM; Kahneman et al., 2004), mood yesterday questions, global questions of life satisfaction and the Parenting Stress Index (Abidin, 1995). All measures were administered by researchers using laptop computers or paper questionnaires, with the exception of the PSI which was self-completed by the participant. The survey took approximately 50 minutes to complete.

3.4 Instruments

Adapted Day Reconstruction Method (DRM; Kahneman et al., 2004). The DRM was adapted for this study based on the research question, literature review, and piloting. To assist the completion of the DRM, participants were asked to keep a diary of the study day broken down into episodes across the morning, afternoon, and evening. Participants used their diary as a prompt to describe each of the day's episodes in terms of the time it began and ended, the type of activity they were participating in - in terms of 21 possibilities[10], where they were - in terms of three possibilities[11], and who they were interacting with, either in person or on the phone - in terms of 15 possibilities[12]. Participants were also asked to rate each episode in terms of 12 affect states including 5 positive states (*happy*, *affectionate*, *competent*, *relaxed, in control*), and 7 negative states (*depressed, impatient*, *criticized*, *angry*, *frustrated*, *irritated, stressed*) on a 7-point Likert scale from *not at all* to *very strongly*. Episodes were demarcated collaboratively by the participant and the researcher in order to provide the most accurate breakdown of the day.[13] On average, episodes lasted 80 minutes, and participants recorded approximately 11 episodes per day, which is in line with prior research employing the DRM (e.g. Daly et al., 2010).

The 12 individual affect states are examined separately across the entire day and are also averaged to create positive and negative affect scores. The difference between positive and negative affect is also calculated to provide an overall measure of utility, known as net affect. All scores are weighted by episode length, such that longer episodes contribute more towards an individual's affect state than shorter episodes.

[10] Grooming/care, exercising, attending training, paid work, preparing food, eating, housework, computer/email/internet, socialising, on the phone/skype, watching TV, relaxing, sleeping, commuting, shopping, taking care of child(ren), playing with child(ren), putting child(ren) to bed, getting child(ren) dressed, feeding child(ren), and other.

[11] Home, work, on the road, and elsewhere.

[12] Alone, *PFL* child, other child(ren), spouse/partner, own parent(s), other relatives, partner's parent(s), partner's child(ren), partner's relatives, friends, clients/customers, other people's child(ren), work colleagues, health professional(s), and other.

[13]While the DRM is typically self-administered, collaborative administration was deemed most appropriate to limit barriers to participation arising from literacy difficulties.

To overcome the potential issue of different participants interpreting the affect states in a different manner, we also use the *U-index*. If participants anchor themselves at different points along the Likert scale, interpersonal comparisons may be meaningless. Thus, Kahneman and Krueger (2006) propose the *U-Index* which captures the proportion of time a participant spends in an unpleasant state. An episode is categorized as unpleasant if the highest rated affect state is a negative one. Crucially, the *U-Index* only relies on an ordinal, as opposed to a cardinal, ranking of feelings. Therefore, all participants need not view a certain point on the scale as being precisely equivalent, but rather, they only need to have the same ranking of affect states. If we denote negative affect as *NA* and positive affect as *PA*, with *K* negative affect states and *L* positive affect states then the *U-Index* for person i during episode j is defined by:

$$U_{ij} = \begin{cases} 1 \quad if \ \max\{NA_{ij}^{K}\} > max\{PA_{ij}^{L}\} \\ 0 \quad if \ \max\{PA_{ij}^{L}\} \geq max\{NA_{ij}^{K}\} \end{cases}$$

The *U-Index* is also weighted by episode length. The resulting score represents the proportion of time where a respondent's strongest emotion was a negative one.

For all scores derived from the DRM, we compare the treatment and control groups for the entire day and for subsets of episodes broken down by the time the participant was with and without the *PFL* target child.

Measures of mood yesterday. To explore the utility of a less intensive proxy of experienced affect, participants were asked to provide global ratings of their mood for the study day. Specifically, participants were asked to indicate the percentage of time they spent in *a bad mood, a little low or irritable*, *in a mildly pleasant mood*, and *in a very good mood* in relation to the day overall and separately in terms of the time they spent with their child(ren). A binary mood variable is created where being in a *mildly pleasant mood* and being in a *very good mood* are considered positive, while being in a *bad mood* and being *a little low or irritable* are not.

Global life satisfaction. To assess participants' global evaluations of their well-being, three life satisfaction questions were included. Participants were asked to indicate the degree to which they were satisfied with their "life as a whole", "life at home", and their "life as a parent" on a 4-point Likert scale from *very unsatisfied* to *very satisfied*. Three binary variables (satisfied plus very satisfied versus unsatisfied plus very unsatisfied) are created.

Parenting Stress Index Short Form (PSI; Abidin, 1995).[14] The PSI includes 36 items rated on a 5-point Likert scale ranging from *strongly disagree* to *strongly agree*. The scale yields a total stress score and three subscale scores: Parental Distress, Parent-Child Dysfunctional Interaction, and Difficult Child.[15] Responses are summed to generate scores for each of the subscales (scoring range 12 – 60) and the Total Stress score (scoring range 36 – 180). A binary variable is also created to represent mothers scoring above a cut-off of 90, indicating a high level of stress.[16] The PSI also contains a measure of defensive responding (Abidin, 1995) derived from the widely used Crowne-Marlowe Social Desirability Scale. These questions pertain to routine parenting experiences, a denial of which can be interpreted as defensive, rather than accurate, responding. A score of 10 or below on this scale indicates defensive responding. Both a cut-off and a continuous score of defensive responding are computed.

4. Econometric framework

4.1 Empirical approach

This study adopts an intention-to-treat approach, regardless of the number of home visits delivered or Triple P attendance. The standard treatment effect framework describes the observed outcome Y_i of participant $i \in I$ by:

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

where $I = \{1 \dots N\}$ denotes the sample space, D_i denotes the treatment assignment for participant i ($D_i = 1$ for the intention-to-treat sample , $D_i = 0$ otherwise) and $(Y_i(0),\, Y_i(1))$ are potential outcomes for participant i. We test the null hypothesis of no treatment effect on maternal well-being via:

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i \tag{2}$$

[14] Nine participants did not complete the PSI at the time of their interview. For these participants PSI scores from their most recent interview conducted as part of the main evaluation were employed. On average, these PSI measures were administered 4.6 months prior to the present study. When these participants are removed from the analysis the results do not change.

[15] Cronbach's alpha is used to assess the internal consistency of the PSI. Total Stress Score (36 items, α=0.90), Parental Distress (12 items, α=0.90), Parent-Child Dysfunctional Interaction (12 items, α=0.90), and Difficult Child (12 items α=0.89). These indicate a high degree of internal consistency.

[16] In accordance with the manual, subdomain and total scores were not computed for participants who were missing data on more than one item on a given subscale. This affected one participant on the Parent Distress subscale, two participants on the Parental Child Dysfunctional Interaction subscale, seven participants on the Difficult Child subscale and eight participants on Total and Cut-Off scores.

Equation 2 is estimated using t-tests/OLS regressions for continuous outcomes and chi-squared tests/logistic regressions for binary outcomes, both excluding and including relevant group differences. Permutation-based hypothesis testing is also used as it does not depend on distributional assumptions and thus facilitates the estimation of treatment effects in small samples (Ludbrook and Dudley, 1998). A permutation test relies on the assumption of exchangeability under the null hypothesis. If the null hypothesis is true, which implies that the program has no impact on well-being, then taking random permutations of the treatment indicator does not change the distribution of outcomes for the treatment or control group. Permutation tests work by calculating the observed test statistic by comparing the outcomes of the treatment and control group. Then, the data are repeatedly shuffled so that the treatment assignment of some participants is switched between the groups. The p-value for the permutation test is computed by examining the proportion of permutations that have a test statistic more extreme than the observed test statistic in the original sample. For the unconditional models, permutation tests, based on 100,000 replications are used to estimate the program's impact.

The permutation procedure relies on the exchangeability properties of the joint distribution of outcomes and treatment assignment. When this testing is applied to a randomized sample, the exchangeability property is easily achieved. When the exchangeability property is not obvious, e.g. the two groups differ on certain characteristics, a conditional inference can be implemented using a revised version of a permutation test that relies on restricted classes of permutations. This procedure uses *the conditional exchangeability property* and tests for program effects, while controlling for a set of variables upon which the joint distribution of outcomes and treatment assignment is exchangeable.

