UCD GEARY INSTITUTE
DISCUSSION PAPER SERIES

Can early intervention policies improve wellbeing? Evidence from a randomized controlled trial

Michael Daly
Behavioural Science Centre,
Stirling Management School,
Stirling University

Liam Delaney
Behavioural Science Centre,
Stirling Management School,
Stirling University
UCD School of Economics & UCD Geary Institute,
University College Dublin

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

Nick Fitzpatrick
UCD Geary Institute,
University College Dublin

Christine O'Farrelly
UCD Geary Institute,
University College Dublin

Geary WP2014/10
October, 2014

Can Early Intervention Policies Improve Well-being? Evidence from a randomized controlled trial [*]

Michael Daly[1], Liam Delaney[1, 2], Orla Doyle[2*], Nick Fitzpatrick[3], and Christine O'Farrelly[3]

[1] Behavioural Science Centre, Stirling Management School, Stirling University, FK94LA, United Kingdom.

[2] UCD School of Economics & UCD Geary Institute, University College Dublin, Belfield, Dublin 4, Ireland.

[3]UCD Geary Institute, University College Dublin, Belfield, Dublin 4, Ireland.

*Corresponding Author:

Orla Doyle,

UCD Geary Institute,

UCD, Dublin 4.

Orla.Doyle@ucd.ie

[*] This study was funded by the Irish Research Council through the Government of Ireland Collaborative Project Scheme. We would also like to thank the Northside Partnership for funding the main evaluation of the Preparing for Life programme 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 Developmental Methodology, San Diego" conference are gratefully acknowledged.

Abstract

Many authors have proposed incorporating measures of well-being into evaluations of public policy. Yet few evaluations use experimental design or examine multiple aspects of well-being, thus the causal impact of public policies on well-being is largely unknown. In this paper we examine the effect of an intensive early intervention program on maternal well-being in a targeted disadvantaged community. Using a randomized controlled trial design we estimate and compare treatment effects on global well-being using measures of life satisfaction, experienced well-being using both the Day Reconstruction Method (DRM) and a measure of mood yesterday, and also a standardized measure of parenting stress. The intervention has no significant impact on negative measures of well-being, such as experienced negative affect as measured by the DRM and global measures of well-being such as life satisfaction or a global measure of parenting stress. Significant treatment effects are observed on experienced measures of positive affect using the DRM, and a measure of mood yesterday. The DRM treatment effects are primarily concentrated during times spent without the target child which may reflect the increased effort and burden associated with additional parental investment. Our findings suggest that a maternal-focused intervention may produce meaningful improvements in experienced well-being. Incorporating measures of experienced affect may thus alter cost-benefit calculations for public policies.

Keywords: Well-Being, Randomised Controlled Trial, Early Intervention.
JEL Classification: I00, I39
This Version: 7th October 2014

1. Introduction

Understanding the impact of early intervention on the life-long development of children is an increasingly important focus of modern policymakers. One potential externality of such intervention 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 maternal utility, and indirectly, through impacting areas such as improved child health and development. Understanding how to quantify these changes in utility is essential to providing a full account of the costs and benefits of public policies.

The identification of utility effects can be hampered by evaluation design. Most evaluations of public policies are non-experimental and thus cannot infer a causal impact on utility. Randomized controlled trials are widely considered the most robust means of determining impact (Craig et al., 2008), yet few experimental policy evaluations have attempted to incorporate comprehensive measures of utility into estimates of treatment effects. Another issue concerns the measurement of utility. A large body of literature has examined the determinants of global well-being using retrospective assessments of evaluative (e.g. life satisfaction) and hedonic (e.g. happiness) well-being. Such measures are often elicited as single-item questions asking respondents to rate their well-being generally or over several weeks using ordinal scales. 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 such as early intervention programs.

In this paper, we report findings from a study designed to evaluate the utility effects of an early intervention 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 home visiting program plus group parent training or a control group that receives low level supports common to both groups. This study is the first to examine the effect of a policy intervention on common measures of experienced and global well-being using an experimental design. This distinction has been described by Kahneman as reflecting the difference between "living life" and "thinking about life" (Kahneman & Riis, 2005). In this study, global well-being is captured using measures of life satisfaction and a measure of general parenting stress

which reflects the type of measurement most frequently employed in studies of early intervention programs. Experienced well-being is captured using daily reports of positive and negative affect derived from the Day Reconstruction Method and a measure of mood yesterday. Our study provides detailed comparisons of the effect of early intervention across different global and experience based measures of well-being and draws conclusions about the welfare effects on mothers. In addition, utilising the methodology of Heckman et al. (2010), we employ permutation testing to address issues relating to the small sample size. As an additional robustness test we use a stepdown procedure to mitigate the likelihood of accepting a false positive due to multiple comparisons.

Our results indicate a treatment effect for participants' reports of experienced positive affect across episodes of the study day, yet only for time spent without the target child. The treatment group have similar levels of positive affect during episodes with and without their target child, while the control group experience a fall in positive affect during episodes when they are without their target child. Similarly, we find a treatment effect on an experienced measure of positive mood for the study day, yet not for time spent with child(ren). Consistent with the early intervention literature, there is no impact on negative aspects of well-being including both experienced negative affect and a global measure of parenting stress. In addition, while higher proportions of the treatment group compared to the control group report being satisfied with their lives across three different domains, these differences did not reach significance.

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

2. Background and Literature

2.1 Well-Being and Evaluation of Public Policy

The use of well-being measures in public policy has been widely debated in recent years (OECD, 2013). One driver of this debate is concern that purely financial measures of utility, such as employment and consumption, do not adequately capture utility, particularly in the presence of various types of bounded rationality (e.g. hyperbolic

discounting, loss aversion) and externalities (e.g. Beshears et al., 2008). Scholars from a wide range of disciplines have called for subjective 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 the causal determinants of well-being, and in particular, the specific impact of the public good being valued. 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 goods with unknown values, it has become increasingly common to pilot test provision of the good using random assignment (Duflo et al., 2008).

2.2 Maternal Welfare and Home Visiting 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). However, less work has concentrated on the effect of targeted interventions on the welfare of parents. While early intervention programs may have an impact on the economic well-being of parents, such effects are complex. For example, effects on employment and consumption measures may be ambiguous if substitution effects occur which result in a change in priorities due to the intervention. An early intervention program may potentially lead to reduced employment amongst participating parents, due to a conscious decision to spend more time with their children. Thus, measuring a parent's welfare directly may provide a more informative measure of whether their utility has been affected by the intervention.

Home visiting programs (HVPs), which are a commonly used form of early intervention that work directly with mothers, may particularly have an impact on maternal welfare. Studies that have examined this issue show effects for certain outcomes but not others. The prevailing pattern, based on meta-analytic findings, suggests that the effects of HVPs are concentrated on parenting with positive program effects identified on parenting behaviours, attitudes, and skills (Filene et al., 2013; Sweet and Appelbaum, 2004). There is also evidence, albeit less consistent, for improvements in maternal 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 maternal psychological well-being, and the direction of this effect is ambiguous. On the one hand, HVPs may improve maternal well-being if the supports delivered by the home visitor foster a therapeutic alliance which acts as a pathway for promoting well-being (Ammerman et al., 2010). Alternatively, drawing on the family investment theory (Becker, 1991), HVPs may have deleterious effects on maternal well-being if the intervention promotes substantial parental investment in the child. This would come at a cost of increased maternal time, effort, and emotional outlays in the short-run, with the expectation that such investments would increase maternal utility in the long run.

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

Comparatively fewer studies have examined the impact of HVPs on positive aspects of maternal 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, behaviour, 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 programs such as *ProKind* (Jungman et al., 2011) and the *Nurse Family Partnership* (Kitzman et al., 1997), have demonstrated positive treatment effects for self-efficacy, no effects were observed on standardized measures of self-efficacy and self-esteem employed in the *Healthy Families America* (Mitchell-Herzfeld et al., 2005), *Early Intervention Program for Adolescent Mothers* (Koniak-Griffin et al., 2002), *Parents as Teachers* (Wagner and Clayton, 1999), and the *Family Partnership Model* (Barlow et al., 2007) studies. Collectively, this evidence has led to the inference that it may be easier for HVPs to alter parenting behaviours than emotional states (Brooks-Gunn and Markman, 2005).

