Human Capital and Economic Opportunity Global Working Group

Working Paper Series

Working Paper No. 2015-015

Can Early Intervention Improve Maternal Well-being? Evidence from a Randomized Controlled Trial

Orla Doyle
Liam Delaney
Christine O'Farrelly
Nick Fitzpatrick
Michael Daly

December, 2015

Human Capital and Economic Opportunity Global Working Group
Economics Research Center
University of Chicago
1126 E. 59th Street
Chicago IL 60637
www.hceconomics.org

Can Early Intervention Improve Maternal Well-being? Evidence from a Randomized Controlled Trial*

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

[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.

[d] Centre for Mental Health, Imperial College London, Commonwealth Building, Hammersmith Hospital Campus, Du Cane Road, London W12 0NN, UK.

* Acknowledgments: 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.

ABSTRACT

This study estimates the effect of a targeted policy intervention on global and experienced measures of maternal well-being. Participants from a disadvantaged community are randomly assigned during pregnancy to an intensive home visiting parenting program or a control group. The intervention has no impact on global well-being as measured by life satisfaction and parenting stress or experienced negative affect using episodic reports derived from the Day Reconstruction Method (DRM). Treatment effects are observed on measures of experienced positive affect from the DRM and a measure of mood yesterday. This suggests that early intervention may produce some improvements in experienced well-being.

Keywords: Well-Being, Randomised Controlled Trial, Early Intervention.

JEL Classification: C12, C93, I01, I39, J13

Trial Registration:

AEA RCT Registry: AEARCTR-0000066,

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

ISRCTN register (ISRCTN04631728

http://www.controlled-trials.com/ISRCTN04631728/)

I. 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 improvements in 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 limit inferences regarding causality. Randomized controlled trials are widely considered the most robust means of determining impact (Craig et al., 2008), yet few experimental policy evaluations incorporate 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 used 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 standardized 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.

As the DRM also incorporates time use data, it allows us to measure maternal 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 parental 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 utilized and, as a robustness test, we apply a stepdown procedure to mitigate the likelihood of accepting a false positive due to multiple hypothesis testing. Finally, we estimate unconditional models, in addition to conditional models which allow us to control for any baseline imbalance.

We identify an individual treatment effect on experienced reports of happiness across episodes of the study day as measured by the DRM. In almost all specifications, this applies to episodes both with and without the target child. The treatment group have similar levels of happiness during episodes with and without their target child, while the control group experience a fall in happiness during episodes without their child. In the conditional models,

we also find that the treatment group report being more relaxed during episodes without the target child. The treatment group also report higher levels of overall positive affect during episodes without the target child in the unconditional models. We also 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 statistical significance. We identify no differences between the treatment and control groups 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 II outlines the conceptual issues involved in measuring well-being and their relevance for the evaluation of early intervention programs. In Section III we provide details of the early intervention under investigation and the well-being measures employed. Section IV outlines our empirical model and statistical methods. Section V presents the results, and Section VI concludes.

II. Background and Literature

A. 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).

B. 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 such effects may exist, they are likely to be complex. For example, the impact of an early childhood program on parental employment and consumption may be ambiguous if substitution effects. The 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 aim to mediate gains for children by working directly with parents, may particularly have an impact on parental utility. The prevailing pattern, based on meta-analytic findings of primarily experimental studies, suggests that the effects of HVPs for parents 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 parental 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 to the parents foster a therapeutic alliance which acts as a pathway for promoting parental 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. This has been driven by a substantial literature demonstrating the burden that stress and depression exert 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 of mainly experimental studies 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 parental 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 experimentally designed 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).

C. Global Versus Experienced Measures of Well-being

A critical issue for evaluations of public policies, including early intervention programs, is the question of how well-being should be measured. The possibility that experienced measures of well-being 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 experienced 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, albeit non-experimental 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

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

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 yesterday. This measure 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).

III. Experimental Treatment and Methods

A. 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 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).

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 presenting 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 well-being due to the nature of the mentor-mother relationship and the help 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 took part in Group Triple P which consists of five 2-hour group discussion sessions and three individual phone calls facilitated by the mentors.

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 evaluation design has been published elsewhere (Doyle, 2013).

B. Participants

The original RCT study enrolled pregnant women from a suburban community in Dublin, Ireland, which had above national average rates of unemployment, school dropout, 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.[2] A computerised randomisation program was used, with no stratification or block techniques.

Of the original 233 participants, 192 were eligible to participate in the well-being sub-study as they had not voluntarily or involuntarily dropped out of the original study at the time of data collection.[3] Appendix Figure A1 depicts the recruitment of participants in the original trial and the present sub-study. Mothers were invited to take part in the sub-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, 101 (treatment = 46; control = 55) took part in the sub-study, 34 refused[4], 2 agreed but did not participate, and 54 could not

[2] As stated in the trial registry (www.controlled-trials.com/ISRCTN04631728/), originally 300 parents were to be included in the RCT and the quasi-experimental component - 100 parents in the randomized treatment group, 100 parents in the randomised control group, and 100 parents from a non-randomized external comparison group from another community. We oversampled the RCT group (115 treatment group and 118 control group), and recruited 99 parents for the quasi-experimental study. Participants from the quasi-experimental study are not included here as they were not invited to participate in this sub-study.