Conditional permutation testing first partitions the sample into subsets, termed *orbits*, each consisting of participants with common background measures. Under the null hypothesis of no treatment effect, treatment and control outcomes have the same distributions within an orbit. Thus, the exchangeability assumption is restricted to strata defined by the controls. In our conditional analysis we include eight control variables. One binary variable is used to produce the orbits: the child's gender. However, using orbits proves problematic with multiple conditioning variables, as the strata become too small, leading to a lack of variation within each orbit. To circumvent this problem and obtain restricted permutation orbits of reasonable size, we assumed a linear relationship between the remaining seven[17] conditioning variables and the outcomes.

Thus, we partition the data into orbits on the basis of the child's gender and then regress each outcome on the seven variables assumed to share a linear relationship with the outcomes. Next, the residuals are permuted, based on 100,000 replications, from this regression within the orbits. This method is referred to as the Freedman–Lane procedure (Freedman and Lane, 1983) and was found to be statistically sound in a series of Monte Carlo studies (Anderson and Legendre, 1999). Heckman et al., (2010) applied this procedure to an analysis where the randomization was compromised so that the exchangeability property was not guaranteed. The results presented in Section 5 include both conditional and unconditional permutation testing *p*-values from two-tailed tests.

4.2 Additional analysis

Analysing the impact of the program on multiple well-being measures increases the likelihood of a Type-1 error and studies of RCTs have been criticized for overstating treatment effects due to this 'multiplicity' effect (Pocock et al., 1987). To address this issue and assess the robustness of our results, we employ the stepdown procedure described in Romano and Wolf (2005). The stepdown procedure involves calculating a t-statistic for each null hypothesis in a family of outcomes and placing them in descending order. Using the permutation testing method, the largest observed t-statistic is compared with the distribution of maxima permuted t-statistics. If the probability of observing this statistic by chance is high ($p \geq 0.1$), we fail to reject the joint null hypothesis that the treatment has no impact on any outcome in the family of measures being tested. If the probability of observing this t-statistic is low ($p < 0.1$), we reject the joint null hypothesis and proceed by excluding the most significant individual hypothesis and test the subset of hypotheses that remain for joint significance. This process of dropping the most significant individual hypothesis continues until only one hypothesis remains. 'Stepping down' through the hypotheses allows us to isolate the hypotheses that lead to a rejection of the null. This method is superior to the Bonferroni adjustment method as it accounts for interdependence across outcomes.

In this study the well-being measures are placed into 14 families for the individual permutation tests.[18] The stepdown procedure is then conducted on the families where we

[17] The control set is composed of a participant's program duration, an emotional attachment score, a self-efficacy score, the number of neighbours known by the participant, whether or not the participant exercises at least three times per week, a community service use variable, and whether or not the participants' pregnancy was planned.

identify significant individual differences. The outcome measures included in each family should be correlated and represent an underlying construct. However, outcomes which are derived from the same measure should not be included in the same stepdown family. For this reason, we apply the stepdown procedure to 9 of the 14 families.[19]

In addition to examining differences in well-being, we also explore patterns of time use across the treatment and control groups regarding interactions, locations, and activities. We calculate the proportion of episodes involving interactions with the *PFL* child, the participant's partner, and other family members.[20] In terms of locations, we examine the proportion of episodes which take place in the home and in the workplace. Finally, we calculate the proportion of episodes where the participant was looking after and playing with their children[21] and where they were relaxing/socializing[22], engaging in housework/cooking[23], exercising or commuting.

We apply two-tailed tests for both the individual and stepdown tests as we are not proposing a specific directional hypothesis regarding the program's impact on well-being.

5. Results

5.1 Correlation across well-being measures

Appendix Table A2 presents individual level correlations for the well-being measures. By construction, net affect exhibits a strong positive correlation with positive affect and is negatively correlated with both negative affect and the U-Index. Additionally, positive affect exhibits a moderate negative correlation with both negative affect and the U-Index. As one would expect, negative affect is strongly associated with the U-Index. The experienced measure of mood yesterday is also moderately correlated with the four measures of well-being derived from the DRM. However, a previous study found a higher degree of association between similar measures (Christodoulou, Schneider, and Stone, 2014). In

[18] Overall net affect, the U-Index, overall positive affect, positive emotions during the day as a whole, positive emotions during time spent with the *PFL* child, positive emotions during time without the *PFL* child, overall negative affect, negative emotions during the day as a whole, negative emotions during time spent with the *PFL* child, negative emotions during time without the *PFL* child, mood, life satisfaction PSI total scores, and PSI subdomains.

[19] For example, as the measure of net affect during times spent with the *PFL* child and the measure of net affect during time spent without the *PFL* child, are both constructed from overall net affect measure, it is not possible to test the joint significance of these three variables in the same stepdown family. The 5 groups that were ineligible for stepdown analysis were: net affect, the U-Index, overall positive affect, overall negative affect, and PSI total scores.

[20] This category includes the participants' parents, other relatives of the participant, their partners' parents and their partners' other relatives and it not include the participants' children or the participants' partner.

[21] This category includes getting children dressed, feeding children, getting children ready for bed, and caring for children.

[22] This category includes socialising, relaxing, browsing the computer, talking on the phone, and watching TV.

[23] This category includes doing housework, preparing food, and shopping.

addition, the global measure of life satisfaction displays only weak correlations with the experienced measures of net affect, positive affect, the U-Index, and the measure of mood yesterday. Life satisfaction is significantly negatively correlated with negative affect and total stress as measured by the PSI, but the magnitude of these associations is modest. The PSI is also correlated with net, positive, and negative affect, and mood yesterday, but is not related to the U-Index. This analysis suggests that global and experienced measures of well-being may represent different concepts.

5.2 Descriptive statistics on affect measures[24]

For each episode, respondents report a score for a range of affect states which are classified as being either positive (*happy, competent, relaxed, affectionate, in control*) or negative (*impatient, frustrated, depressed, irritated, angry, stressed, criticized*). To generate descriptive statistics, the positive and negative affect values are standardized for the entire sample to have a zero mean and a standard deviation of one. Every episode recorded is assigned an hour corresponding to the midpoint of the episode. For each midpoint hour from 08:00 to 22:00, the average positive and negative affect is calculated separately for the treatment and control groups.

Figure 1 illustrates the pattern of average positive affect over the course of the study day and shows that the treatment group report higher positive affect scores at every hour, compared to the control group.

[24] In order to gauge the normality of the study day, participants were asked to rate how the study day compared to that day of the week typically, on a five-point Likert scale from *much worse, to much better*, both overall and separately in terms of the time they spent with their child(ren). Participants were also asked to rate how anxious they felt on the study day compared to that day of the week typically, on a five-point Likert scale from *a lot less anxious, to a lot more anxious*, both overall and separately in terms of the time they spent with their child(ren). There were no differences found between the treatment and control groups on either of these variables suggesting the study took place on an a typical day. The majority of participants reported that the study day was either typical or better compared to that day of the week usually, both for the day as a whole (79%) and separately in terms of time spent with their child(ren) (83%). The majority of participants also reported that they felt less anxious on the study day compared to that day of the week usually, both for the day as a whole (57%) and separately in terms of time spent with child(ren) (88%).

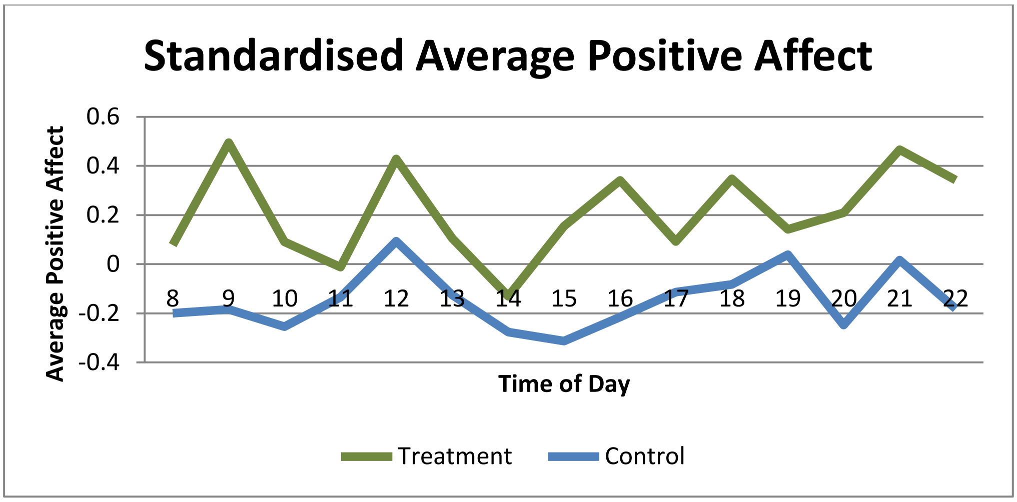


Fig.1. Standardized average positive affect for treatment and control groups.