2.3 Global versus Experienced Measures of Well-being

A critical issue for evaluations of public policies is the question of how well-being should be measured. A large body of literature has emerged on the use of global measures of subjective well-being such as evaluations of life or domain satisfaction and retrospective accounts of happiness. Well-being research has relied heavily on such global retrospective judgements which have the strong advantage of providing information regarding the person's appraisal of their circumstances and their feelings about them; however, a large debate exists about the consistency of such evaluations. 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 et al., 2001; Kahneman & 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 such 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.

Kahneman introduced the concept of experienced utility as distinct from decision utility to capture this important difference (Dolan and Kahneman, 2008). He argues 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. The experience sampling approach is the most widely used method for capturing flows of experienced utility. This method collects 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 (e.g. Henquet et al., 2010; Bylsma et al., 2011; Peeters et al., 2006; Thompson et al., 2012; Palmeier Claus et al., 2012; Bowen et al., 2013). Kahneman et al. (2004) proposed the use of the DRM as an alternative means of recording diurnal fluctuations in experienced measures of well-being in a less burdensome manner than the experienced sampling approach. The DRM is completed in a single session during which participants divide the previous day into discrete activities or episodes which are then rated across several positive and negative emotional/affective states. The DRM has the advantage of eliciting events over an entire day without interfering with the activities of the day or placing administrative or respondent burden associated with carrying equipment to record events as required by experienced sampling. The DRM has been used in a variety of settings, including measuring time use and emotional well-being among the unemployed (Knabe et al., 2010; Krueger and Mueller, 2012), examining individuals with optimal mental health (Catalino and Fredrickson, 2011), and studying women during the transition to motherhood (Hoffenaar et al., 2010).

The possibility that experienced measures of well-being may have different determinants to global measures of well-being has been addressed in a number of studies. Knabe et al. (2011) have argued that the negative effects of unemployment may depend on whether self-reported life satisfaction measures or diurnal measures are employed 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.

Another important distinction when measuring well-being using ratings of experienced episodes, 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. An advantage of the DRM is its ability to elicit respondents' ratings of a series of episodes across their previous day on several dimensions of both positive and negative affect.

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

Although the DRM is arguably less burdensome than experience sampling, it nonetheless requires considerable participant effort (Atz, 2013). Consequently, interest has developed in less intensive measures of experienced wellbeing that are still robust to cognitive biases which affect global measures of decision utility. One proposed option is a measure of mood yesterday. This requires individuals to provide an overall appraisal of a given emotional state across the course of the study day, and thus may be a more practical alternative than DRM in large scale surveys. 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 & Mackie, 2013).

3. Experimental Treatment and Econometric Design

3.1 Experimental Set-up

Participants were randomly assigned to an intervention group receiving the *Preparing for Life* (*PFL*) HVP (*PFL* & The Northside Partnership, 2008) and the Triple P Positive Parenting Program (Sanders et al., 2003), or a control group. The treatment aims to

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

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 information, social support, access to other community services, and direct instruction on parenting practices (Howard and Brooks-Gunn, 2009). The program was developed in response to evidence that children from the catchment area were lagging behind their peers in terms of cognitive and non-cognitive skills at school entry (Doyle et al., 2012). *PFL* is a manualized program which is grounded in the theories of human attachment (Bowlby, 1969), socio-ecological development (Bronfenbrenner, 1979), and social-learning (Bandura, 1977). The trial is registered with controlled-trials.com (ISRCTN04631728).

3.1.1 Treatment

PFL 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 weekly supervision thereafter. Each family is assigned the same mentor over the course of the treatment where possible. The home visits are tailored based on the age of the child and the needs of the family and are guided by a set of Tip Sheets which present best-practice information on pregnancy, parenting, and child health and development.

This study refers to the impact of the treatment on maternal well-being and includes participants who were engaged with the program for at least two and a half years. The program is anticipated to impact maternal well-being due to the nature of the mentor-mother relationship and the supports provided. Specifically, the mentors aim to 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, during the prebirth-12 month period a Tip Sheet provides information on the prevalence and symptoms of post-natal 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 Tip Sheet on self-care delivered between 12-24 months suggests that mothers reward themselves by relaxing and doing something that makes them feel good.

The treatment group were invited to participate in the 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 and can be delivered at different levels. 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 treatment participants who availed of Triple P took part in Group Triple P which consists of five 2-hour group discussion sessions and three individual phone calls facilitated by the mentors.

3.1.2 Common Supports

While the HVP and the Triple P program is the treatment under investigation, both the treatment and control group 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. The control group also has access to a support worker who can help them avail of community services if needed, while this function is provided by the mentors for the treatment group. Further information on the program and the design of the evaluation has been published elsewhere (Doyle, 2013).

3.2 Participants

The original RCT study enrolled pregnant women from a suburban community in Dublin, Ireland, which had above national average rates of unemployment, early school leavers, lone parent households, and public housing (Doyle, 2013). All pregnant women from this community regardless of parity were eligible for voluntary participation. Recruitment took place between 2008 and 2010 through two maternity hospitals or self-referral in the community. In total, 233 participants were recruited and an unconditional probability randomization procedure assigned 115 participants to the treatment group and 118 to the control group. A computerised randomisation program was used, with no stratification or block techniques.

Of the original 233 participants, 192 were eligible to participate in the present study as they had not voluntarily or involuntarily dropped out of program and/or evaluation at the time of data collection[2]. Appendix Figure 1 depicts the recruitment of participants in the original trial and the present study.

Mothers were invited to take part in the present 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 the daily lives of parents in the *PFL* program and to learn about the different emotions parents experience during a typical day. Of the 192 target participants, 102 (treatment = 46; control = 56) took part, 34 refused[3], 2 agreed but did not participate, and 54 could not be reached by telephone, text, or letter[4]. The participants were at various stages in the program when they completed the present study; the youngest child was 24.6 months and the oldest child was 62.5 months old[5].

Participants who chose to take part do not differ from those who refused to participate on 95% of the baseline characteristics collected during pregnancy (108/114)[6]. Significant differences on 5% of measures indicated that mothers who chose to take part in the present study were somewhat more disadvantaged than those who did not participate. For example, mothers who participated reported consuming more drinks per week, availing of a greater number of certain services, being more open [as per the TIPI (Gosling et al., 2003)], having their activity impaired by illness, being in receipt of social welfare payments, and meeting the risk cutoff for lack of empathy towards their child's needs [as per the AAPI (Bavolek and Keene, 2002)].

Appendix Table 2 presents descriptive statistics on the participating sample using baseline data disaggregated by treatment status. The treatment and control mothers were largely equivalent on the majority of demographic indicators, with the exception of baby's gender. On average, mothers were between 25 and 26 years old, and had one non-*PFL* child. Approximately half of participants were first time mothers, over 55% lived in public housing, and approximately 40% had not completed a second level

[2] 32 participants (treatment = 17; control = 15) dropped out of the program and/or the evaluation and a further 9 (treatment = 6; control = 3) involuntarily chose to drop out of program due to miscarriage, death, child death, or moving out of the catchment area at the time of data collection for the present study.
[3] The leading reason for refusal was lack of time, particularly amongst working participants.
[4] Of the 92 participants who did not participate in the present study, 83 completed a baseline interview, 70 completed a 6 month interview, 66 completed a 12 month interview, 57 completed an 18 month interview and 65 completed a 24 month interview.
[5] Length of time in the program is controlled for in all analysis.
[6] Two-tailed tests were conducted, p-values <0.10 were considered significant.

education and identified themselves as being unemployed. However, a significantly higher proportion of treatment mothers had a boy as their *PFL* child (48%) than control mothers (31%). A more detailed analysis of differences between the participating treatment and control groups on 114 baseline characteristics identified that the groups did not differ on 92% (105/114) of measures. We control for three of these nine measures in all subsequent analysis (the biological father's employment status, whether or not the pregnancy was planned, and a measure of the mother's emotional attachment)[7]. In addition, we control for the infant's gender and the length of time spent by participants in the program at the time of the study interview. Program duration differs for each participant as interviews for this study were conducted within a one year period, and recruitment into the program took place over two and a half years.