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

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

be reached by telephone, text, or letter. Many of those who could not be reached for the sub-study could also not be reached in the original study and are classified as 'disengaged'. In addition, many of those who did not participate in the sub-study did engage in further waves of data collection as part of the original study.[5] The participants were at various stages in the program when they participated in the sub-study; the youngest child was 24.6 months and the oldest child was 62.5 months old. Thus, program duration differs for each participant as data collection for the sub-study was conducted over a one year period.[6]

In order to test selection into the sub-study, we compare those who participated to those who did not on 48 baseline measures of socio-demographics, health, parenting, and psychometrics. Participants who chose to take part in the sub-study did not differ from those who did not on 96% of the baseline characteristics (46/48).[7] Significant differences on 2 (4%) measures indicated that mothers who chose to take part in this sub-study were more open [as per the Ten Item Personality Index (TIPI; Gosling et al., 2003)], and more likely to have their activity impaired by illness. This suggests that there was no systematic selection into the sub-study based on a wide range of observable characteristics.

Appendix Table A1 presents descriptive statistics on the participating sample for a selection of the baseline variables disaggregated by treatment status. 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

[5] 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, 65 completed a 24 month interview, 53 completed a 36 month interview and 47 completed a 48 month interview.

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

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

40% had not completed second level education and identified themselves as being unemployed. However, a significantly higher proportion of treatment mothers had a boy as their *PFL* target child (48%) than control mothers (31%).

A detailed analysis examining differences between the treatment and control groups who participated in the sub-study found that the groups did not differ on 92% (44/48) of baseline measures. This suggests that the randomization assumption is still valid among the sample who participated in the sub-study. Significant baseline differences on the 4 (8%) measures indicate that the treatment group were less likely to exercise at least three times per week and had lower self-efficacy scores [as per the Pearlin Self Efficacy Scale (Pearlin and Schooler, 1978)] compared to the control group. Control participants had higher emotional attachment scores [as per the Vulnerable Attachment Style Questionnaire (VASQ; Bifulco et al., 2003)] and were more likely to know multiple neighbours compared to the treatment participants.

Given the limited sample size, it is not optimal to control for all variables upon which the two groups differ, therefore, the Bayesian Information Criterion (BIC; Schwarz, 1978) is used to determine which covariates are included as controls in the conditional analysis. The BIC, which measures goodness of fit, is estimated for different combinations of baseline variables which exhibit imbalance between the groups, while accounting for the number of variables included in the model. First, the four baseline characteristics which the two groups differ upon, along with the infant's gender (where a statistical difference across the treatment and control group was also identified at birth), are included as potential controls and the BIC is calculated and stored. Next, one variable is excluded and the BIC is calculated and compared to the stored BIC. If the new BIC is more than 2 points smaller than the stored BIC (a lower BIC indicates a model with greater prediction), the new BIC is stored and the process continues by testing all possible combinations of variables until the optimal set of

baseline predictors has been identified. A similar method is adopted in Campbell et al. (2014). We also include the length of time exposed to the program at the time of data collection in the BIC model. The set of variables which result in the lowest BIC is infant gender, program duration, emotional attachment, number of neighbours known and exercise. Therefore, the only variable excluded by the BIC is self-efficacy.[8]

C. 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 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 (PSI; 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.

[8] Similar results are obtained when self-efficacy is included as a control. Results available upon request.

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

D. 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

[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* target 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.

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.

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:

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

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 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). The mood variable is created by combining the proportion of time the participants reported being in a *mildly pleasant mood* or a *very good mood*.

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). 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.[14] 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,

[14] 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.

indicating a high level of stress.[15] 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.

IV. Econometric Framework

A. 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:

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

where $I = \{1 \dots N\}$ 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:

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

[15] 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 the six control variables identified using the BIC procedure. 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 five[16] 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 five 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 IV include both conditional and unconditional permutation testing *p*-values from two-tailed tests.

B. 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

[16] The control set is composed of a participant's program duration, an emotional attachment score, the number of neighbours known by the participant, whether or not the participant exercises at least three times per week.

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.[17] The stepdown procedure is then conducted on the families where significant individual differences are identified. The outcome measures included in each family should be correlated and represent an underlying construct. However, outcomes which

[17] 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.

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.[18]

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* target child, the participant's partner, and other family members.[19] 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[20] and where they were relaxing/socializing[21], engaging in housework/cooking[22], and exercising or commuting.

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

---

[18] 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.

[19] This category includes the participants' parents, other relatives of the participant, their partners' parents and their partners' other relatives.

[20] Includes getting children dressed, feeding children, getting children ready for bed, and caring for children.

[21] Includes socialising, relaxing, on computer, talking on the phone, and watching TV.

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

V. Results

A. 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 expected, 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 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 the global and experienced measures of well-being may represent different concepts.

B. Descriptive Statistics on Affect Measures[23]

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.

---

[23] 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%).

[Insert Figure 1 here]

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).

[Insert Figure 2 here]

C. Estimation of Treatment Effects

The tables present estimates of treatment effects for experienced and global measures of maternal well-being. 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.[24] Given the few 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* target 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 time spent with children. The treatment group reports spending a higher proportion of their study day in a positive mood, relative to the control group, this difference is statistically significant in the conditional models only.

[24] For continuous outcomes t-tests refer to OLS coefficients, for binary outcomes logistic regressions are estimated and chi-squared tests are used.

In terms of the DRM measures, on average, participants in both groups report a net affect score of approximately 3 which implies that participants experience positive emotions three points more intensively on the 0-6 Likert scale than negative emotions. 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* target child. No significant treatment effects are identified for the net affect or U-Index measures.