Conversely, Figure 2 indicates that there is no clear difference in negative affect between the two groups. Both the treatment and control groups display a similar pattern of mid-morning and mid-afternoon peaks, followed by an evening decline as is typical (e.g. Daly et al., 2010; Stone et al., 2006).

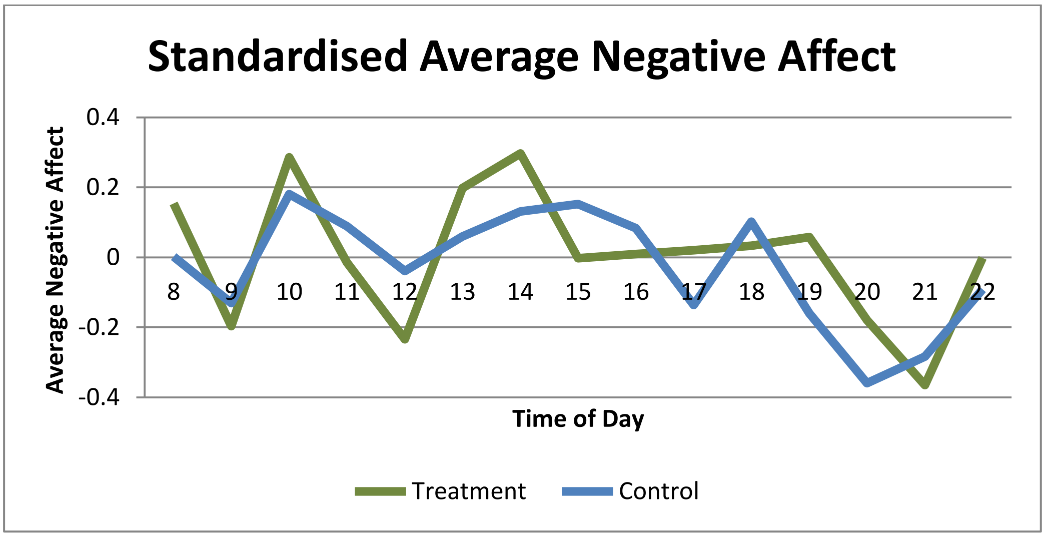


Fig.2. Standardized average negative affect for treatment and control groups.

5.3 Estimation of Treatment Effects

Below we present estimates of treatment effects for experienced measures of mood yesterday, net affect, the U-Index (Table 1), positive affect (Table 2), and negative affect (Table 3) scores. Table 4 presents the results using global measures of life satisfaction and the standardized measure of parenting stress. The unconditional means and standard deviations are reported throughout. Four columns of *p*-values are presented in each table representing the statistical significance of the estimated treatment effect from an unconditional t-test/chi-squared test, an unconditional permutation test, a conditional t-test/chi-squared test, and a conditional permutation test, respectively.[25] Given the observed differences between the treatment and control groups at baseline the conditional results represent the most reliable set of findings. Overall, the t-tests and the permutation tests produce very similar results.

Table 1 compares the treatment and control groups in terms of their mood yesterday, net affect, and U-Index for the day as a whole and also time spent with and without the *PFL* child. It shows that both groups report spending approximately three-quarters of the study day in a positive mood. This increases to four-fifths when participants restricted their judgements to the time spent with children. Furthermore, the treatment group reports spending a significantly higher proportion of the study day in a positive mood than the control group in the conditional models.

In terms of the DRM measures, on average, participants in both groups report a net affect score of approximately 3 over the course of the study day. This implies that participants experience positive emotions three points more intensively on the 0-6 Likert scale than negative emotions. Therefore, it is unsurprising that both groups spend approximately only 10% of their day in an episode where the strongest emotion is a negative one, as shown by the U-Index. Both groups experience a slight decline in net affect and a corresponding slight rise in the U-Index in episodes when they are without their *PFL* child. No significant treatment effects are identified for the net affect or U-Index measures.

Table 1

Treatment effects for experienced well-being: Mood yesterday, net affect and U-index.

	M_{TREAT} (*SD*)	M_{CONTROL} (*SD*)	*Unconditional*		*Conditional*	
			p^1	p^2	p^1	p^2
Mood Yesterday						
Portion of day spent in a positive mood	0.76 (0.18)	0.71 (0.25)	0.321	0.308	0.011**	0.014**
Portion of day spent with children in a positive mood	0.83 (0.21)	0.84 (0.19)	0.821	0.827	0.510	0.471
Net affect	3.03 (1.41)	2.84 (1.37)	0.355	0.512	0.330	0.377
Net affect during time spent with *PFL*	2.98	2.95	0.829	0.917	0.601	0.787

[25] For continuous outcomes t-tests refer to OLS coefficients, for binary outcomes logistic regressions were estimated and chi-squared tests assessed the significance of observed treatment coefficients.

Child	(1.58)	(1.38)				
Net affect during time spent without *PFL* child	3.00 (1.78)	2.68 (1.59)	0.141	0.356	0.266	0.219
U-Index	0.10 (0.14)	0.09 (0.18)	0.965	0.689	0.745	0.704
U-Index during time spent with *PFL* child	0.10 (0.16)	0.08 (0.18)	0.506	0.461	0.571	0.457
U-Index during time spent without *PFL* child	0.11 (0.24)	0.12 (0.27)	0.429	0.875	0.979	0.776

Notes: The sample size is 101 (Treatment=46, Control=55), except when we restrict the analysis to time spend without the *PFL* child, as 5 control participants (Treatment=46, Control=50) did not record any episodes without their *PFL* child, and apart from Mood Yesterday (Treatment=45, Control=55). 'M' indicates the unconditional mean. 'SD' indicates the unconditional standard deviation. [1] two-tailed t-test p-value [2] two-tailed p-value from an individual permutation test with 100,000 replications, * p < .10, ** p < .05, *** p < .01

Table 2 compares the treatment and control groups in terms of their overall positive affect and individual positive affect states for the day as a whole and also time spent with and without the *PFL* child. Overall, feelings of competence and control receive the highest ratings, while feeling relaxed receives the lowest. This pattern differs slightly depending on whether participants were in episodes with/without their *PFL* child, with participants reporting substantially higher levels of affection during episodes with the *PFL* child. A treatment effect is identified for overall positive affect in all 4 models; however it is only significant for the time spent without the *PFL* child. This difference is primarily driven by a decline in the control group's positive affect during episodes in which they are not with their *PFL* child, while the treatment group is slightly more stable in terms of positive affect during episodes with or without their *PFL* child.

In terms of the individual positive affect states, we find that treatment participants report significantly higher levels of happiness for the day overall and during times spent without the *PFL* child in all 4 models. In 3 models, the treatment group also report higher levels of happiness during times spent with the *PFL* child. However, this result is not present in the conditional permutation model which represents our best estimate of program impact. The groups do not significantly differ on the remaining four positive affect states.

Tests comparing positive affect states when with and without the *PFL* child (not reported) show that participants from both groups are significantly less affectionate during episodes without their *PFL* child, yet the control group experience a larger decline. Additionally, control group participants feel significantly less in control when they are without their *PFL* child than when they are with the *PFL* child, while treatment participants are significantly more relaxed when without, compared to with, their *PFL* child.

Table 2

Treatment effects for experienced well-being: Positive affect.