3.3 Data Collection

The study procedure was approved by the institution's human research ethics committee and maternity hospitals' respective ethics committees. The survey was piloted between November 2012 and January 2013 with a convenience sample of parents (n = 5), *PFL* program staff (n = 7), and *PFL* pilot families (n = 5). Data collection commenced in February 2013 and ended in November 2013 when the target sample was exhausted. Participants were visited in their homes or a community centre (based on the participants' preference) by a researcher on two occasions over a three day period[8]. On the first day participants were given diaries and asked to record the next day's activities (study day). On the third day the interview was completed. Participants were given a €20 (~$27) voucher as a thank you for their participation.

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

[7] We do not control for the remaining six baseline differences, which include three other emotional attachment scores, two service use variables and the number of neighbours known by the participant, as they are either captured by the other control variables, or are unlikely to influence the outcome of interest.

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

3.4 Instruments

Adapted Day Reconstruction Method (DRM; Kahneman et al., 2004). The DRM was adapted for the present 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[9]. 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 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 field researcher in order to provide the most accurate breakdown of the day[13]. On average, the episodes lasted 80 minutes, and participants recorded approximately 11 episodes per day, which is in line with prior research employing the DRM (e.g. Daly et al., 2010).

The affect scores provided by each respondent can be analysed in a number of ways. Individual affect states can be examined separately across the entire day and can also be averaged to create overall positive and negative scores, known as positive and negative affect respectively. Positive and negative affect scores, as well as the individual affect states, are weighted by episode length. This means that longer episodes contribute more towards an individual's overall affect state than shorter episodes. In this study, positive and negative affect and individual affect states are considered for the entire day and for episodes where the participant is with their *PFL* child and episodes when they are not with their *PFL* child.

In order 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

[9] A copy of the diary given to participants and the appended DRM are in Appendix A.

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

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

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

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

themselves at different points along the Likert scale, interpersonal comparisons are meaningless (Kahneman and Krueger 2006). 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 states was 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:

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

As is the case for the individual affect states and the summary affect measures, the *U-Index* is weighted by episode length. The resulting score represents the proportion of time during the day where a respondent's strongest emotion was a negative one. In the present study, we compare the treatment and control groups on their *U-Index* for the entire day, and we also calculate the *U-Index* for subsets of episodes broken down by the time the participant was with and without the *PFL* child.

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

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

binary satisfaction variables (satisfied plus very satisfied versus unsatisfied plus very unsatisfied) were created.

Parenting Stress Index Short Form (PSI; Abidin, 1995).[14] Participants self-completed a paper version of the PSI (unless they requested assistance from the researcher). The PSI includes 36 items rated on a 5-point Likert scale ranging from *strongly disagree* to *strongly agree*. The scale yields a total stress score and three subscale scores: Parental Distress, Parent-Child Dysfunctional Interaction, and Difficult Child[15]. Responses were 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 was also created to represent mothers scoring above a cut-off of 90, indicating a high level of stress[16]. The PSI also contains a measure of defensive responding (Abidin, 1995) derived from the widely used Crowne-Marlowe Social Desirability Scale. These questions pertain to routine parenting experiences, a denial of these experiences 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 were computed.

4. Econometric Framework

4.1 Empirical Approach

This study adopts an intention-to-treat approach and estimates the impact of the *PFL* treatment on maternal well-being via:

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

where D_i denotes the treatment assignment for participant i (D_i = 1 for the treatment group, D_i = 0 otherwise) and $Y_i(1)$ is the potential outcome for participant i if in the

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

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

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

treatment group and $Y_i(0)$ is the potential outcome for participant i if in the control group.

The average treatment effect (ATE) is thus defined as:

$$ATE = \frac{1}{N}\sum_{n=1}^{N}(Y_i(1) - Y_i(0)\ ) \quad (2)$$

Using randomisation, the ATE is:

$$\widehat{ATE} = E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \quad (3)$$

and the relationship between Y_i and D_i can be estimated as:

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i \quad (4)$$

4.2 *Testing Procedure*

Permutation-based hypothesis testing is used to estimate equation 4. It is more suitable than standard bivariate tests, such as t-tests, 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, 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 firstly 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 a permutation test is computed by examining the proportion of permutations that have a test statistic greater than or equal to the observed statistic in the original sample. For the current study, permutation tests, based on 100,000 replications, using a regression framework, are used to estimate the program's impact on maternal well-being.

The permutation testing 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 testing 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. Heckman et al., (2010) applied this procedure to an analysis where the randomization was compromised so that the exchangeability property was not guaranteed.

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. We include five control variables.[17] Two binary variables are used to produce the orbits; the biological father's employment status and the child's gender. This method proves problematic however with many 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 three conditioning variables and the outcomes. The first linear conditioning variable reflects the amount of time spent in the *PFL* program, the second linear control variable relates to whether or not the pregnancy was planned, and the final linear control is a measure of the mother's emotional attachment.

We partition the data into orbits on the basis of the father's unemployment status and child's gender and then regress the outcome on the three variables assumed to share a linear relationship with the outcome measure. Next, the residuals are permuted from this regression within the orbits. This method is referred to as the Freedman–Lane procedure (Freedman and Lane, 1983). In a series of Monte Carlo studies, this procedure was found to be statistically sound (Anderson and Legendre, 1999).

4.3 Robustness Checks

Analysing the impact of the program on multiple well-being measures increases the likelihood of a Type-1 error and studies of RCTs have been criticized for overstating treatment effects due to this 'multiplicity' effect (Pocock et al., 1987). To address this problem and assess the robustness of our results, we employ the stepdown procedure

[17] The rational for including these particular controls is outlined in Section 3.1.

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 cluster 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 13 families for the individual permutation tests[18]. The stepdown procedure is then conducted on the families where we identify significant individual differences and the procedure can be suitably applied. The outcome measures included in each family should be correlated and represent an underlying construct. However, outcomes which are derived from the same measure should not be included in the same stepdown family. For this reason we cannot apply the stepdown procedure to all outcome measures. For example, as the measure of positive affect during times spent with the *PFL* child and the measure of positive affect during time spent without the *PFL* child, are both constructed from overall positive affect measure, it is not possible to test the joint significance of these three variables in the same stepdown family. In total, 9 of the 13 groups are suitable for stepdown analysis[19].

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

[18] 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, the U-Index, life satisfaction PSI total scores, and PSI subdomains.

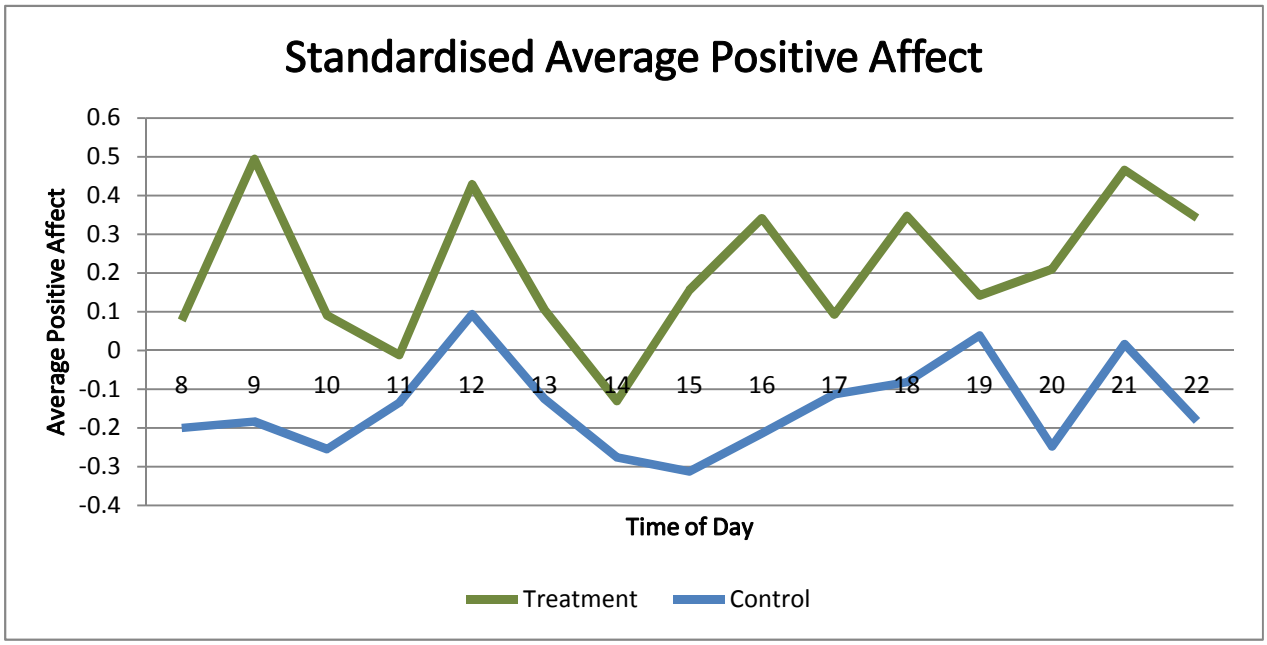
[19] The 4 groups that were ineligible for stepdown analysis were: overall positive affect, overall negative affect, the U-Index, and PSI total scores.