[Insert Table 1 here]

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* target 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 in the unconditional models for overall positive affect for episodes spent without the *PFL* child. In the conditional models, the p-values are slightly larger and not statistical significant at conventional levels. The emergence of this difference during episodes without the presence of the *PFL* child is primarily driven by a decline in the control group's positive affect during those episodes, 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 the unconditional and conditional models. In 3 models, the treatment group also report higher levels of happiness during times spent with the *PFL* child. However,

the higher p value means that this effect does not reaching conventional levels of significance in the conditional permutation model. Additionally, in the conditional permutation model, the treatment group are significantly more relaxed during episodes without their child relative to the control group. The groups do not significantly differ on the remaining positive affect states.

Tests comparing positive affect states when with and without the *PFL* target 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.

[Insert Table 2 here]

Table 3 compares the treatment and control groups in terms of their negative affect and individual negative affect states for the entire day and for time 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).

[Insert Table 3 here]

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.[25]

In terms of participants' reports of parenting stress (PSI), the treatment and control groups report comparable levels and approximately 10% report clinically significant levels. 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.

[Insert Table 4 here]

D. Additional Analysis

Stepdown analysis - Table 5 presents the unconditional and conditional stepdown results for the measures upon which we identified significant differences in the individual tests. The first p-value in the conditional mood yesterday stepdown family is significant following adjustment for multiple comparisons, and is driven by the significant finding for the portion

[25] 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. In addition, when ordered logit models are calculated using the original 4-point scale, there is a statistical significance difference between the treatment and control group regarding satisfaction with life as a parent in the unconditional and conditional models. There are still no differences for satisfaction overall or satisfaction at home.

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.

[Insert Table 5 here]

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 9% 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) and a higher proportion of their episodes in work, yet less than 6% of all episodes are spent at work. There are also no differences in time use in terms of daily activities (relaxing/socializing, housework/cooking, commuting, exercising). These results suggest that the higher positive affect experienced by the treatment group may be driven by differences in the quality of the episodes rather than differences in time use.

[Insert Table 6 here]

VI. 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. Yet, 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 maternal 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. When positive DRM affect states are examined separately, we observe an individual treatment effect for happiness for the day overall and when participants are with and without the *PFL* child. Participants also report feeling more relaxed during episodes without the *PFL* child. However, these results do not survive the stepdown procedure. These findings are broadly consistent with the result concerning 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.[26] 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,

[26] The DRM and the yesterday mood question are not directly equivalent as 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.

based predominantly on deficit measures of well-being, that HVPs do not influence parents' emotional states (Brooks-Gunn and Markman, 2005).

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. This lends some supports to the finding that the treatment group feel more relaxed than the control group when without the PFL child. While there are no differences in the amount of time participants spend with their children in 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, however see also Nelson et al., 2013). 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 for the findings 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. It is also possible that gains to

maternal well-being, and happiness in particular, are accrued indirectly, via the program's impact on the children's cognitive, emotional and physical well-being (see Doyle and the *PFL* Evaluation Team, 2013). However, directionality may be obscured 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 the global assessments of well-being such as life satisfaction. 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 which recognizes 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 that the life satisfaction question should have encouraged 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. Interestingly, Eibach and Mock (2011) observed that the cost of parenthood - in this case monetary - appears to motivate parents to idealize global judgements of how rewarding parenting is. It is also possible, as discussed by Knabe and Rätzel (2011), 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 our results. 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, the treatment 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, some methodological issues should be acknowledged. The study relies on self-report measures which can be contaminated by social desirability when participants are not 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 trials is small sample size. This issue is a particular concern in the present study as the sample in the sub-study is smaller than in the original *PFL* trial. Yet we demonstrate that few factors predict selection into the sub-study, and the randomization assumption of baseline equivalence still holds in the reduced sample. In addition, 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 after adjustment.

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 may develop between positive affect and behaviors linked to successful outcomes such as higher quality relationships, superior income and productivity, greater community participation, and improved health and mortality (Lyubomirsky, King, and Diener, 2005; 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 RCTs 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 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 psychic benefit of successful program outcomes on top of the core measures being targeted.

References

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

Ammerman, R.T., F.W. Putnam, N.R. Bosse, A.R. Teeters, and J.B. Van Ginkel. 2010. "Maternal Depression in Home Visitation: A Systematic Review." *Aggression and Violent Behavior* 15 (3): 191-200.

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

Atz, U. 2013. "Evaluating Experience Sampling of Stress in a Single-subject Research Design." *Personal and Ubiquitous Computing* 17 (4): 639-652.

Bandura, A. 1977. "Self-efficacy: Toward a Unifying Theory of Behavioural Change." *Psychological Review* 84 (2): 191-215.

Bavolek, S.J., and R.G. Keene. 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., J.J. Choi, D. Laibson, and B.C. Madrian. 2008. "How Are Preferences Revealed?" *Journal of Public Economics* 92 (8-9): 1787-1794.

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

Bowen R.C., Y. Wang, L. Balbuena, A. Houmphan, and M. Baetz. 2013. "The Relationship Between Mood Instability and Depression: Implications for Studying and Treating Depression." *Medical Hypotheses* 81 (3): 459-462.

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

Brooks-Gunn, J., and L.S. Markman. 2005. "The Contribution of Parenting to Ethnic and Racial Gaps in School Readiness." *The Future of Children* 15 (1): 139-167.