	M_{TREAT} (*SD*)	M_{CONTROL} (*SD*)	*Unconditional*		*Conditional*	
			p^1	p^2	p^1	p^2
Overall						
Positive affect	3.94 (0.96)	3.66 (0.95)	0.151	0.150	0.163	0.187
Positive affect during time spent with *PFL* child	3.97 (1.02)	3.77 (1.00)	0.336	0.336	0.298	0.390
Positive affect during time spent without *PFL* child	3.84 (1.13)	3.48 (0.92)	0.088*	0.090*	0.099*	0.095*
Positive affect states						
Happy	4.03 (1.00)	3.59 (1.12)	0.043**	0.041**	0.054*	0.054*
Affectionate	3.75 (1.49)	3.43 (1.38)	0.271	0.273	0.234	0.232
Competent	4.40 (1.04)	4.18 (1.12)	0.324	0.320	0.371	0.416
In Control	4.25 (1.16)	4.04 (1.19)	0.379	0.378	0.541	0.629
Relaxed	3.24 (1.16)	3.04 (1.16)	0.410	0.409	0.337	0.370
Positive affect states during time spent with PFL child						
Happy	3.99 (1.22)	3.59 (1.17)	0.094*	0.096*	0.075*	0.121
Affectionate	4.25 (1.42)	3.98 (1.40)	0.340	0.341	0.257	0.341
Competent	4.34 (1.09)	4.13 (1.22)	0.358	0.353	0.395	0.443
In Control	4.25 (1.20)	4.13 (1.17)	0.607	0.607	0.985	0.983
Relaxed	2.94 (1.34)	3.00 (1.21)	0.834	0.836	0.788	0.861
Positive affect states during time spent without PFL child						
Happy	3.98 (1.07)	3.50 (1.25)	0.045**	0.045**	0.073*	0.055*
Affectionate	3.08 (1.89)	2.57 (1.59)	0.159	0.162	0.194	0.154
Competent	4.31 (1.40)	4.16 (1.15)	0.550	0.553	0.360	0.397
In Control	4.17 (1.44)	4.00 (1.29)	0.522	0.522	0.404	0.457
Relaxed	3.67 (1.59)	3.18 (1.27)	0.100	0.103	0.199	0.203

Notes: The sample size is 101 (Treatment=46, Control=55), except when we restrict the analysis to time spend without the *PFL* child, as 5 control participants (Treatment=46, Control=50) did not record any episodes without their *PFL* child. 'M' indicates the unconditional mean. 'SD' indicates the unconditional standard deviation. [1] two-tailed t-test p-value. [2] two-tailed p-value from an individual permutation test with 100,000 replications, * p < .10, ** p < .05, *** p < .01

Table 3 compares the treatment and control groups in terms of their negative affect and individual negative affect states for the entire day and the time participants spent with and without their *PFL* child. No significant treatment effects are identified in any of the models. While the pattern across groups is less consistent than positive affect, both treatment and control participants tend to give higher ratings regarding feeling stressed and impatient, with depressed and criticised receiving the lowest ratings. Overall, ratings of negative affect states seem to be slightly less intense when participants were not with their *PFL* child, although none of these differences are significant for either group (not reported).

Table 3

Treatment effects for experienced well-being: Negative affect.

Negative Affect	M_{TREAT} (*SD*)	M_{CONTROL} (*SD*)	*Unconditional*		*Conditional*	
			p^1	p^2	p^1	p^2
Overall						
Negative affect	0.91 (0.79)	0.82 (0.76)	0.547	0.551	0.999	0.946
Negative affect during time spent with *PFL* child	0.98 (0.88)	0.82 (0.73)	0.309	0.323	0.714	0.571
Negative affect during time spent without *PFL* child	0.84 (0.97)	0.80 (0.92)	0.831	0.919	0.869	0.732
Negative affect states						
Stressed	1.47 (1.25)	1.24 (1.08)	0.320	0.329	0.710	0.660
Irritated	1.29 (1.12)	1.08 (1.05)	0.338	0.343	0.773	0.847
Frustrated	1.26 (1.02)	1.10 (1.00)	0.422	0.426	0.889	0.843
Angry	0.66 (0.84)	0.55 (0.84)	0.504	0.510	0.939	0.901
Impatient	1.27 (1.15)	1.32 (1.02)	0.829	0.830	0.794	0.792
Depressed	0.23 (0.37)	0.28 (0.50)	0.627	0.622	0.429	0.511
Criticized	0.18 (0.40)	0.16 (0.36)	0.781	0.786	0.611	0.777
Negative affect states during time spent with PFL child						
Stressed	1.61 (1.45)	1.25 (1.08)	0.155	0.167	0.465	0.385
Irritated	1.36 (1.22)	1.04 (0.98)	0.153	0.164	0.289	0.350
Frustrated	1.37 (1.19)	1.11 (1.00)	0.233	0.245	0.578	0.468
Angry	0.66 (0.87)	0.56 (0.85)	0.584	0.593	0.894	0.828
Impatient	1.43 (1.26)	1.36 (1.09)	0.783	0.787	0.980	0.783
Depressed	0.24 (0.53)	0.24 (0.49)	0.989	0.990	0.315	0.595

Criticised	0.22 (0.49)	0.17 (0.39)	0.600	0.611	0.875	0.915
Negative affect states during time spent without PFL child						
Stressed	1.36 (1.61)	1.23 (1.31)	0.672	0.674	0.928	0.865
Irritated	1.16 (1.38)	1.03 (1.33)	0.634	0.636	0.921	0.784
Frustrated	1.10 (1.31)	1.07 (1.29)	0.895	0.896	0.687	0.590
Angry	0.70 (1.21)	0.58 (1.15)	0.620	0.625	0.949	0.970
Impatient	1.15 (1.46)	1.12 (1.29)	0.932	0.934	0.922	0.816
Depressed	0.26 (0.57)	0.44 (0.91)	0.255	0.256	0.615	0.525
Criticised	0.14 (0.58)	0.13 (0.34)	0.922	0.929	0.745	0.700

Notes: The sample size is 101 (Treatment=46, Control=55), except when we restrict analysis to time spend without *PFL* child, as 5 control participants (Treatment=46, Control=50) did not record any episodes without their *PFL* child. 'M' indicates the unconditional mean. 'SD' indicates the unconditional standard deviation. ^{1} two-tailed t-test p-value ^{2} two-tailed p-value from an individual permutation test with 100,000 replications, * p < .10, ** p < .05, *** p < .01

Table 4 presents estimates of treatment effects for the global measures of life satisfaction and the standardized measure of parenting stress. In terms of life satisfaction, the vast majority of participants in both groups report that they are satisfied with their life overall, as a parent, and at home. A slightly higher proportion of treatment participants report that they are satisfied with their life in all three categories than control participants, however, none of these differences are statistically significant.[26]

In terms of participants' reports of parenting stress (PSI), the treatment and control groups report comparable levels of parenting stress and approximately 10% of participants in both groups report stress levels that are considered to be clinically significant. However, there are no significant treatment effects for any of the PSI scores. In addition, 24% of the treatment group and 27% of the control group meet the cut off for defensive responding suggesting that these participants may be positively biasing their responses based on their perception of socially desirable parenting experiences. Importantly, however, there are no significant differences between the groups in terms of defensive responding, suggesting no evidence of systematic mis-reporting by the treatment and control groups.

[26] Note that only 9 participants across both groups report being either *unsatisfied* or *very unsatisfied* with their life overall compared to 91 reporting being *satisfied* or *very satisfied* (the comparable figures for satisfaction as a parent and satisfaction with home life are 7 and 8 respectively), thus the small cell size in the binary variables should be noted when interpreting the results.

Table 4

Treatment effects for global well-being: Life satisfaction and Parenting Stress Index.

	N (n_{TREAT}/ n_{CONTROL})	M_{TREAT} (*SD*)	M_{CONTROL} (*SD*)	*Unconditional*		*Conditional*	
				p^1	p^2	p^1	p^2
Life Satisfaction							
Satisfaction with life as a parent	100 (45/55)	0.98 (0.15)	0.89 (0.31)	0.126	0.118	0.570	0.542
Satisfaction with home life	100 (45/55)	0.96 (0.21)	0.89 (0.31)	0.251	0.234	0.627	0.849
Satisfaction with life overall	100 (45/55)	0.93 (0.25)	0.89 (0.31)	0.465	0.477	0.908	0.674
PSI subdomains							
Parent-Child Dysfunctional Interactions	99 (45/54)	18.04 (5.44)	17.22 (5.40)	0.402	0.456	0.748	0.876
Difficult Child	94 (43/51)	22.42 (8.34)	22.18 (7.03)	0.944	0.881	0.501	0.560
Parental Distress	100 (45/55)	24.82 (8.39)	24.67 (8.50)	0.907	0.932	0.652	0.558
Total Parental Stress	93 (42/51)	64.52 (18.17)	64.02 (17.95)	0.888	0.894	0.566	0.550
Stress Cut-off	93 (42/51)	0.10 (0.30)	0.08 (0.27)	0.752	0.827	0.458	0.929
Defensive Responding	93 (42/51)	14.76 (5.24)	14.64 (5.05)	0.967	0.972	0.805	0.667
Defensive Responding Cut-off	93 (42/51)	0.24 (0.43)	0.27 (0.45)	0.731	0.694	0.995	0.639

Notes: 'N' indicates the sample size. 'M' indicates the unconditional mean. 'SD' indicates the unconditional standard deviation. [1] two-tailed t-test p-value [2] two-tailed p-value from an individual permutation test with 100,000 replications,* p < .10, ** p < .05, *** p < .01

5.4 Additional analysis

5.4.1 Stepdown analysis

Table 5 presents the unconditional and conditional stepdown results for the measures upon which we identified significant differences according to the individual tests in Tables 1-4. The first p-value in the conditional mood yesterday stepdown family is significant following adjustment for multiple comparisons, and is driven by the significant individual finding for the portion of day spent in a positive mood. In contrast, the stepdown families for positive affect states for the day as a whole or for episodes with and without their *PFL* child are not significant when the unconditional and conditional stepdown procedure is applied.