5. Results

5.1 Descriptive Statistics on Affect Measures[20]

For each episode, respondents report a score, on a scale of 0-6, 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 for each respondent 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 for the two groups and shows that the treatment group report higher positive affect scores at every hour, compared to the control group.



[20] In order to gauge the normality of the study day, participants were also 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 DRM took place on an a typical day. The majority of participants reported that the study day was either typical or better compared to that day of the week usually, both for the day as a whole (79%) and separately in terms of time spent with their child(ren) (83%). The majority of participants also reported that they felt less anxious on the study day compared to that day of the week usually, both for the day as a whole (57%) and separately in terms of time spent with child(ren) (88%).

Fig.1. Standardized average positive affect for treatment and control groups across the study day.

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

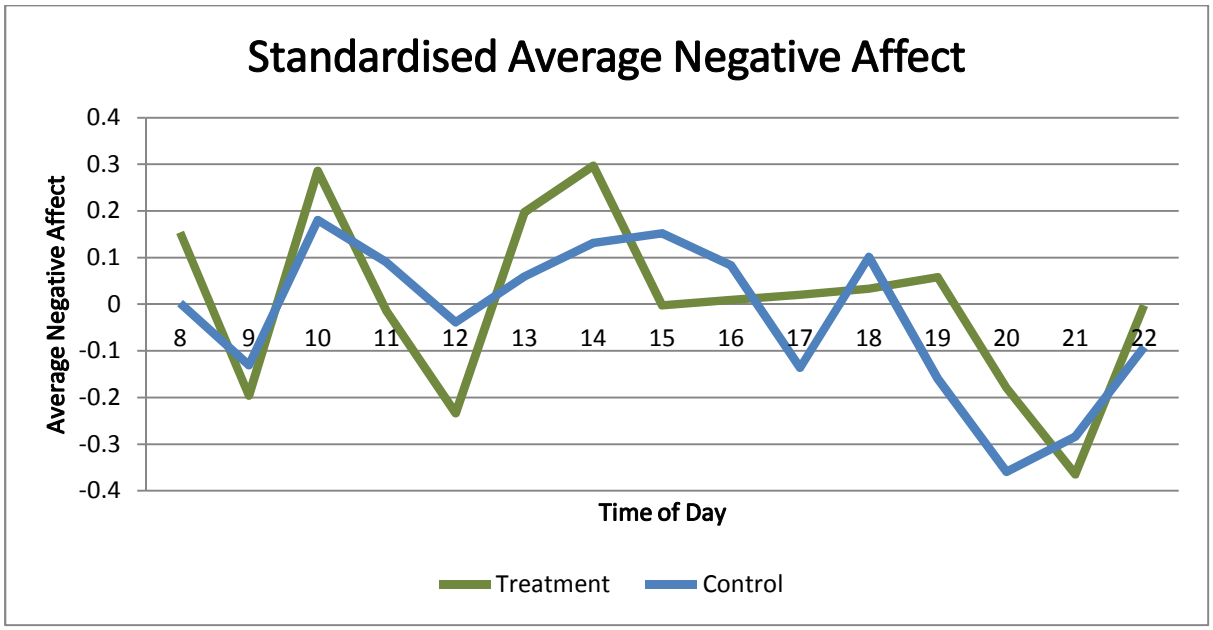


Fig.2. Standardized average negative affect for treatment and control groups across the study day.

5.2 Estimating Treatment Effects

Tables 1, 2, and 3 present estimates of treatment effects for experienced measures of positive affect, negative affect, and U-index scores. All scores are weighted by episode length and encompass all episodes recorded. Tables 4 and 5 present the results using global measures of life satisfaction and mood, and the standardized measure of parenting stress.

Table 1 compares the treatment and control groups in terms of their overall positive affect and individual positive affect states for the day as a whole and also time spent with and without the *PFL* child. Overall, feelings of competence and control receive the highest ratings across both groups, 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.

One treatment effect is identified for overall positive affect; however it is only significant for the time spent without the *PFL* child. The two groups do not significantly differ in terms of positive affect over the entire day or during episodes spent with their *PFL* child. The significant group difference is primarily driven by a decline in the control group's positive affect during episodes in which they are not with their *PFL* child, while the treatment group is 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 higher levels of happiness for the day overall and during times spent without the *PFL* child when compared with the control group. The groups do not significantly differ on the remaining four positive affect states for the day overall or the time spent with the *PFL* child, compared to the control group. However, the treatment group report feeling significantly more affectionate, competent, in control, and relaxed during time spent without the *PFL* child, compared to the control group.

Tests comparing positive affect states when with and without the PFL child (not reported) find that participants from both groups are significantly less affectionate during episodes without their *PFL* child, as we would expect, 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 they are without their *PFL* child than when they are with the *PFL* child.

The observed treatment effects for time spent without the *PFL* child may be driven by differences in time use between the two groups during the episodes in question. Yet both the treatment and the control group spend approximately the same proportion of their without *PFL* child episodes at home; 57% and 56% respectively. Both groups also spend 25% of their time socializing when they are separated from their *PFL* child. However, the control group are slightly more likely to be alone during episodes spent without their *PFL* child than the treatment group (32% versus 25%). Overall, these results suggest that time use differences may not drive the observed treatment effects.

Table 1.

Positive affect results for the treatment and control groups.

	N (n_{TREAT}/ $n_{CONTROL}$)	M_{TREAT} (*SD*)	$M_{CONTROL}$ (*SD*)	p^1
Overall				
Positive Affect	101 (46/55)	3.94 (0.96)	3.66 (0.95)	0.177
Positive Affect during time spend with *PFL* Child	101 (46/55)	3.97 (1.02)	3.77 (1.00)	0.448
Positive Affect during time spend without *PFL* child	101 (46/55)	3.84 (1.13)	3.16 (1.33)	**0.006*** **
Positive affect states				
Happy	101 (46/55)	4.03 (1.00)	3.59 (1.12)	**0.056** **
Affectionate	101 (46/55)	3.75 (1.49)	3.43 (1.38)	0.266
Competent	101 (46/55)	4.40 (1.04)	4.18 (1.12)	0.448
In Control	101 (46/55)	4.25 (1.16)	4.04 (1.19)	0.501
Relaxed	101 (46/55)	3.24 (1.16)	3.04 (1.16)	0.347
Positive affect states during time spent with PFL child				
Happy	101 (46/55)	3.99 (1.22)	3.59 (1.17)	0.114
Affectionate	101 (46/55)	4.25 (1.42)	3.98 (1.40)	0.547
Competent	101 (46/55)	4.34 (1.09)	4.13 (1.22)	0.508
In Control	101 (46/55)	4.25 (1.20)	4.13 (1.17)	0.852
Relaxed	101 (46/55)	2.94 (1.34)	3.00 (1.21)	0.995
Positive affect states during time spent without PFL child				
Happy	101 (46/55)	3.98 (1.07)	3.18 (1.56)	**0.005*** **
Affectionate	101 (46/55)	3.08 (1.89)	2.34 (1.69)	**0.020** **
Competent	101 (46/55)	4.31 (1.40)	3.78 (1.63)	**0.072** **
In Control	101 (46/55)	4.17 (1.44)	3.63 (1.69)	**0.067** **
Relaxed	101 (46/55)	3.67 (1.59)	2.89 (1.53)	**0.011*** **

Notes: 'N' indicates the sample size. 'M' indicates the mean. 'SD' indicates the standard deviation. [1] two-tailed p-value from an individual permutation test with 100,000 replications. ** p < .05, *** p < .01

Table 2 compares the treatment and control groups in terms of their negative affect and individual negative affect states for the entire day and the time participants spent with and without their *PFL* child. No significant treatment effects are identified. 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 than the other negative affect states, with depressed and criticised receiving the lowest ratings. Overall, ratings of negative affect states seem to be slightly less intense when participants were not with their *PFL* child, although none of these differences are significant for either group (not reported).