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

Bylsma. L.M., A. Taylor-Clift, and J. Rottenberg. 2011. "Emotional Reactivity to Daily Events in Major and Minor Depression." *Journal of Abnormal Psychology* 120 (1): 155-167.

Campbell, F., G. Conti, J.J. Heckman, S.H. Moon, R. Pinto, E. Pungello, and Y. Pan. 2014. "Early Childhood Investments Substantially Boost Adult Health*." Science* 343 (6178): 1478-1485.

Catalino, L.I., and B.L Fredrickson. 2011. "A Tuesday in the Life of a Flourisher: The Role of Positive Emotional Reactivity in Optimal Mental Health." *Emotion* 11 (4): 938-950.

Christodolou, C., S. Schneider, and A.A. Stone. 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.

Craig, P., P. Dieppe, S. Macintyre, I. Nazareth, and M. Petticrew. 2008. "Developing and Evaluating Complex Interventions: The New Medical Research Council Guidance." *British Medical Journal* 337: 1655.

Crnic, K.A., and C. Low. 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., and J.D Henry. 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.

Daly, M., L. Delaney, P.P. Doran, C. Harmon, and M. MacLachlan. 2010. "Naturalistic Monitoring of the Affect-Heart rate Relationship: A Day Reconstruction Study." *Health Psychology* 29 (2): 186-195.

Deaton, A., and A.A Stone. 2013. "Grandpa and the Snapper: The Wellbeing of the Elderly who Live with Children." NBER Working Paper No 19100, June.

Deaton, A., and A.A Stone. 2014. "Evaluative and Hedonic Wellbeing Among Those With and Without Children at Home." *Proceedings of the National Academy of Science* 111: 1328-1333.

Diener E., and M.E.P Seligman. 2004. "Beyond Money: Toward an Economy of Well-being." *Psychological Science in the Public Interest* 5 (1): 1-30.

Diener, E., and L. Tay. 2014. "Review of the Day Reconstruction Method (DRM)." *Social Indicators Research* 116 (1): 255-267.

Dockray, S., N. Grant, A.A. Stone, D. Kahneman, J. Wardle, and A. Steptoe. 2010. "A Comparison of Affect Ratings Obtained with Ecological Momentary Assessment and the Day Reconstruction Method." *Social Indicators Research* 99 (2): 269-283.

Dolan, P., and D. Kahneman. 2008. "Interpretations of Utility and their Implications for the Valuation of Health." *The Economic Journal* 118: 215–234.

Dolan, P., R. Layard, R. Metcalfe. 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., and M.P. White. 2007. "How Can Measures of Subjective Well-being be Used to Inform Public Policy." *Perspective on Psychological Science* 2 (1): 71-85.

Dolan, P., and M.P. White. 2009. "Accounting for the Richness of Daily Activities." *Psychological Science* 20 (8): 1000-1008.

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* XLI: 92-111.

Doyle, O., L. McEntee, and K.A. McNamara. 2012. "Skills, Capabilities, and Inequalities at School Entry in a Disadvantaged Community." *European Journal of Psychology of Education* 27 (1): 133-154.

Duflo, E., R. Glennerster, and M. Kremer. 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.

Eibach, R.P., and S.E Mock. 2011. "Idealizing Parenthood to Rationalise Parental Investments." *Psychological Science*, 22: 203-208.

Elgar, F.J., P.J. McGrath, D.A. Waschbusch, S.H. Stewart, and L.J. Curtis. 2004. "Mutual Influences on Maternal Depression and Child Adjustment Problems." *Clinical Psychology Review* 24: 441-459.

Filene, J.H., J.W. Kaminski, L.A. Valle, and P. Cachat. 2013. "Components Associated with Home Visiting Program Outcomes: A Meta-analysis." *Pediatrics* 132 (2): S100-S109.

Forgeard, M.J.C., E. Jayawickreme, M.L. Kern, and M.E.P. Seligman. 2011. "Doing the Right Thing: Measuring Well-being for Public Policy." *International Journal of Well-being* 1 (1): 79-106.

Fredrickson, B.L., and T. Joiner. 2002. "Positive Emotions Trigger Upward Spirals Toward Emotional Well-being." *Psychological Science* 13 (2): 172-175.

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

Frey, B.S., and A. Stutzer. 2002. "What Can Economists Learn from Happiness Research?" *Journal of Economic Literature* 40 (2): 402-435.

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

Gosling, S.D., P.J. Rentfrow, and W.B. Swann, 2003. "A Very Brief Measure of the Big-Five Personality Domains." *Journal of Research in Personality* 37 (6): 504-528.

Gruber, J., and S. Mullainathan. 2005. "Do Cigarette Taxes Make Smokers Happier?" *Advances in Economic Analysis and Policy* 5(1): 1 – 43.

Heckman, J.J, S.H. Moon, R. Pinto, P. Savelyev, and A. Yavitz. 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.

Henquet, C., J. van Os, R. Kuepper, P. Delespaul, M. Smits, J.A. Campo, and I. Myin-Germeys. 2010. "Psychosis Reactivity to Cannabis Use in Daily Life: An Experience Sampling Study." *British Journal of Psychiatry* 196 (6): 447-453.

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

Hoffenaar, P.J., F. van Balen, and J. Hermanns. 2010. "The Impact of Having a Baby on the Level and Content of Women's Well-being." *Social Indicators Research* 97 (2): 279-295.

Kahneman, D., and A. Deaton. 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.

Kahneman, D., and A.B. Krueger. 2006. "Developments in the Measurement of Subjective Well-being." *The Journal of Economic Perspectives* 20 (1): 3-24.