Table 5

Stepdown results.

	Unconditional p^1	*Conditional* p^2
Mood Yesterday		
Portion of day spent in a positive mood	~	0.027*
Positive affect states		
Happy	0.138	0.174
Positive affect states during time spent with PFL child		
Happy	0.294	~
Positive affect states during time spent without PFL child		
Happy	0.162	0.189

Notes: [1] two-tailed p-value from an unconditional stepdown permutation test with 100,000 replications.[2] two-tailed p-value from a conditional stepdown permutation test with 100,000 replications
* p < .10, ** p < .05, *** p < .01

5.4.2 Time Use

The few observed treatment effects may be driven by differences in time use across the two groups. Yet, as shown in Table 6, the treatment group spend approximately the same proportion of episodes with their *PFL* child (62%) as do the control group (66%). In addition, there are no differences regarding the proportion of episodes spent caring for or playing with their children, with both groups spending approximately 10% of their episodes playing with their children. The conditional results show that the treatment group are significantly more likely to spend a given episode with their relatives (excluding their children and partner). However, both groups spend a similar proportion of episodes alone and with their partners. The conditional results also show there are no differences by location, with both groups spending roughly two thirds of their episodes at home, and less than 6% of their episodes in work. There are also no differences in time use in terms of daily activities (relaxing/socializing, housework/cooking, commuting), apart from exercising where the control group spend a greater proportion of their episodes exercising. Note, however, that the proportion of episodes spent exercising is minimal. Overall, these results suggest that the higher positive affect experienced by the treatment group may be driven by the differences in the quality of episodes rather than differences in time use.

Table 6

Time use amongst treatment and control groups.

	$\%_{\text{TREAT}}$	$\%_{\text{CONTROL}}$	*Unconditional* p^1	*Conditional* p^2
Interaction				
With *PFL* child	61.89	66.28	0.125	0.262
With partner	16.70	22.09	0.019**	0.235
With relatives	22.99	16.45	0.008***	0.003***
Alone	9.49	10.89	0.445	0.201
Location				
At home	66.60	64.95	0.564	0.997
At work	5.89	3.16	0.029**	0.108
Activities				
Looking after children	44.20	46.84	0.399	0.377
Playing with children	8.84	8.97	0.962	0.574
Relaxing/socializing	24.95	25.42	0.881	0.653
Housework/cooking	26.92	29.40	0.376	0.533
Commuting	12.77	13.95	0.540	0.846
Exercising	1.57	2.16	0.501	0.000***

Notes: Unconditional percentages are reported**.** [1] two-tailed p-value from an individual unconditional permutation test with 100,000 replications.[2] two-tailed p-value from an individual conditional permutation test with 100,000 replications. * p < .10, ** p < .05, *** p < .01

6. Conclusion

Kahneman et al. (2004) has proposed that aggregated measures of experienced affect can be utilized as a measure of policy effectiveness and Dolan and White (2007) also discuss the possibility that such measures replace traditional quality of life questions in health care evaluations. However, to date, no study has attempted to integrate these insights into a formal policy evaluation.

This paper examines the utility effects of a targeted early intervention program using multiple measures of well-being. Based on the individual treatment effect results, we find some evidence that the *PFL* intervention generates higher levels of experienced positive affect using a Day Reconstruction Method, primarily for times when participants are without their target child. Interestingly, when positive DRM affect states are examined separately, we observe an individual treatment effect for happiness for the day overall and when participants are without the *PFL* child, however these results does not survive the stepdown procedure. These results are broadly consistent with participants' judgments for their overall levels of positive mood yesterday, where we observe a significant treatment effect in both the individual and stepdown results, yet not during times spent with children.[27] There are no treatment effects for negative aspects of well-being, irrespective of the measure used, including experienced negative affect, individual negative affect states, U-index scores, and parenting stress. Lastly, although higher proportions of the treatment group compared to the control group report being satisfied with their lives across three domains, these differences did not reach significance.

The concentration of the few identified treatment effects amongst positive, yet not negative, measures of well-being is broadly in keeping with the existing HVP literature. Systematic reviews have found that home visiting is typically not effective in ameliorating negative emotional states (Sweet and Appelbaum, 2004; Ammerman et al., 2010). Thus our findings are consistent with the view that targeted and intensive therapeutic supplements are needed in order for HVPs to alleviate negative affect states such as depression (Ammerman et al., 2010). In particular, the mentors in the *PFL* trial are not trained counsellors or clinical psychologists. Notwithstanding this, our findings demonstrate that a HVP may have an impact on some dimensions of positive affect, which questions the prevailing assumption, based predominantly on deficit measures of well-being, that HVPs do not influence parents' emotional states (Brooks-Gunn and Markman, 2005).

Understanding why the intervention has some impact on affect states during times spent without the target child (as demonstrated by the individual result for positive affect and stepdown result for mood yesterday), may be linked to the family investment theory. The intervention aims to heighten parents' awareness of being actively engaged when interacting with their child. If such investment confers an increased effort and burden on the parents, treatment mothers may particularly value times when they are not actively being a parent. While there are no differences in the amount of time participants spend with their children in

[27] Note that the DRM and the yesterday mood question are not directly equivalent given that the DRM is broken down by time spent with and without the *PFL* child, while the mood question was asked for the day as a whole and times spent with any of the participants' children.

either group, the level and intensity of their engagement may be enhanced by the intervention. Support for this interpretation can be drawn from previous DRM research which demonstrates that spending time with one's children is amongst the least enjoyable and least pleasurable activities that individuals engage in (Dolan and White, 2009; Kahneman et al., 2004). The transition to motherhood also appears to create an upward shift in experienced positive affect for leisure activities, suggesting that free time becomes more valuable when contrasted with the demands of parenting (Hoffenaar et al., 2010). Consequently, if treated parents become more effortful in an activity that is inherently low in pleasure – parenting - they may derive more pleasure from times when they are not engaging in this activity.

A second related pathway is that the intervention, through Tip Sheets and mentor support, encourages mothers to use their non-parenting time for self-care, relaxation, and social relationships. These supports may result in positive emotional experiences as rich social relationships are integral to optimizing happiness (Diener and Seligman, 2004), and socializing and relaxing typically receive the highest ratings of experienced positive affect on the DRM (Kahneman et al., 2004). While there are no differences in time use between the two groups, it is possible that the quality of these non-parenting experiences differ in some unobserved way. Finally, it is also possible that gains to maternal well-being are accrued indirectly, via the program's identified impact on the children's cognitive, emotional and physical well-being (see Doyle and the *PFL* Evaluation Team, 2013). However, directionality may be obscured here due to the dynamic and bidirectional interplay between child and maternal well-being (Elgar et al., 2004).

Another key question concerns the intervention's effect on daily experiences of well-being, including experienced affect and assessments of yesterday's mood, but not more global assessments of well-being such as life satisfaction.[28] The first possibility is that the DRM provides a more sensitive measure of well-being which avoids the cognitive filters that impinge upon global assessments of life satisfaction. Such filters may operate less intensively on measures of yesterday's mood (see Stone and Mackie, 2013). Another hypothesis is that global and experienced well-being are independent constructs, as is reflected in the recent conceptual shift to recognize experienced well-being and global/evaluative well-being as distinct psychological phenomena (Diener and Tay, 2014; Kahneman et al., 2010). Applied to our study, the absence of treatment effects for global well-being may be considered counterintuitive, if we believe the life satisfaction question should have encouraged

[28] While the treatment effects on the global measures did not reach significance, a clear pattern was discernible as the treatment group report higher levels of satisfaction on all three domains.

participants to focus on their participation in the program, its association with greater parenting competency, and anticipation of future benefits. Indeed, while Dolan and White (2009) found that spending time with children was low in pleasure, it was thought of as rewarding. Thus, the authors postulate that parenting may have a more positive influence on global aspects of well-being by providing individuals with a sense of purpose, connection, and contribution to personal goals. Another potential reason for this finding, discussed by Knabe and Rätzel (2011), is that participants habituate quickly to their circumstances - in this case treatment status - and thus the effects on global well-being may dissipate over time as, on average, the participants have spent four years in the program.