Table 2.

Negative affect results for the treatment and control groups.

Negative Affect	N (n_{TREAT}/ $n_{CONTROL}$)	M_{TREAT} (*SD*)	$M_{CONTROL}$ (*SD*)	p^1
Overall				
Negative Affect	101 (46/55)	0.91 (0.79)	0.82 (0.76)	0.982
Negative Affect during time spent with *PFL* child	101 (46/55)	0.98 (0.88)	0.82 (0.73)	0.588
Negative Affect during time spent without *PFL* child	101 (46/55)	0.84 (0.97)	0.73 (0.91)	0.862
Negative affect states				
Stressed	101 (46/55)	1.47 (1.25)	1.24 (1.08)	0.742
Irritated	101 (46/55)	1.29 (1.12)	1.08 (1.05)	0.803
Frustrated	101 (46/55)	1.26 (1.02)	1.10 (1.00)	0.885
Angry	101 (46/55)	0.66 (0.84)	0.55 (0.84)	0.889
Impatient	101 (46/55)	1.27 (1.15)	1.32 (1.02)	0.559
Depressed	101 (46/55)	0.23 (0.37)	0.28 (0.50)	0.466
Criticized	101 (46/55)	0.18 (0.40)	0.16 (0.36)	0.968
Negative affect states during time spent with PFL child				
Stressed	101 (46/55)	1.61 (1.45)	1.25 (1.08)	0.409
Irritated	101 (46/55)	1.36 (1.22)	1.04 (0.98)	0.336
Frustrated	101 (46/55)	1.37 (1.19)	1.11 (1.00)	0.479

Angry	101 (46/55)	0.66 (0.87)	0.56 (0.85)	0.819
Impatient	101 (46/55)	1.43 (1.26)	1.36 (1.09)	0.992
Depressed	101 (46/55)	0.24 (0.53)	0.24 (0.49)	0.725
Criticised	101 (46/55)	0.22 (0.49)	0.17 (0.39)	0.729
Negative affect states during time spent without PFL child				
Stressed	101 (46/55)	1.36 (1.61)	1.12 (1.30)	0.936
Irritated	101 (46/55)	1.16 (1.38)	0.94 (1.30)	0.986
Frustrated	101 (46/55)	1.10 (1.31)	0.97 (1.27)	0.807
Angry	101 (46/55)	0.70 (1.21)	0.53 (1.11)	0.912
Impatient	101 (46/55)	1.15 (1.46)	1.02 (1.27)	0.835
Depressed	101 (46/55)	0.26 (0.57)	0.40 (0.88)	0.340
Criticised	101 (46/55)	0.14 (0.58)	0.12 (0.33)	0.864

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

Table 3 compares the treatment and control groups in terms of their U-index scores across the day as a whole and the time spent with and without the *PFL* child and no significant treatment effects are found. Both groups spend approximately 10% of their day in an unpleasant state and this is broadly consistent across time spent with and without the *PFL* child.

Table 3.

U-Index results for the treatment and control groups.

	N (n_{TREAT}/ $n_{CONTROL}$)	M_{TREAT} (*SD*)	$M_{CONTROL}$ (*SD*)	p^1
Overall				
U-Index	101 (46/55)	0.10 (0.14)	0.09 (0.18)	0.965
U-Index during time spend with *PFL* Child	101 (46/55)	0.10 (0.16)	0.08 (0.18)	0.506
U-Index during time spend without *PFL* Child	101 (46/55)	0.11 (0.24)	0.11 (0.26)	0.582

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

Table 4 presents estimates of treatment effects for the measures of mood yesterday and life satisfaction questions. It shows that both groups report that they spent approximately three-quarters of the study day in a positive mood. This increases to four-fifths when participants restricted their judgements to the time spent with children. Furthermore, the treatment group reports spending a significantly higher proportion of the study day in a positive mood than the control group. In 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. Note that only 9 participants across both groups report being either *unsatisfied* or *very unsatisfied* with their life overall compared to 91 reporting being *satisfied* or *very satisfied* (the comparable figures for satisfaction as a parent and satisfaction with home life are 7 and 8 respectively), thus the small cell size in the binary variables should be noted when interpreting the results.

Table 4.

Measures of mood yesterday mood and life satisfaction results for the treatment and control groups.

	N (n_{TREAT}/ $n_{CONTROL}$)	M_{TREAT} (SD)	$M_{CONTROL}$ (SD)	p^1
Mood				
Portion of Day Spent in a Positive Mood	99 (45/54)	0.76 (0.18)	0.71 (0.25)	**0.036****
Portion of Time Spent with Children in a Positive Mood	101 (46/55)	0.83 (0.21)	0.84 (0.19)	0.867
Life Satisfaction				
Satisfaction with Life as a Parent	100 (45/55)	0.98 (0.15)	0.89 (0.31)	0.167
Satisfaction with Home Life	100 (45/55)	0.96 (0.21)	0.89 (0.31)	0.400
Satisfaction with Life Overall	100 (45/55)	0.93 (0.25)	0.89 (0.31)	0.662

Notes: 'N' indicates the sample size. 'M' indicates the mean. 'SD' indicates the standard deviation. [1] two-tailed p-value from an individual permutation test with 100,000 replications, ** $p < .05$

Finally, Table 5 presents estimates of treatment effects for participants' reports of parenting stress (PSI). It shows that the treatment and control groups report comparable levels of parenting stress and approximately 10% of participants in both groups report stress levels that are considered to be clinically significant. However, there are no significant treatment effects for any of the five PSI scores.

Table 5.

Parenting stress index results for treatment and control groups.

	N (n_{TREAT}/ $n_{CONTROL}$)	M_{TREAT} (*SD*)	$M_{CONTROL}$ (*SD*)	p^1
PSI subdomains				
*Parent-Child Dysfunctional Interactions	99 (45/54)	18.04 (5.44)	17.23 (5.40)	0.575
*Difficult Child	94 (43/51)	22.42 (8.34)	22.18 (7.03)	0.850
*Parental Distress	100 (45/55)	24.82 (8.39)	24.67 (8.50)	0.656
*Total Parental Stress	93 (42/51)	64.52 (18.17)	64.02 (17.95)	0.850
*Stress Cut-off	93 (42/51)	0.10 (0.30)	0.08 (0.27)	0.739
Defensive Responding	93 (42/51)	14.76 (5.24)	14.64 (5.05)	0.712
Defensive Responding Cut-off	93 (42/51)	0.24 (0.43)	0.27 (0.45)	0.950

Notes: 'N' indicates the sample size. 'M' indicates the mean. 'SD' indicates the standard deviation. [1] two-tailed p-value from an individual permutation test with 100,000 replications. * indicates the variable was reverse coded for the testing procedure.

Table 5 also shows that 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.

5.3 Robustness Checks

Table 6 presents stepdown results for the measures upon which we identified significant differences according to the individual tests in Tables 1-5. The variables

within each stepdown family are ordered by relative magnitude within their respective family of outcomes. The first outcome in a group has the largest t-statistic and is the first variable to be dropped as we stepdown through the hypotheses.

Table 6 shows that the two groups do not have significantly different levels of positive affect states for the day as a whole when the stepdown procedure is applied. In contrast, the positive affect states during time spent without *PFL* child stepdown family does survive adjustment for multiple comparisons. The first p-value in this category (Happy) is the result of jointly testing all 5 outcomes in the without *PFL* child stepdown family. The observed significant stepdown p-value is driven by the five individual significant findings. The next adjusted p-value (Relaxed) is the result of excluding the happy variable from the joint hypothesis test and testing the remaining 4 positive affect states collectively. We continue to stepdown through the outcomes in this family until only one measure remains (in this case Competent). The stepdown p-value for this last measure is the same as the individual test p-value for that measure in Table 1. The first p-value in the mood stepdown family is also significant following adjustment for multiple comparisons, and is driven by the significant individual finding for the portion of day spent in a positive mood.