Kahneman, D., A.B. Krueger, D.A. Schkade, N., Schwarz, and A.A. Stone. 2004. "A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method." *Science* 306 (5702): 1776-1780.

Kahneman, D., and J. Riis. 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., H. Kikuchi, and Y. Yamamoto. 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.

Kitzman, H., D.L. Olds, C.R. Henderson, C. Hanks, R. Cole, R. Tatelbaum,...K. Barnard. 1997. "Effect of Prenatal and Infancy Home Visitation by Nurses on Pregnancy Outcomes, Child Injuries and Repeated Childbearing. A Randomised Controlled Trial." *Journal of the American Medical Association* 278 (8): 644–652.

Knabe, A., and S. Rätzel. 2011. "Scarring or Scaring? The Psychological Impact of Past Unemployment Risk." *Economica* 78 (310): 283-293.

Knabe, A., S. Rätzel, R. Schöb, and J. Weimann. 2010. "Dissatisfied With Life But Having A Good Day: Time-use and Well-being of the Unemployed." *The Economic Journal* 120 (547): 867-889.

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

Levinson, A. 2012. "Valuing Public Goods Using Happiness Data: The Case of Air Quality." *Journal of Public Economics* 96 (9-10): 869-880.

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

Luechinger, S. 2009. "Valuing Air Quality Using the Life Satisfaction Approach." *The Economic Journal* 119 (536): 482-515.

Lyubomirsky, S., L. King, and E. Diener. 2005. "The Benefits of Frequent Positive Affect: Does Happiness Lead to Success?" *Psychological Bulletin* 131 (6): 803-8055.

Miret, M., F.F. Caballero, A. Mathur, N. Naidoo, P. Kowal, J.L. Ayuso-Mateos, and S. Chatterji. 2012. "Validation of a Measure of Subjective Well-being: An Abbreviated Version of the Day Reconstruction Method." PLoS ONE 7(8).

Mitchell-Herzfeld, S., C. Izzo, R. Greene, E. Lee, and A. Lowenfels. 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., A. Fiori-Cowley, R. Hooper, and P. Cooper. 1996. "The Impact of Postnatal Depression and Associated Adversity on Early Mother Infant Interactions and Later Infant Outcome." *Child Development* 67 (5): 2512 -2526.

Nelson, S.K., K. Kushlev, T. English., , E.W. Dunn, and S. Lyubomirsky. 2013. "In Defense of Parenthood: Children are Associated with More Joy than Misery." *Psychological Science* 24: 3-10.

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.

Palmier-Claus, J.E., J. Ainsworth, M. Machin, C. Barrowclough, G. Dunn, et al., 2012. "The Feasibility and Validity of Ambulatory Self-report of Psychotic Symptoms Using a Smartphone Software Application." *BMC Psychiatry* 12: 172.

Pearlin, L.I., and C. Schooler. 1978. "The Structure of Coping." *Journal of Health and Social Behaviour*, 19: 2-21.

Peeters, F., J. Berkhof, P. Delespaul, J. Rottenberg, and N.A. Nicolson. 2006. "Diurnal Mood Variation in Major Depressive Disorder." *Emotion* 6 (3): 383-391.

Pocock, S.J., M.D. Hughes, and R.J. Lee. 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., and M. Wolf. 2005. "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing." *Journal of the American Statistical Association* 100: 94–108.

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

Sanders, M.R., J.N. Kirby, C.L. Tellegen, and J.J Day. 2014. "The Triple P-Positive Parenting Program: A Systematic Review and Meta-analysis of a Multi-level System of Parenting Support." *Clinical Psychology Review* 34 (4): 337-357.

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

Steptoe, A., E.L. Gibson, M. Hamer, and J. Wardle. 2007. "Neuroendocrine and Cardiovascular Correlates of Positive Affect Measured by Ecological Momentary Assessment and by Questionnaire." *Psychoneuroendocrinology* 32: 56.-64.

Stiglitz, J., A. Sen, and J.P. Fitoussi. 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., and C. Mackie. 2013. "Subjective Well-Being: Measuring Happiness, Suffering, and Other Dimensions of Experience." National Research Council, National Academies Press, Washington, DC.

Stone, A.A., J.E. Schwartz, D. Schkade, N. Schwarz, A. Krueger, and D. Kahneman. 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.

Stone, A.A., and S. Shiffman. 1994. "Ecological Momentary Assessment in Behavioural Medicine." *Annals of Behavioural Medicine* 16 (3): 199-202.

Sweet, M.A., and M.I. Appelbaum. 2004. "Is Home Visiting an Effective Strategy? A Meta-analytic Review of Home Visiting Programs for Families with Young Children." *Child Development* 75 (5): 1435-1456.

Tellegen, A., D. Watson, and L. Clark. 1999. "On the Dimensional and Hierarchical Structure of Affect." *Psychological Science* 10: 297-303.

Thompson, R.J., J. Mata, S.M. Jaeggi, M. Buschkuehl, J. Jonides, and I.H. Gotlib. 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.

Watson, D., L.A. Clark, and A. Tellegen. 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.