Given the absence of experimental studies examining the causal impact of policy interventions on experienced well-being, it is difficult to give precise comparisons to the magnitude of the finding on positive well-being. However, useful reference points may be provided by non-experimental studies. Comparing our individual happiness effect to the well-being effects observed in the original DRM study (Kahneman et al., 2004), we identify a similar magnitude to the effect of commuting (.49 points less than average well-being) and being alone (.48 points less than average). In addition, it is noteworthy that treated participants' average levels of happiness for times when they are without the study child (3.98), are very similar to those reported in Kahneman et al.'s original sample of employed women (3.96; Stone et al., 2006). This suggests that the treatment may raise the levels of well-being of a disadvantaged group closer to those that are typical of the population.

While this study is the first to our knowledge to test for the causal impact of a policy intervention on multiple measures of well-being, a number of methodological issues should be acknowledged. A common criticism of experimental trials is the use of self-report measures, which can be contaminated by social desirability when participants cannot be blinded to their treatment status. Experienced and global well-being, by definition, demands self-report. However, our results show that there are no systematic differences in social desirability between the treatment and control groups according to the defensive responding validity measure embedded within the PSI. An additional issue which is common to many experimental trials is small sample size. This issue is a particular concern in this study as the sample is smaller and relatively more disadvantaged than in the original *PFL* trial. Yet the sample size is equivalent to seminal studies of other early intervention programs, such as the Perry Preschool program and the Abecedarian program (see Heckman et al. (2010) and Campbell et al. (2014) for a discussion on the use of small samples in experimental trials). The permutation testing method helps to address this issue and is conditional on salient group differences. A further concern frequently associated with studies of HVPs, is the risk of overstating the program's impact due to multiple hypothesis testing. We address this using the stepdown procedure and highlight the significance of failing to account for this issue. The stepdown analysis shows that only the result for mood yesterday remains significant.

If the identified treatment effect for experienced positive mood is valid, this may confer meaningful benefits for mothers. Evidence suggests that positive emotions create an upward positive spiral in emotional well-being by enhancing an individual's cognitive coping strategies (Fredrickson and Joiner, 2002). Over time a causal relationship is believed to develop between positive affect and behaviors linked to more successful outcomes such as higher quality relationships, superior income and productivity, greater community participation, and improved health and mortality (Lyubomirsky, King, and Diener, 2005, see also Steptoe, Gibson, Hamer, and Wardle, 2007). Thus, the treatment effect identified here may have important implications for the cost-benefit analysis of the *PFL* program and similar HVPs in the future.

Using randomized controlled trials to examine the well-being effects of policy interventions is a growing area for economics. Our findings demonstrate the importance of measurement and conceptualization of well-being and of inferential techniques. Further research is needed to reconcile differences in treatment effects on global versus experienced measures of well-being and on positive and negative affect. These issues are important across many domains, including labor market and health interventions where there is also likely to be a substantial psychic benefit of successful program outcomes on top of the core measures being targeted. The issues discussed here point to the importance of conducting rigorous investigations into the impact of public policies on well-being.

References

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

Ammerman, R.T., Putnam, F.W., Bosse, N.R., Teeters, A.R., Van Ginkel, J.B., 2010. Maternal depression in home visitation: A systematic review. Aggression and Violent Behavior 15 (3), 191-200. doi: 10.1016/j.avb.2009.12.002

Anderson, M.J., Legendre, P., 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. Journal of Statistical Computation and Simulation 62 (3), 271–303. doi: 10.1080/00949659908811936

Atz, U., 2013. Evaluating experience sampling of stress in a single-subject research design. Personal and Ubiquitous Computing. Personal and Ubiquitous Computing 17 (4), 639-652. doi: 10.1007/s00779-012-0512-7

Bandura, A., 1977. Self-efficacy: Toward a unifying theory of behavioural change. Psychological Review 84 (2), 191-215. doi: 10.1037/0033-295X.84.2.191

Bavolek, S.J., Keene, R.G., 1999. Adult-adolescent parenting inventory - AAPI-2: Administration and development handbook. Family Development Resources, Inc, Park City, UT.

Becker, G.S., 1991. A Treatise on the Family (enlarged ed.). Cambridge, MA: Harvard University Press

Beshears J, Choi J.J, Laibson D., Madrian B.C., 2008. How Are Preferences Revealed? Journal of Public Economics 92 (8-9), 1787-1794. doi: 10.1016/j.jpubeco.2008.04.010

Bowen R.C., Wang, Y., Balbuena, L., Houmphan, A., Baetz, M., 2013. The relationship between mood instability and depression: Implications for studying and treating depression. Medical Hypotheses 81 (3), 459-462. doi: 10.1016/j.mehy.2013.06.010

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

Brooks-Gunn, J., Markman, L.S., 2005. The contribution of parenting to ethnic and racial gaps in school readiness. The Future of Children 15 (1), 139-167. doi: 10.1353/foc.2005.0001

Bowlby J., 1969. Attachment and Loss, Volume I: Attachment. Basic Books, New York, NY.

Bylsma. L.M., Taylor-Clift, A., Rottenberg, J., 2011. Emotional reactivity to daily events in Major and Minor Depression. Journal of Abnormal Psychology 120 (1), 155-167. doi: 10.1037/a0021662

Campbell F, Conti G, Heckman J.J., Moon, S.H., Pinto, R., Pungello, E., Pan, Y., 2014. Early childhood investments substantially boost adult health. Science 343, (6178), 1478-1485. 10.1126/science.1248429

Catalino, L.I., Fredrickson, B.L., 2011. A Tuesday in the life of a flourisher: The role of positive emotional reactivity in optimal mental health. Emotion, 11 (4), 938-950, doi: 10.1037/a0024889.

Christodolou, C., Schneider, S., Stone, A.A., 2014. Validation of a brief yesterday measure of hedonic well-being and daily activities: Comparison with the Day Reconstruction Method. Social Indicators Research, 115, 907-917. doi: 10.1007/s11205-013-0240-z

Craig, P. Dieppe, P., Macintyre, S., Nazareth, I., Petticrew, M., 2008. Developing and evaluating complex interventions: The new Medical Research Council guidance. BMJ 337, a1655. doi: 10.1136/bmj.a1655

Crnic, K.A., Low, C., 2002. Everyday stresses and parenting. In: Bornstein M. (ed.), Handbook of Parenting: Volume 5, Practical Issues in Parenting, (2nd ed.), 243-68. Lawrence Erlbaum Associates, Mahwah, NJ.

Crawford, J. R., Henry, J.D., 2004, The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. British Journal of Clinical Psychology, 43: 245–265. doi: 10.1348/0144665031752934

Daly, M., Delaney, L., Doran, P.P., Harmon, C., MacLachlan, M., 2010. Naturalistic monitoring of the affect-heart rate relationship: a Day Reconstruction study. Health Psychology, 29 (2), 186-195. doi: 10.1371/journal.pone.0043887

Deaton, A., Stone, A.A., 2013. Grandpa and the snapper: the wellbeing of the elderly who live with children. NBER Working Paper No 19100, June.

Deaton, A., Stone, A.A., 2014. Evaluative and hedonic wellbeing among those with and without children at home. PNAS, 111, 1328-1333.

Diener E., Seligman, M.E.P., 2004. Beyond money: Toward an economy of well-being. Psychological Science in the Public Interest 5 (1), 1-30. doi: 10.1111/j.0963-7214.2004.00501001.x

Diener, E., Tay, L., 2014. Review of the Day Reconstruction Method (DRM). Social Indicators Research 116 (1), 255-267. doi: 10.1007/s11205-013-0279-x

Dockray, S., Grant, N., Stone, A.A., Kahneman, D., Wardle, J., Steptoe, A., 2010. A comparison of affect ratings obtained with Ecological Momentary Assessment and the

Day Reconstruction Method. Social Indicators Research 99 (2), 269-283. doi: 10.1007/s11205-010-9578-7

Dolan, P., Kahneman, D., 2008. Interpretations of utility and their implications for the valuation of health. The Economic Journal 118, 215–234. doi: 10.1111/j.1468-0297.2007.02110.x

Dolan, P., Layard, R., Metcalfe, R., 2011. Measuring subjective well-being for public policy: recommendations on measures. Center for Economic Performance, Special Paper no. 23. Retrieved from: http://cep.lse.ac.uk/pubs/download/special/cepsp23.pdf

Dolan, P., White, M.P., 2007. How can measures of subjective well-being be used to inform public policy. Perspective on Psychological Science 2 (1), 71-85. doi: 10.1111/j.1745-6916.2007.00030.x

Dolan, P., White, M.P., 2009. Accounting for the richness of daily activities. Psychological Science 20 (8), 1000-1008. doi: 10.1111/j.1467-9280.2009.02392.x

Doyle, O., 2013., Breaking the cycle of deprivation: An experimental evaluation of an early childhood intervention. Journal of the Statistical and Social Inquiry Society of Ireland 2013; XLI: 92-111.