Table 6.

Stepdown results for significant group differences in positive affect and mood.

	Stepdown Test p^2
Positive affect states	
Happy	0.186
In Control	0.501
Competent	0.567
Relaxed	0.608
Affectionate	0.608
Positive affect states during time spent without PFL child	
Happy	**0.016****
Relaxed	**0.033****
Affectionate	**0.041****
Competent	**0.072***
In Control	**0.094***
Mood	
Portion of Day Spent in a Positive Mood[1]	**0.072***
Portion of Time Spent with Children in a	0.867

Positive Mood[2]

Notes: [1] two-tailed p-value from a stepdown permutation test with100,000 replications, * p < .10, ** p < .05.

6. Conclusion

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

This paper examines the utility effects of an early intervention program using multiple measures of well-being. We find that participants who receive the *PFL* intervention report higher levels of experienced positive affect using a Day Reconstruction Method than the control group, for times when participants are without their study child. This result is broadly consistent with participants' global judgments for their overall levels of positive mood, where we observe a significant treatment effect for the study day, yet not during times spent with children.[21] Interestingly, when individual positive DRM affect states are examined, we observe a treatment effect for happiness for the day overall, however this result does not survive the stepdown procedure. There are no treatment effects for mothers' negative well-being irrespective of measurement including overall experienced negative affect, individual negative affect states, U-index scores which measure time spent in an unpleasant state, and general ratings of parenting stress as measured by a standardized instrument. 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 program 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

[21] Note that the DRM and the global mood question are not directly equivalent given that the DRM is broken down by time spent with and without *PFL* child, whereas the global mood question was asked for the day as a whole and with any of the participants' children. This limits our ability to make direct comparisons across the two measures.

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 can have an impact on positive affect, thus, contradicting the prevailing assumption, based predominantly on deficit measures of well-being, that HVPs do not influence parents' emotional states (Brooks-Gunn and Markman, 2005).

Understanding why the intervention has an impact on affect states during times spent without the study child may be linked to the family investment theory. The intervention aims to heighten parents' awareness of the importance of being actively engaged when interacting with their child. If such investment confers an increased effort and burden on the parents in the short-run, treatment mothers may particularly value times when they are not actively being a parent. While there are no differences in the amount of time participants spend with their children in 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 (Kahneman et al., 2004; White and Dolan 2009). 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 the activity.

A second related pathway is that the intervention, through Tip Sheets and mentor support, encourages mothers to use their non-parenting time for self-care, relaxation, and social relationships. These supports may result in positive emotional experiences as rich social relationships are integral to optimising happiness (Diener and Seligman, 2002), and socialising and relaxing typically receive the highest ratings of experienced positive affect on the DRM (Kahneman et al., 2004). Yet, this explanation is less likely given that time use between the groups appears broadly similar, although it is possible that the quality of these experiences differ in some unobserved way.

Another key question concerns why the intervention generates treatment effects for daily experiences of well-being, including experienced affect and assessments of

yesterday's mood, but not more evaluative assessments of well-being such as life satisfaction[22]. 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 yesterday's mood measures (see Stone & Mackie, 2013). Another hypothesis is that global and experienced well-being are independent constructs, as is reflected in the recent conceptual shift to recognize experienced well-being and global/evaluative well-being as distinct psychological phenomena (Diener and Tay, 2014; Kahneman et al., 2010). Applied to our study, the absence of treatment effects for global well-being may be considered counterintuitive if we believe the question should have encouraged participants to focus on their participation in the program, its association with greater parenting competency, and anticipation of future benefits – as part of participants' appraisals of their general life circumstances. 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 evaluative aspects of well-being by providing individuals with a sense of purpose, connection, and contribution to personal goals. Another potential reason for this finding, discussed by Knabe and Rätzel (2011), is that participants habituate quickly to their circumstances - in this case treatment status - and thus the effects on global well-being may dissipate over time.

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

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

the generally lower levels of well-being among women living in disadvantaged communities (Ammerman et al., 2010), this treatment effect is positive from both an absolute and relative perspective. While further research is needed to benchmark these effects against causal estimates of income and other policy-relevant variables, these suggest relatively large positive well-being effects.[23]

While this study is the first to our knowledge to elucidate the causal impact of a public policy on experienced affect, a number of methodological issues should be acknowledged. A common criticism of experimental trials is the use of self-report measures, which can be contaminated by social desirability when participants cannot be blinded to their treatment status. Subjective well-being, by definition, demands self-report. However, our results show that there are no systematic differences in social desirability between the treatment and control groups according to the defensive responding validity measure embedded within the PSI.

An additional issue which is common to many experimental trials is small sample size. This issue is a particular concern in the present study as the sample is smaller and relatively more disadvantaged than the sample in the original *PFL* trial. The permutation testing method helps to address this issue and is conditional on salient group differences. A further issue frequently associated with studies of HVP, is the risk of overstating the program's impact due to multiple hypothesis testing. This is addressed in the present study by the stepdown procedure, which highlights the significance of failing to account for this issue.

Furthermore, increased socioeconomic risk is often a prohibitive factor for recruitment (Korfmacher et al., 2008) and is associated with lower maternal well-being (Kaplan et al., 1987). In this way our results demonstrate that treatment effects extend to trial participants who may be most in need of support. It is also important to note that at the time of data collection, participants had received various levels of treatment, which precludes our ability to test the effects of the full *PFL* treatment on well-being.

If the identified treatment effect for experienced positive affect is valid, this could 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 & Joiner, 2002). Over time a causal relationship is believed to develop between positive affect and behaviors linked to more successful

[23] See also Krueger (ed) 2009 for within-person comparisons of the effect of being in different situations.

outcomes such as higher quality relationships, superior income and productivity, greater community participation, and improved health and mortality (Lyubomirsky, King, & Diener, 2005). Thus, the treatment effects identified here may have important implications for the cost-benefit analysis of the *PFL* program and similar HVPs in the future.

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

References

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

Ammerman, R.T., Putnam, F.W., Bosse, N. R., Teeters, A. R., Van Ginkel, J. B. 2009. Maternal depression in home visitation: A systematic review. Aggression and Violent Behavior 15 (3), 191-200. doi: 10.1016/j.avb.2009.12.002

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

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

Barlow, J., Davis, H., McIntosh, E., Jarrett, P., Mockford, C., Stewart-Brown, 2007. Role of home visiting in improving parenting and health in families at risk of abuse and neglect: Results of a multicentre random controlled trial and economic evaluation. Archives of Disease in Childhood 92 (3), 229-233. doi: 10.1136/adc.2006.095117

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Dockray, S., Grant, N., Stone, A.A., Kahneman, D., Wardle, J., Steptoe, A., 2010. A comparison of affect ratings obtained with Ecological Momentary Assessment and the Day Reconstruction Method. Social Indicators Research 99 (2), 269-283. doi: 10.1007/s11205-010-9578-7

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Gruber, J., and Mullainathan, S., 2005. Do Cigarette Taxes Make Smokers Happier? Advances in Economic Analysis and Policy 5(1), 1 – 43. doi: 10.2202/1538-0637.1412

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

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

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

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

Jungmann, T., Kurtz, V., Brand, T., von Klitzing, K, Sierau, S., Lutz, P., Sandner, M. (2011). International Meeting on Nurse Family Partnership. Implementation and Evaluation Research: Results of the German Pilot Project, Pro Kind. [Lecture] June 23/24 2011. Amsterdam.

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

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

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

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

Kaplan, G. A., Robert, R., Camacho, T., Coyne, J. C., 1987. Psychosocial predictors of depression: Prospective evidence from the human population laboratory studies. American Journal of Epidemiology, 125 (2), 206-220.

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

Kitzman, H., Olds, D.L., Henderson, C.R., Hanks, C., Cole, R., Tatelbaum, R.,...Barnard, K., 1997. Effect of prenatal and infancy home visitation by nurses on pregnancy outcomes,

child injuries and repeated childbearing. A randomised controlled trial. JAMA 278 (8), 644–652. doi: 10.1001/jama.1997.03550080054039.