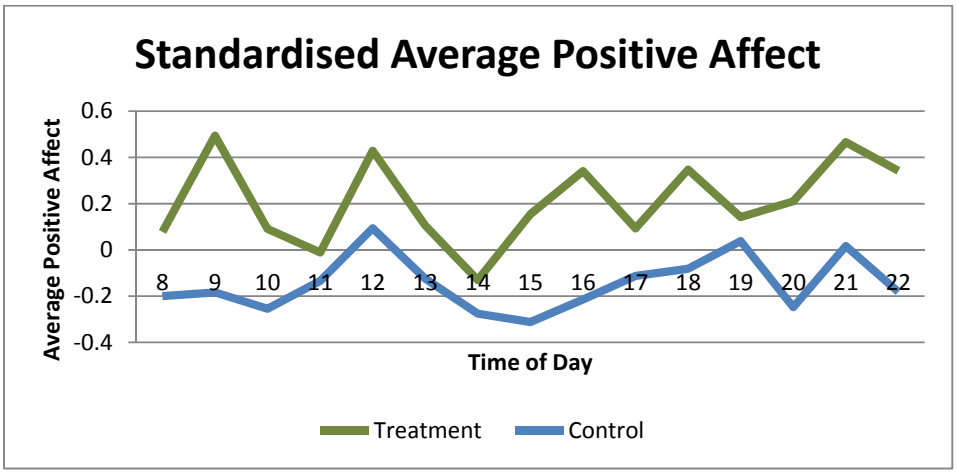


Figure 1.

Standardized average positive affect for treatment and control groups.

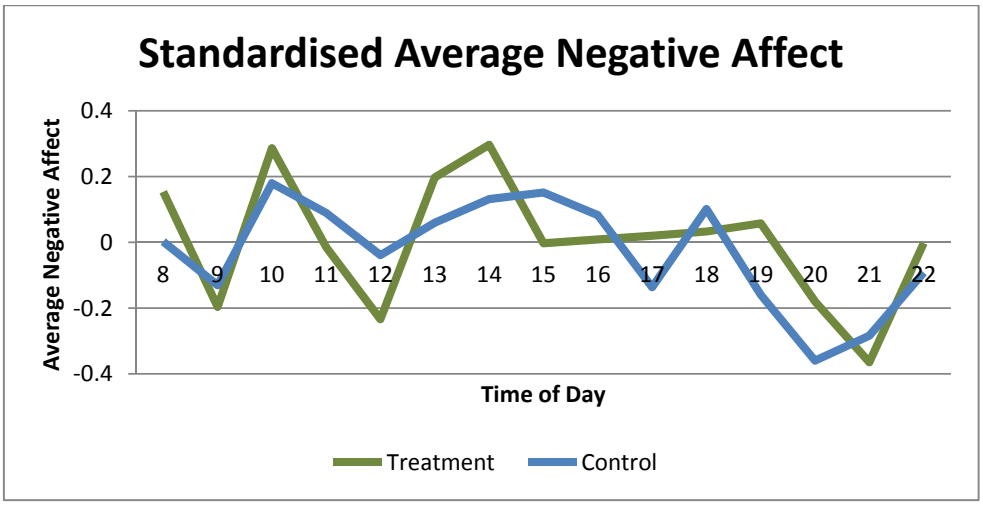


Figure 2.

Standardized average negative affect for treatment and control groups.

Table 1

Treatment Effects for Experienced Well-being: Mood Yesterday, Net Affect and U-Index

	$M_{TREATMENT}$ (*SD*)	$M_{CONTROL}$ (*SD*)	p^a	p^b	p^a	p^b
			Unconditional		Conditional	
Mood Yesterday						
Portion of day spent in a positive mood	0.76 (0.18)	0.71 (0.25)	0.321	0.308	0.047**	0.035**
Portion of day spent with children in a positive mood	0.83 (0.21)	0.84 (0.19)	0.821	0.827	0.783	0.673
Net Affect						
Net Affect	3.03 (1.41)	2.84 (1.37)	0.511	0.512	0.329	0.269
Net affect during time spent with *PFL* child	2.98 (1.58)	2.95 (1.38)	0.916	0.917	0.603	0.637
Net affect during time spent without *PFL* child	3.00 (1.78)	2.68 (1.59)	0.353	0.356	0.355	0.188
U-Index						
U-Index	0.10 (0.14)	0.09 (0.18)	0.686	0.689	0.777	0.315
U-Index during time spent with *PFL* child	0.10 (0.16)	0.08 (0.18)	0.453	0.461	0.703	0.758
U-Index during time spent without *PFL* child	0.11 (0.24)	0.12 (0.27)	0.874	0.875	0.907	0.235

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 did not record any episodes without their *PFL* child therefore n = 101 (Treatment=46, Control=50), and apart from Mood Yesterday (Treatment=45, Control=55). 'M' indicates the unconditional mean. 'SD' indicates the unconditional standard deviation. [a] two-tailed t-test p-value [b] two-tailed p-value from an individual permutation test with 100,000 replications. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table 2