Doyle, O., McEntee, L., McNamara, K.A., 2012. Skills, capabilities, and inequalities at school entry in a disadvantaged community. European Journal of Psychology of Education 27 (1), 133-154. doi: 10.1007/s10212-011-0072-7.

Duflo, E., Glennerster, R., Kremer, M., 2008. Using randomization in development economics research: A toolkit. In Handbook of Development Economics, Volume 4, ed. T. P. Schultz and John Strauss, 3895–3962. Oxford, Elsevier.

Elgar, F.J., McGrath, P.J., Waschbusch, D.A., Stewart, S.H., Curtis, L.J., 2004. Mutual influences on maternal depression and child adjustment problems. Clinical Psychology Review, 24, 441-459.

Filene, J.H., Kaminski, J.W., Valle, L.A., Cachat, P., 2013. Components associated with home visiting program outcomes: A meta-analysis. Pediatrics, 132 (2), S100-S109. doi: 10.1542/peds.2013-1021H

Forgeard, M.J.C., Jayawickreme, E., Kern, M. L., Seligman, M.E.P., 2011. Doing the right thing: Measuring well-being for public policy. International Journal of Well-being 1 (1),http://internationaljournalofwellbeing.org/ijow/index.php/ijow/article/viewArticle/4

Fredrickson, B.L., Joiner, T., 2002. Positive emotions trigger upward spirals toward emotional well-being. Psychological Science 13 (2), 172-175. doi: 10.1111/1467-9280.00431

Freedman, D., Lane, D., 1983. A nonstochastic interpretation of reported significance levels. Journal of Business and Economic Statistics 1 (4), 292–298. doi: 10.2307/1391660.

Frey, B.S. Stutzer, A., 2002. What can economists learn from happiness research? Journal of Economic Literature 40, (2), 402-435.

Gertler, P., Heckman, J.J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., Chang, S.M., Grantham-McGregor, S., 2014. Labor market returns to an early childhood stimulation intervention in Jamaica. Science 344 (6187), 998 – 1001. doi: 10.1126/science.1251178

Gosling, S.D., Rentfrow, P.J., Swann, W.B., 2003. A very brief measure of the big-five personality domains. Journal of Research in Personality 37 (6), 504-528. doi: 10.1016/S0092-6566(03)00046-1

Gruber, J., Mullainathan, S., 2005. Do cigarette taxes make smokers happier? Advances in economic analysis and policy 5(1), 1 – 43. doi: 10.2202/1538-0637.1412

Heckman, J., Moon, S.H., Pinto, R., Savelyev, P. Yavitz, A., 2010. Analyzing social experiments as implemented: A re-examination of the evidence from the High Scope Perry Preschool Program. Quantitative Economics 1 (1), 1-46. doi: 10.3982/QE8.

Henquet, C., van Os, J., Kuepper, R., Delespaul, P., Smits, M., Campo, J.A., Myin-Germeys, I., 2010. Psychosis reactivity to cannabis use in daily life: an experience sampling study. British Journal of Psychiatry 196 (6), 447-453. doi: 10.1192/bjp.bp.109.072249.

Howard KS, Brooks-Gunn J., 2009. The role of home-visiting programs in preventing child abuse and neglect. The Future of Children 19 (2), 119-46.

Hoffenaar, P.J., van Balen, F. Hermanns, J., 2010. The impact of having a baby on the level and content of women’s well-being. Social Indicators Research 97 (2), 279-295. doi: 10.1007/s11205-009-9503-0

Kahneman, D., Deaton, A., 2010. High income improves evaluation of life but not emotional well-being. Proceedings of the National Academy of Sciences of the USA 107 (38), 16489-16493. doi: 10.1073/pnas.1011492107.

Kahneman, D., Krueger, A.B., 2006. Developments in the measurement of subjective well-being. The Journal of Economic Perspectives 20 (1), 3-24. doi: 10.1257/089533006776526030.

Kahneman, D., Krueger, A.B., Schkade, D.A., Schwarz, N, Stone, A.A., 2004. A survey method for characterizing daily life experience: The Day Reconstruction Method. Science 306 (5702), 1776-1780. doi: 10.1126/science.1103572

Kahneman, D., Riis, J., 2005. Living and thinking about it: two perspectives on life. In: Felicia A, Huppert N. Baylis, Keverne B. (Eds). The science of well-being. Oxford University Press, Oxford, pp. 285–304.

Kim, J., Kikuchi, H., Yamamoto, Y., 2013. Systematic comparison between ecological momentary assessment and day reconstruction method for fatigue and mood states in healthy adults. British Journal of Health Psychology 18, 155-167. doi: 10.1111/bjhp.12000.

Kitzman, H., Olds, D.L., Henderson, C.R., Hanks, C., Cole, R., Tatelbaum, R.,...Barnard, K., 1997. Effect of prenatal and infancy home visitation by nurses on pregnancy outcomes, child injuries and repeated childbearing. A randomised controlled trial. JAMA 278 (8), 644–652. doi: 10.1001/jama.1997.03550080054039.

Knabe, A., Rätzel, S., 2011. Scarring or scaring? The psychological impact of past unemployment risk. Economica 78 (310), 283-293. doi: 10.1111/j.1468-0335.2009.00816.x

Knabe, A., Rätzel, S., Schöb R., Weimann, J., 2010. Dissatisfied with life but having a good day: Time-use and well-being of the unemployed. The Economic Journal 120 (547), 867-889. doi: 10.1111/j.1468-0297.2009.02347.

Krueger, A., Mueller, A., 2012. Time Use, Emotional Well-Being, and Unemployment: Evidence from Longitudinal Data. American Economic Review, 102 (3), 594-99. doi: 10.1257/aer.102.3.594

Levinson, A., 2012. Valuing public goods using happiness data: The case of air quality. Journal of Public Economics 96 (9-10), 869-880. doi: 10.1016/j.jpubeco.2012.06.007

Ludbrook, J. Dudley, H., 1998. Why permutation tests are superior to t and F tests in biomedical research. The American Statistician 52, (2) 127-132. doi: 10.1080/00031305.1998.10480551

Luechinger, S., 2009. Valuing air quality using the life satisfaction approach. The Economic Journal 119 (536), 482-515. doi: 10.1111/j.1468-0297.2008.02241.x

Lyubomirsky, S., King, L. Diener, E., 2005. The benefits of frequent positive affect: Does happiness lead to success? Psychological Bulletin 131 (6), 803-8055. doi: 10.1037/0033-2909.131.6.803

Miret, M., Caballero, F.F., Mathur, A., Naidoo, N., Kowal, P., Ayuso-Mateos, J.L., Chatterji, S., 2012. Validation of a measure of subjective well-being: An abbreviated version of the Day Reconstruction Method. PLoS ONE 7(8). doi: 10.1371/journal.pone.0043887

Mitchell-Herzfeld, S., Izzo, C., Greene, R., Lee, E., Lowenfels, A., 2005. Evaluation of Healthy Families New York: First year program impacts. Office of Children and Family Services Bureau of Evaluation and Research, New York, NY.

Murray, L., Fiori-Cowley, A., Hooper, R., Cooper, P., 1996. The impact of postnatal depression and associated adversity on early mother infant interactions and later infant outcome. Child Development, 67 (5), 2512 -2526. doi: 10.1111/j.1467-8624.1996.tb01871.x

OECD, 2013. Guidelines on measuring subjective well-being. Paris: OECD. Retrieved from http://www.oecd.org/statistics/Guidelines on Measuring Subjective Well-being.pdf

Olds, D.L., 2006. The Nurse-Family Partnership: An evidence based prevention intervention. Infant Mental Health Journal 27 (1), 5-25. doi: 10.1002/imhj.20077

Palmier-Claus, J. E., Ainsworth, J., Machin, M., Barrowclough, C., Dunn, G., et al., 2012. The feasibility and validity of ambulatory self-report of psychotic symptoms using a smartphone software application. BMC psychiatry12: 172. doi:10.1186/1471-244X-12-172

Peeters, F., Berkhof, J., Delespaul, P., Rottenberg, J., Nicolson, N.A., 2006. Diurnal mood variation in major depressive disorder. Emotion 6 (3), 383-391. doi: 10.1037/1528-3542.6.3.383

Pocock, S.J., Hughes, M.D., Lee, R.J., 1987. Statistical problems in the reporting of clinical trials. New England Journal of Medicine 317 (7), 426-432.