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

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

Koniak-Griffin, D., Anderson, N. L., Brecht, M.-L., Verzemniekes, I., Lesser, J., Kim, S. 2002. Public health nursing care for adolescent mothers: Impact on infant health and selected maternal outcomes and 1 year postbirth. Journal of Adolescent Health 30 (1), 44-54. doi: 10.1016/S1054-139X(01)00330-5

Kormacher, J., Green, B. Staerkel, F., Peterson, C., Cook, G., Roggman, L., Faldowski, R. A., Schiffman, R. 2008. Parent involvement in early childhood home visiting. Child and Youth Care Forum, 37 (4), 171-196. doi: 10.1007/s10566-008-9057-3

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

Krueger, A.B., ed. (2009) Measuring the Subjective Well-Being of Nations: National Accounts of Time Use and Well-Being (Chicago: University of Chicago Press)

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

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

Luechinger, S. 2009. Valuing Air Quality Using the Life Satisfaction Approach. Economic Journal 119 (536), 482-515. doi: 10.1111/j.1468-0297.2008.02241.x

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

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

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

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

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

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

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

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

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

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

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

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

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

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

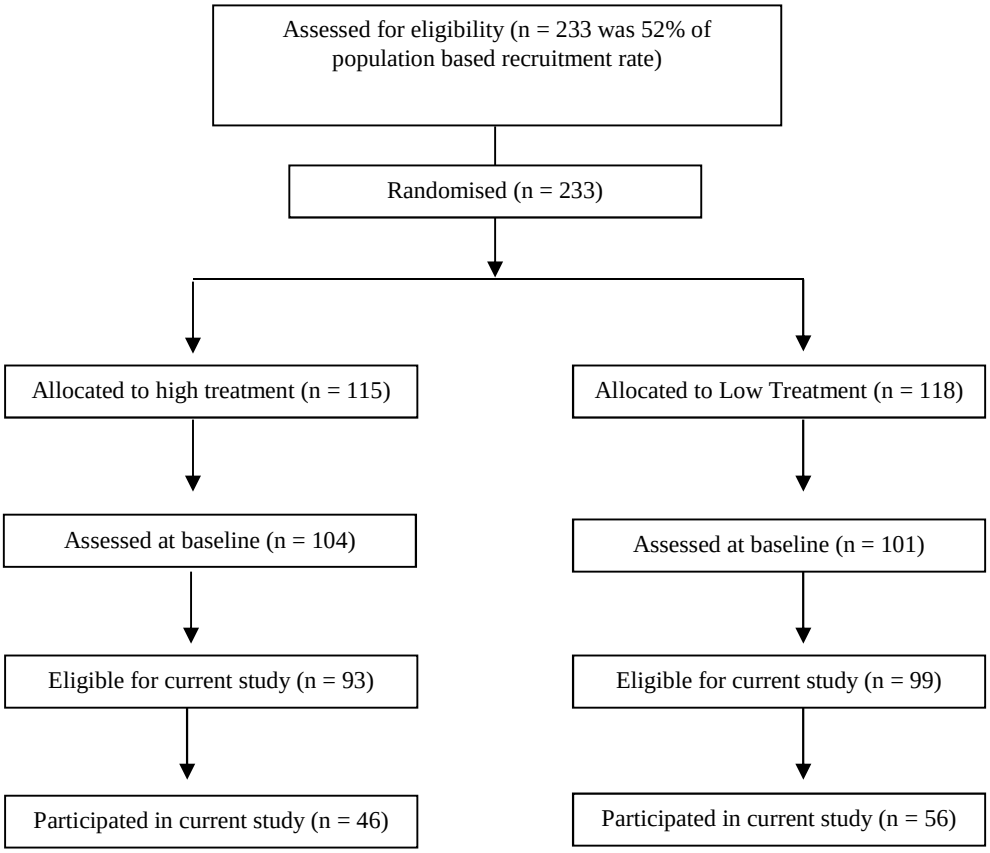
Stone, A, A., Shiffman, S. 1994. Ecological momentary assessment in behavioural medicine. Annals of Behavioural Medicine. 16 (3) 199-202.

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

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

Wagner, M. M., Clayton, S. L. 1999. The Parents as Teachers Program: Results from two demonstrations. The Future of Children, 9 (1), 91-115. 10.2307/1602723

Appendix Figure 1



Appendix Table 1: Descriptive statistics regarding participants' characteristics

	Baseline Interview			
	N [a] (n_{TREAT}/ $n_{CONTROL}$)	M_{TREAT} (*SD*)	$M_{CONTROL}$ (*SD*)	P-Value
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
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
Education				
Lower than second level education	101 (46/55)	0.41 (0.50)	0.44 (0.50)	0.82
Second level education	101 (46/55)	0.20 (0.40)	0.25 (0.44)	0.49
Primary degree/non-degree qualification	101 (46/55)	0.39 (0.49)	0.31 (0.47)	0.39

Notes. **'**N' indicates the sample size. 'M' indicates the mean. 'SD' indicates the standard deviation.
[a] One participant did not complete a baseline interview, $p < .05$

Appendix A: Survey Instrument

Preparing For Life

Northside Partnership & UCD Geary Institute
"A Day in the Life of a Parent" Study

Day Reconstruction Method

Diary Pages

On the next three pages, please describe yesterday. Think of your day as a continuous series of scenes or episodes in a film. Give each episode a brief name that will help you remember it (for example, “bringing child to school”, or “at lunch with B”, where B is a person or a group of people). Write down the approximate times at which each episode began and ended. The episodes usually last between 15 minutes and 2 hours, but this is just a guideline. The end of an episode might be going to a different location, ending one activity and starting another, or a change in the people you interacted with.

There is one page for each part of the day – Morning (from waking up until just before lunchtime), Afternoon (from lunchtime to just before dinner) and Evening (from dinner until you went to bed). There is room to list 10 episodes for each part of the day, although you may not need that many, depending on your day. It is not necessary to fill up all of the spaces – use the breakdown of your day that makes the most sense to you and best captures what you did and how you felt. Try to remember each episode in detail, and write a few words that will remind you of exactly what was going on. Also, try to remember how you felt, and what your mood was like during each episode. What you write down only has to make sense to you, and to help you remember what happened when you are answering the questions in Section 3.

Morning

This covers the time from when you woke up until just before lunchtime. Remember you don't have to fill in all ten episodes – just however many you need.

Episode Number:	Time it began:	Time it ended:	Notes to yourself: What happened? How did you feel?
1M			
2M			
3M			
4M			
5M			
6M			
7M			
8M			
9M			
10M			

Afternoon

This covers the time from lunch until just before dinner.

Episode Number:	Time it began:	Time it ended:	Notes to yourself: What happened? How did you feel?
1A			
2A			
3A			
4A			
5A			
6A			
7A			
8A			
9A			
10A			

Evening

This covers the time from when you had dinner until just before you went to sleep.

Please look over your diary in **Section 2** once more. Are there any other episodes that you would like to revise or add more notes to? Is there an episode that you would want to break up into two parts? If so, please go back and make the necessary changes. When you are happy with your diary, please let the researcher know and we will continue with **Section 3**.

Episode Number:	Time it began:	Time it ended:	Notes to yourself: What happened? How did you feel?
1E			
2E			
3E			
4E			
5E			
6E			
7E			
8E			
9E			
10E			

DRM Survey

Section 1: General

First we would like to ask you some general questions about your life.

Taking all things together, how satisfied are you with your life as a whole these days?			
Very unsatisfied	Unsatisfied	Satisfied	Very Satisfied
How satisfied are you with your life at home?			
Very unsatisfied	Unsatisfied	Satisfied	Very Satisfied
How satisfied are you with your life as a parent?			
Very unsatisfied	Unsatisfied	Satisfied	Very Satisfied

Please answer these questions by giving the answer that best describes how you feel.

Section 2: Yesterday

We would like to learn what you did and how you felt yesterday. Not all days are the same – some are better, some are worse and others are pretty typical. Here we are only asking you about **yesterday**.

Because many people find it difficult to remember what exactly they did yesterday, we will do this in three steps. First of all, please tell us a little about yesterday:

What day was it yesterday?	
What time (approximately) did you wake up at yesterday?	__:__
What time (approximately) did you go to sleep?	__:__

We would like you to write down what your day was like during this time, as if you were writing in your diary. Where were you during the day? What did you do and how did you feel? Answering these questions on the next page will help you to break down your day.