Treatment Effects for Experienced Well-being: Positive Affect

	$M_{\text{TREATMENT}}$ (*SD*)	M_{CONTROL} (*SD*)	p^a	p^b	p^a	p^b
			Unconditional		Conditional	
Overall						
Positive affect	3.94	3.66	0.151	0.150	0.214	0.188
	(0.96)	(0.95)				
Positive affect during time	3.97	3.77	0.336	0.336	0.373	0.414
spent with *PFL* Child	(1.02)	(1.00)				
Positive affect during time	3.84	3.48	0.088*	0.090*	0.184	0.122
spent without *PFL* child	(1.13)	(0.92)				
Positive affect states						
Happy	4.03	3.59	0.043**	0.041**	0.064*	0.044**
	(1.00)	(1.12)				
Affectionate	3.75	3.43	0.271	0.273	0.530	0.430
	(1.49)	(1.38)				
Competent	4.40	4.18	0.324	0.320	0.402	0.408
	(1.04)	(1.12)				
In Control	4.25	4.04	0.379	0.378	0.432	0.444
	(1.16)	(1.19)				
Relaxed	3.24	3.04	0.410	0.409	0.322	0.302
	(1.16)	(1.16)				
Positive affect states during time spent with PFL child						
Happy	3.99	3.59	0.094*	0.096*	0.091*	0.108
	(1.22)	(1.17)				
Affectionate	4.25	3.98	0.340	0.341	0.562	0.588
	(1.42)	(1.40)				
Competent	4.34	4.13	0.358	0.353	0.393	0.412
	(1.09)	(1.22)				

In Control	4.25	4.13	0.607	0.607	0.756	0.761
	(1.20)	(1.17)				
Relaxed	2.94	3.00	0.834	0.836	0.910	0.960
	(1.34)	(1.21)				
Positive affect states during time spent without PFL child						
Happy	3.98	3.50	0.045**	0.045**	0.074*	0.038*
	(1.07)	(1.25)				
Affectionate	3.08	2.57	0.159	0.162	0.450	0.293
	(1.89)	(1.59)				
Competent	4.31	4.16	0.550	0.553	0.675	0.640
	(1.40)	(1.15)				
In Control	4.17	4.00	0.522	0.522	0.599	0.547
	(1.44)	(1.29)				
Relaxed	3.67	3.18	0.100	0.103	0.120	0.094*
	(1.59)	(1.27)				

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 did not record any episodes without their *PFL* child therefore n = 101 (Treatment=46, Control=50). 'M' indicates the unconditional mean. 'SD' indicates the unconditional standard deviation. [a] two-tailed t-test p-value. [b] two-tailed p-value from an individual permutation test with 100,000 replications. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table 3

Treatment Effects for Experienced Well-being: Negative Affect

	$M_{\text{TREATMENT}}$ (*SD*)	M_{CONTROL} (*SD*)	p^a	p^b	p^a	p^b
			Unconditional		Conditional	
Overall						
Negative affect	0.91 (0.79)	0.82 (0.76)	0.547	0.551	0.852	0.894
Negative affect during time spent with *PFL* child	0.98 (0.88)	0.82 (0.73)	0.309	0.323	0.852	0.658
Negative affect during time spent without *PFL* child	0.84 (0.97)	0.80 (0.92)	0.831	0.833	0.857	0.671
Negative affect states						
Stressed	1.47 (1.25)	1.24 (1.08)	0.320	0.329	0.932	0.864
Irritated	1.29 (1.12)	1.08 (1.05)	0.338	0.343	0.734	0.805
Frustrated	1.26 (1.02)	1.10 (1.00)	0.422	0.426	0.866	0.812
Angry	0.66 (0.84)	0.55 (0.84)	0.504	0.510	0.826	0.972
Impatient	1.27 (1.15)	1.32 (1.02)	0.829	0.830	0.590	0.583
Depressed	0.23 (0.37)	0.28 (0.50)	0.627	0.622	0.177	0.196
Criticized	0.18 (0.40)	0.16 (0.36)	0.781	0.786	0.444	0.526
Negative affect states during time spent with PFL child						
Stressed	1.61 (1.45)	1.25 (1.08)	0.155	0.167	0.570	0.438

Irritated	1.36	1.04	0.153	0.164	0.293	0.311
	(1.22)	(0.98)				
Frustrated	1.37	1.11	0.233	0.245	0.601	0.447
	(1.19)	(1.00)				
Angry	0.66	0.56	0.584	0.593	0.717	0.987
	(0.87)	(0.85)				
Impatient	1.43	1.36	0.783	0.787	0.854	0.910
	(1.26)	(1.09)				
Depressed	0.24	0.24	0.989	0.990	0.229	0.421
	(0.53)	(0.49)				
Criticised	0.22	0.17	0.600	0.611	0.529	0.712
	(0.49)	(0.39)				
Negative affect states during time spent without PFL child						
Stressed	1.36	1.23	0.672	0.674	0.746	0.644
	(1.61)	(1.31)				
Irritated	1.16	1.03	0.634	0.636	0.977	0.836
	(1.38)	(1.33)				
Frustrated	1.10	1.07	0.895	0.896	0.827	0.671
	(1.31)	(1.29)				
Angry	0.70	0.58	0.620	0.625	0.945	0.993
	(1.21)	(1.15)				
Impatient	1.15	1.12	0.932	0.934	0.801	0.895
	(1.46)	(1.29)				
Depressed	0.26	0.44	0.255	0.256	0.244	0.176
	(0.57)	(0.91)				
Criticised	0.14	0.13	0.922	0.929	0.871	0.728
	(0.58)	(0.34)				

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 did not record any episodes without their *PFL* child therefore n = 101 (Treatment=46, Control=50). 'M' indicates the unconditional mean. 'SD' indicates the unconditional standard deviation. [a] two-tailed t-test p-value [b] two-tailed p-value from an individual permutation test with 100,000

replications. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table 4

Treatment Effects for Global Well-being: Life Satisfaction and Parenting Stress Index.