Preparing for Life and The Northside Partnership, 2008. *Preparing for Life programme manual*. Preparing for Life and the Northside Partnership, Dublin.

Romano, J., Wolf, M., 2005. Exact and approximate stepdown methods for multiple hypothesis testing. Journal of the American Statistical Association 100, 94–108. doi: 10.1198/016214504000000539

Sanders, M.R., Markie-Dadds, C., Turner, K., 2003. Theoretical, scientific and clinical foundations of the Triple P-Positive Parenting Program: A population approach to the promotion of parenting competence. Parenting Research and Practice Monograph 1, 1-21.

Sanders, M.R., Kirby, J.N., Tellegen, C.L., Day, J.J., 2014. The Triple P-Positive Parenting Program: A systematic review and meta-analysis of a multi-level system of parenting support. Clinical Psychology Review 34 (4), 337-357. doi: 10.1016/j.cpr.2014.04.003

Steptoe, A., Gibson, E.L., Hamer, M., Wardle, J., 2007. Neuroendocrine and cardiovascular correlates of positive affect measured by ecological momentary assessment and by questionnaire. Psychoneuroendocrinology 32, 56.-64.

Stiglitz, J., Sen, A., Fitoussi, J.P., 2009. *Report by the Commission on the Measurement of Economic Performance and Social Progress.* The Commission on the Measurement of Economic Performance and Social Progress, Paris.

Stone, A.A., Mackie, C. 2013. Subjective Well-Being: Measuring Happiness, Suffering, and Other Dimensions of Experience. National Research Council, National Academies Press, Washington, DC.

Stone, A.A., Schwartz, J.E., Schkade, D., Schwarz, N., Krueger, A., Kahneman, D., 2006. A population approach to the study of emotion: Diurnal rhythms of a working day examined with the day reconstruction method. Emotion 6, 139–149. doi: 10.1037/1528-3542.6.1.139

Stone, A.A., Shiffman, S., 1994. Ecological momentary assessment in behavioural medicine. Annals of Behavioural Medicine 16 (3) 199-202.

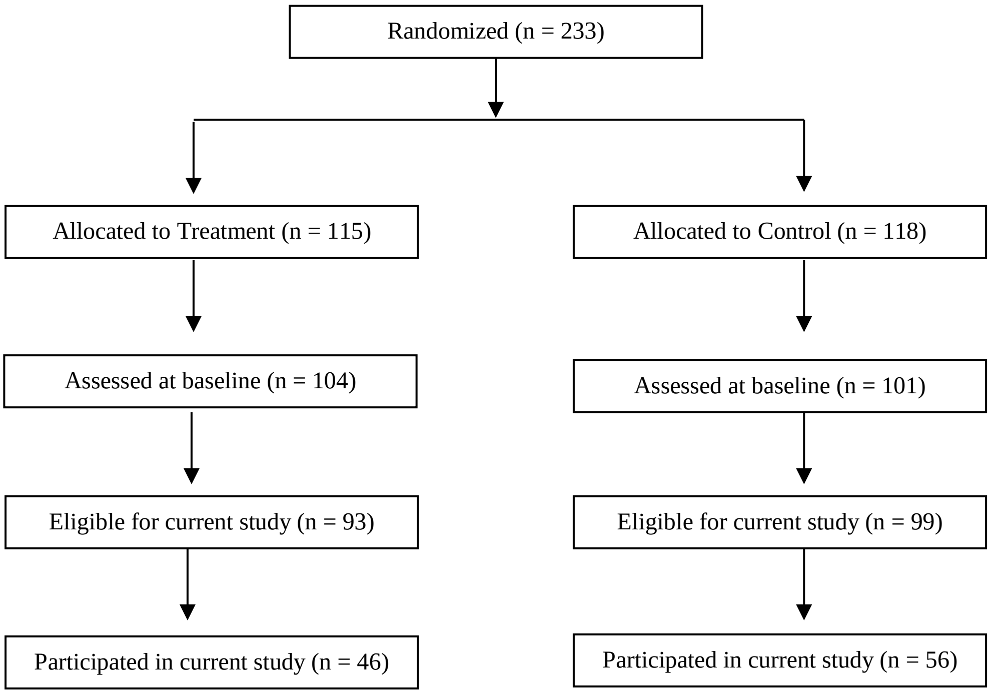
Sweet, M.A., Appelbaum, M.I., 2004. Is home visiting an effective strategy? A meta-analytic review of home visiting programs for families with young children. Child Development 75 (5), 1435-1456. doi: 10.1111/j.1467-8624.2004.00750.x

Tellegen, A., Watson, D., Clark, L., 1999. On the dimensional and hierarchical structure of affect. Psychological Science 10, 297-303, doi:10.1111/1467-9280.00157

Thompson, R.J., Mata, J. Jaeggi, SM., Buschkuehl, M., Jonides, J., Gotlib, I.H., 2012. The everyday emotional experience of adults with Major Depressive Disorder: Examining emotional instability, inertia, and reactivity. Journal of Abnormal Psychology 121 (4), 819-829. doi: 10.1037/a0027978.

Watson, D., Clark, L.A., Tellegen, A., 1988. Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology 54(6), 1063-1070. doi:10.1037/0022-3514.54.6.1063

Appendix Figure A1


Randomized (n = 233)
Allocated to Treatment (n = 115)
Allocated to Control (n = 118)
Assessed at baseline (n = 104)
Assessed at baseline (n = 101)
Eligible for current study (n = 93)
Eligible for current study (n = 99)
Participated in current study (n = 46)
Participated in current study (n = 56)

Appendix Table A1: Descriptive statistics

	Baseline Interview			
	N[a] (n_{TREAT}/ n_{CONTROL})	M_{TREAT} (*SD*)	M_{CONTROL} (*SD*)	P-value
Maternal Age	101 (46/55)	26.00 (5.45)	25.35 (5.75)	0.56
Child gender: Male	101 (46/55)	0.48 (0.51)	0.31 (0.47)	0.08*
Number of non-PFL children	101 (46/55)	1.00 (1.32)	1.05 (1.25)	0.83
First time mother	101 (46/55)	0.50 (0.51)	0.47 (0.50)	0.79
Lives in public housing	101 (46/55)	0.59 (0.50)	0.55 (0.50)	0.68
Married	101 (46/55)	0.17 (0.38)	0.16 (0.37)	0.89
Maternal Work Status				
Employed	101 (46/55)	0.39 (0.49)	0.36 (0.49)	0.78
Looking after family	101 (46/55)	0.13 (0.34)	0.13 (0.34)	0.96
Unemployed	101 (46/55)	0.43 (0.50)	0.40 (0.50)	0.73
Other	101 (46/55)	0.04 (0.21)	0.11 (0.31)	0.23
Maternal Education				
Lower than second level education	101 (46/55)	0.41 (0.50)	0.44 (0.50)	0.82
Second level education	101 (46/55)	0.20 (0.40)	0.25 (0.44)	0.49
Primary degree/non-degree qualification	101 (46/55)	0.39 (0.49)	0.31 (0.47)	0.39

Notes. **'**N' indicates the sample size. 'M' indicates the mean. 'SD' indicates the standard deviation.

[a] One participant did not complete a baseline interview, * p<0.10, ** p<0.05, *** p<0.01

Appendix Table A2: Pairwise Correlations between Well-being Measures

	Net Affect	Positive Affect	Negative Affect	U-Index	Positive Mood Yesterday	Life Satisfaction	PSI Total Stress
Net Affect	1	-	-	-	-	-	-
Positive Affect	0.85***	1	-	-	-	-	-
Negative Affect	-0.75***	-0.28***	1	-	-	-	-
U-Index	-0.71***	-0.40***	0.79***	1	-	-	-
Positive Mood	0.28***	0.22**	-0.41***	-0.25**	1	-	-
Life Satisfaction	0.13	0.03	-0.20*	-0.10	0.06	1	-
PSI Total Stress	-0.34***	-0.34***	0.20*	0.08	-0.38***	-0.19*	1

Notes: The pairwise correlations are calculated at the individual level. For Life Satisfaction the original four category variable is used to calculate the correlation coefficient rather than the two category outcome variable. * p<0.10, ** p<0.05, *** p<0.01