This section is just for you, to help you remember and describe what happened yesterday. It is yours to keep, so your notes are strictly personal and confidential. You do not need to give it to us.

After you have finished writing about your day in this section, we will move on to **Section 3**. In Section 3 we will ask you specific questions about yesterday. In answering these questions we would like you to look at your diary page and the notes you made to remind you of what you did and how you felt.

Section 3: How did you feel yesterday?

How many episodes did you record for the morning?	
How many episodes did you record for the afternoon?	
How many episodes did you record for the evening?	

Before we move on, please look back at your diary pages.

Now, we would like to learn in more detail about **how you felt** during those episodes. For each episode, there are several questions about what you were doing and how you felt. Please use the notes on your diary pages as often as you need to. Please answer the questions for every episode you recorded, beginning with the first episode in the Morning. Each episode is numbered - for example, the first episode of the Morning is number 1M, the third episode of the Afternoon is number 3A, the second episode of the Evening is number 2E, and so forth. It is very important that we get to hear about all of the episodes you experienced yesterday, so please be sure to answer the questions for each episode you recorded. After you have answered the questions for all of your episodes, including the last episode of the day (just before you went to bed), we will go on to Section 4.

First Morning Episode:

Please look at your Diary and select the <u>earliest</u> episode you noted in the Morning.
When did this first episode begin and end (e.g., 7:30am)? Please try to remember the times as precisely as you can.

This is episode number _____, which began at _____ and ended at _____

What were you doing? (please tick <u>all</u> that apply):

grooming/self care		exercising (alone/group)		Other(please specify):
getting child(ren) dressed		attending training (paid/unpaid)		
feeding your child(ren)		paid work		
eating		taking care of your child(ren)		
commuting		playing with your child(ren)		
doing housework		putting child(ren) to bed		
shopping		computer/internet/email (home)		
preparing food		on the phone/skype		
socialising		watching TV		
relaxing		sleeping		

Where were you? (please tick):

Home		Work		On the road		Elsewhere (please specify):

Were you interacting with anyone (including on the phone):

Yes		No (if no, please skip the next question):		

Who were you interacting with (please tick <u>all</u> that apply, and <u>specify</u> where requested):

Your child who is part of the PFL programme		Your other child/children (please tick, & specify ages in box to the right):					
Spouse/partner		Partner's child(ren)		Partner's relative(s)		Clients/customers	
Friend(s)		Other people's child(ren)		Work colleagues		Health professional(s)	
Own parent(s)		Partner's parent(s)					
Other relative(s)		Others (please specify):					

How did you feel during this episode?

Please rate each feeling listed below on the scale given. A rating of 0 means that you did not experience that feeling at all. A rating of 6 means that this feeling was a very important part of the experience. Please include an answer for each feeling. If you did not experience a particular feeling during the episode, please mark 0 for 'not at all'. Please circle the number between 0 and 6 that best describes how you felt.

	Not at all → Very Much						
Impatient	0	1	2	3	4	5	6
Happy	0	1	2	3	4	5	6
Frustrated/Annoyed	0	1	2	3	4	5	6
Depressed/Sad	0	1	2	3	4	5	6
Competent/Capable	0	1	2	3	4	5	6
Irritated	0	1	2	3	4	5	6
Relaxed	0	1	2	3	4	5	6
Affectionate	0	1	2	3	4	5	6
Angry	0	1	2	3	4	5	6
Stressed/Anxious	0	1	2	3	4	5	6
In control	0	1	2	3	4	5	6
Criticised/put down	0	1	2	3	4	5	6
Tired	0	1	2	3	4	5	6

Next Episode:

Please look at your Diary and select the <u>next</u> episode you noted:

This is episode number _____, which began at _____ and ended at _____

How did you feel during this episode?

What were you doing? (please tick <u>all</u> that apply):				
grooming/self care		exercising (alone/group)		Other(please specify):
getting child(ren) dressed		attending training (paid/unpaid)		
feeding your child(ren)		paid work		
eating		taking care of your child(ren)		
commuting		playing with your child(ren)		
doing housework		putting child(ren) to bed		
shopping		computer/internet/email (home)		
preparing food		on the phone/skype		
socialising		watching TV		
relaxing		sleeping		

Where were you? (please tick):						
Home		Work		On the road		Elsewhere (please specify):

Were you interacting with anyone (including on the phone):				
Yes		No (if no, please skip the next question):		

Who were you interacting with (please tick <u>all</u> that apply, and <u>specify</u> where requested):							
Your child who is part of the PFL programme		Your other child/children (please tick, & specify ages in box to the right):					
Spouse/partner		Partner's child(ren)		Partner's relative(s)		Clients/customers	
Friend(s)		Other people's child(ren)		Work colleagues		Health professional(s)	
Own parent(s)		Partner's parent(s)					
Other relative(s)		Others (please specify):					

Please rate each feeling listed below on the scale given. A rating of 0 means that you did not experience that feeling at all. A rating of 6 means that this feeling was a very important part of the experience. Please include an answer for each feeling. If you did not experience a particular feeling during the episode, please mark 0 for 'not at all'. Please circle the number between 0 and 6 that best describes how you felt.

	Not at all → Very Much						
Impatient	0	1	2	3	4	5	6
Happy	0	1	2	3	4	5	6
Frustrated/Annoyed	0	1	2	3	4	5	6
Depressed/Sad	0	1	2	3	4	5	6
Competent/Capable	0	1	2	3	4	5	6
Irritated	0	1	2	3	4	5	6
Relaxed	0	1	2	3	4	5	6
Affectionate	0	1	2	3	4	5	6
Angry	0	1	2	3	4	5	6
Stressed/Anxious	0	1	2	3	4	5	6
In control	0	1	2	3	4	5	6
Criticised/put down	0	1	2	3	4	5	6
Tired	0	1	2	3	4	5	6

Section 4: A Few More Questions about Yesterday

Now that you have told us about your day in detail, we have a few more general questions.

We would like to know overall how you felt and what your mood was like yesterday.	
Thinking only about yesterday, what **percentage** of the time were you:	
In a bad mood	
A little low or irritable	
In a mildly pleasant mood	
In a very good mood	
Total:	100%

Now we would like to know how typical yesterday was for that day of the week (i.e. for a Monday, for a Tuesday and so on). Compared to what that day of the week is usually like, yesterday was... (please circle one):				
Much worse	Somewhat worse	Typical	Somewhat Better	Much Better
Now please tell us whether you felt any anxiety or stress yesterday. Compared to what that day of the week is usually like, yesterday I felt...(Please circle one):				
A lot more anxious	A little more anxious	Typical	A little less anxious	A lot less anxious

Now we would like to know overall how you felt and what your mood was like when you were with your child/children yesterday.	
Thinking only about the time you spent with your child/children yesterday, what percentage of the time were you:	
In a bad mood	
A little low or irritable	
In a mildly pleasant mood	
In a very good mood	
Total:	100%

Now we would like to know how yesterday compared to a typical day with your children. Compared to a typical day with my children, yesterday was (please circle one):				
Much worse	Somewhat worse	Typical	Somewhat Better	Much Better

Now please tell us whether you felt any anxiety related to your children yesterday. Compared to what that day of the week is usually like, yesterday I felt...(Please circle one):				
A lot more anxious	A little more anxious	Typical	A little less anxious	A lot less anxious

During the past month, how would you rate your overall sleep quality (please circle one)?				
Very bad	Fairly bad	OK - neither good nor bad	Fairly good	Very good
During the past month, on average how many hours of actual sleep did you get at night?				____hours
Last night, how many hours of sleep did you get?				____hours
During the past month, how much of a problem has it been for you to keep up enough enthusiasm to get things done?				
No problem at all				
Only a very slight problem				
Somewhat of a problem				
A very big problem				

Finally, please tell us how you felt about this questionnaire by circling your response to the following two questions on the scale below.

Was it difficult to answer the questions? (Please rate your answer on a scale of 1-5, where 1 means “Not at all” and 5 is “very much”):				
1	2	3	4	5
Did you enjoy answering the questions? (Please rate your answer on a scale of 1-5, where 1 means “Not at all” and 5 is “very much”):				
1	2	3	4	5

This part of the study is now completed. Thank you for taking part