	N ($n_{TREATMENT}$/ $n_{CONTROL}$)	$M_{TREATMENT}$ (SD)	$M_{CONTROL}$ (SD)	p^a	p^b	p^a	p^b
				Unconditional		Conditional	
Life Satisfaction							
Satisfaction with Life as a Parent	100 (45/55)	0.98 (0.15)	0.89 (0.31)	0.126	0.118	0.190	0.160
Satisfaction with Home Life	100 (45/55)	0.96 (0.21)	0.89 (0.31)	0.251	0.234	0.303	0.319
Satisfaction with Life Overall	100 (45/55)	0.93 (0.25)	0.89 (0.31)	0.465	0.477	0.650	0.704
PSI subdomains							
Parent-Child Dysfunctional Interactions	99 (45/54)	18.04 (5.44)	17.22 (5.40)	0.402	0.456	0.855	0.735
Difficult Child	94 (43/51)	22.42 (8.34)	22.18 (7.03)	0.944	0.881	0.605	0.697
Parental Distress	100 (45/55)	24.82 (8.39)	24.67 (8.50)	0.907	0.932	0.661	0.548
Total Parental Stress	93 (42/51)	64.52 (18.17)	64.02 (17.95)	0.888	0.894	0.641	0.646
Stress Cut-off	93 (42/51)	0.10 (0.30)	0.08 (0.27)	0.752	0.827	0.601	0.900
Defensive Responding	93 (42/51)	14.76 (5.24)	14.64 (5.05)	0.967	0.972	0.621	0.518
Defensive Responding Cut-off	93 (42/51)	0.24 (0.43)	0.27 (0.45)	0.731	0.694	0.980	0.945

Notes: 'N' indicates the sample size. 'M' indicates the unconditional mean. 'SD' indicates the unconditional standard deviation. [a] two-tailed t-test p-value [b] two-tailed p-value from an individual permutation test with 100,000 replications. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table 5

Stepdown Results

	Stepdown Test p^a	Stepdown Test p^b
Mood Yesterday		
Portion of Day Spent in a Positive Mood	~	0.066*
Positive affect states		
Happy	0.138	0.146
Positive affect states during time spent with PFL child		
Happy	0.294	~
Positive affect states during time spent without PFL child		
Happy	0.162	0.133
Relaxed	~	0.279

Notes: [a] two-tailed p-value from an unconditional stepdown permutation test with 100,000 replications.[b] two-tailed p-value from a conditional stepdown permutation test with 100,000 replications. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

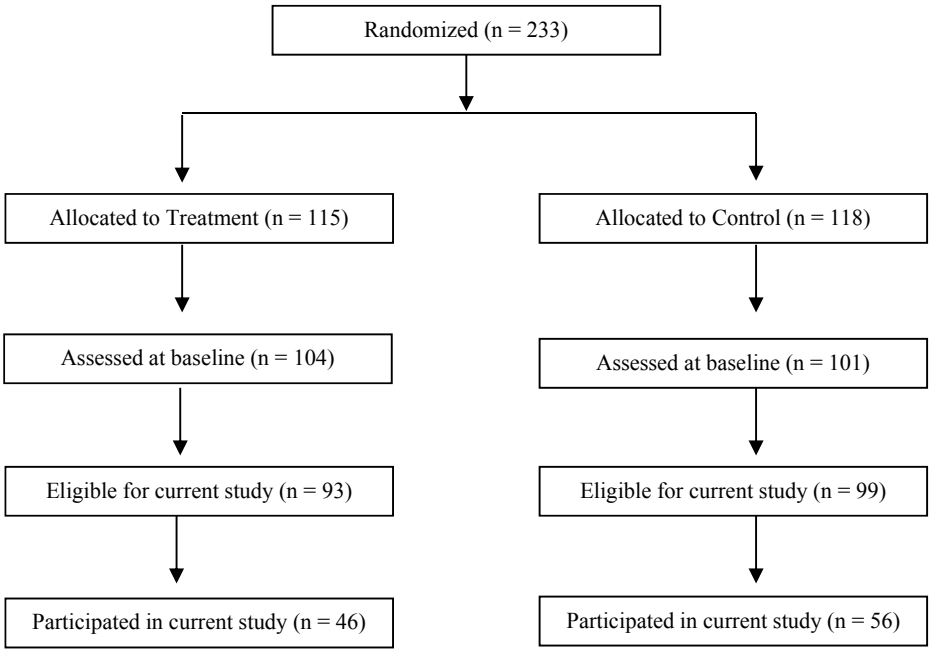
Table 6

Time Use Amongst Treatment and Control Groups

	$\%_{TREATMENT}$	$\%_{CONTROL}$	*Unconditional* p^a	*Conditional* p^b
Interaction				
With *PFL* child	61.89	66.28	0.125	0.214
With partner	16.70	22.09	0.019**	0.244
With relatives	22.99	16.45	0.008***	0.026**
Alone	9.49	10.89	0.445	0.217
Location				
At home	66.60	64.95	0.564	0.554
At work	5.89	3.16	0.029**	0.045**
Activities				
Looking after children	44.20	46.84	0.399	0.369
Playing with children	8.84	8.97	0.962	0.658
Relaxing/socializing	24.95	25.42	0.881	0.927
Housework/cooking	26.92	29.40	0.376	0.685
Commuting	12.77	13.95	0.540	0.598
Exercising	1.57	2.16	0.501	0.370

Notes: Unconditional percentages are reported**.** [a] two-tailed p-value from an individual unconditional permutation test with 100,000 replications.[b] two-tailed p-value from an individual conditional permutation test with 100,000 replications. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Appendix Figure A1



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.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

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.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.