SOEP The Socio-Economic Panel

1062 2019

SOEPpapers
on Multidisciplinary Panel Data Research

SOEP – The German Socio-Economic Panel at DIW Berlin 1062-2019

Short- and Mid-Term Effects of a Parenting Program on Maternal Well-Being: Evidence for more and less advantaged mothers

Georg F. Camehl, C. Katharina Spiess, Kurt Hahlweg

DIW SOEP

SOEPpapers on Multidisciplinary Panel Data Research at DIW Berlin

This series presents research findings based either directly on data from the German Socio-Economic Panel (SOEP) or using SOEP data as part of an internationally comparable data set (e.g. CNEF, ECHP, LIS, LWS, CHER/PACO). SOEP is a truly multidisciplinary household panel study covering a wide range of social and behavioral sciences: economics, sociology, psychology, survey methodology, econometrics and applied statistics, educational science, political science, public health, behavioral genetics, demography, geography, and sport science.

The decision to publish a submission in SOEPpapers is made by a board of editors chosen by the DIW Berlin to represent the wide range of disciplines covered by SOEP. There is no external referee process and papers are either accepted or rejected without revision. Papers appear in this series as works in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be requested from the author directly.

Any opinions expressed in this series are those of the author(s) and not those of DIW Berlin. Research disseminated by DIW Berlin may include views on public policy issues, but the institute itself takes no institutional policy positions.

The SOEPpapers are available at **http://www.diw.de/soeppapers**

ISSN: 1864-6689 (online)

German Socio-Economic Panel (SOEP)
DIW Berlin
Mohrenstrasse 58
10117 Berlin, Germany

Contact: soeppapers@diw.de

Short- and Mid-Term Effects of a Parenting Program on Maternal Well-Being: Evidence for more and less advantaged mothers

Georg F. Camehl[1], C. Katharina Spiess[2,*], and Kurt Hahlweg[3]

Abstract

This paper evaluates how a light-touch parenting program for parents of children below school entry age affects maternal well-being. We first analyze data from a randomized controlled trial focusing on more advantaged parents. Second, we use a sample of mothers from deprived neighborhoods, for which we generate a control group using additional data. Overall, results show a relatively large positive effect of the intervention on maternal well-being, with the largest effects appearing three years after treatment for both groups, while less advantaged families also experience a well-being increase directly after treatment. Mechanisms are further explored.

Keywords: Parenting Program, Family Well-being, Instrumental Variables, Triple P

JEL: I31, I26, J13, C21, C26

Acknowledgments: The authors thank Wolfgang Schulz, Marie LeMouel, Jan Marcus, Sophia Schmitz and Sevrin Waights for their valuable comments on previous versions of this paper. Moreover, we thank participants of the 3rd LEER workshop on Education Economics in Leuven, Belgium, and the 2017 ESPE conference in Glasgow for helpful comments. We also thank Adam Lederer for helpful editorial assistance.

[1] DIW Berlin
[2] DIW Berlin and Freie Universität Berlin, School of Business and Economics
[3] Technical University Braunschweig, Institute for Psychology

*Corresponding author, Email: kspiess@diw.de.

1 Introduction

Early childhood intervention programs can have short- to long-term impacts on child development. Economists extensively evaluate programs focusing directly on fostering cognitive skills or socio-emotional behavior of children. This literature covers the evaluation of targeted programs, such as the Perry Program or Head Start (e.g. Barnett 2011 or Heckman et al. 2010), as well as the effects of universal programs[4]. However, in contrast to universal programs, many successful targeted early intervention programs do not just address young children, but also intensively work with parents – in particular mothers – to improve parenting skills. They combine a focus on children with components directly addressing parenting skills.

In addition to these programs, there are other early childhood intervention programs, which solely address parents, with the goal of improving child development via changing interactions between parents and children. Comparatively fewer studies exist on the effects of these programs on child development (for a summary of these studies, see e.g. Heckman & Mosso 2014). Although these programs address parents, in particular mothers, it is remarkable that there is little evidence how any of these programs affect parental, respectively maternal, well-being, in particular the well-being of parents of different backgrounds. The scarce focus on well-being is surprising, as maternal well-being is an outcome of interest for several reasons: Well-being and how it can be improved is of inherent interest as a core question in social science, including economics. It is also a central priority for policy makers in several countries (see for example Helliwell et al. 2016, OECD 2017). Moreover, maternal well-being is of interest as it improves child development (e.g. Berger & Spiess 2011, Dahlen 2016) and it influences decisions regarding maternal labor supply and fertility (e.g. Sandner 2019). More broadly, well-being is associated with positive health outcomes (e.g. Diener & Chan 2011) and labor productivity (e.g. Oswald et al. 2015).

Given this background, this paper is one of the few analyzing the effects of an early childhood intervention, namely a universal light-touch parent-training intervention, on maternal well-being. We use data from both a randomized controlled trial (RCT) and an offering of the program to a group for which no control group was recruited. While the RCT provided the intervention to a group of more advantaged parents, the other intervention focused on less advantaged mothers. All parents had children from 2.6 to 6 years old and the program was free of charge to participants. All families received the intervention voluntarily. More specifically,

[4] "Universal" means that a program is not explicitly targeted toward a particular group of parents. For some evaluations see e.g. Havnes & Mogstad (2011), Peter et al. (2016), and Backer et al. (2019). See Spiess (2018) for an overview.

we evaluate how one particular parenting program – namely Triple P (Positive Parenting Program) – affects the maternal well-being of two different groups of families.

Triple P is a multilevel parenting and family support strategy designed to reduce the prevalence of behavioral and emotional problems in children. From an economic perspective, it aims at increasing parental investments, which might be associated with parental well-being. An increase in parental investments is supposed to be of particular relevance for children and parents of less advantages families, who live in less privileged environments (e.g. Cunha et al. 2006, Cunha & Heckman 2007). Overall, effective parenting trainings might lead to a reduction in developmental inequalities (e.g. Caucutt et al. 2017), either via a direct improvement in child development or an increase in maternal well-being.

Potential mechanisms for an increase of parental well-being due to parenting training programs might be at least threefold (see also 6.1.2.): One direct channel could be that parenting trainings change parental abilities and skills, which leads to changes in parental behavior and parenting quality. This might result in an increasing well-being as parents realize that their parenting quality improved. An indirect channel could be that the training changed child behavior due to changes in parental skills, subsequently increasing parental well-being as parents' utility increases as child behavior improves. These mechanisms are, for instance, covered in an early model by Belsky (1984). A third mechanism might be the increasing social support mothers receive due to the program, resulting in a well-being increase as well. Independent of the mechanisms, we argue that potential improvements in maternal well-being could be an additional benefit improving the cost-efficiency of the program, even though the program's ultimate goal is to improve child outcomes.

Our main contribution is the analysis of a parent-training intervention of the type that is rarely analyzed using econometric methods, which focuses on the causal identification of the treatment. In particular, we consider maternal well-being, which is an important outcome for mothers, children, the families, and the economy as a whole. We do this by focusing on subjective satisfaction as a measure for well-being, an approach widely used in the economic literature. Not only can we measure short-term effects, but also mid-term effects. We use several data sets: RCT data for more advantaged families, data from an additional offering of Triple P in a deprived neighborhood, representative survey data and administrative data. We analyze the program effects on more *and* less advantaged mothers. We are able to measure the effect of the provision of the program and the actual treatment. This is a rarely used approach for this type of intervention study (for an exception e.g. Kim et al. 2018). Moreover, we contribute to the body of research on methodology of program evaluation, inspired by LaLonde

(1986), by combining data from a small-scale trial of the program with survey data to generate a control group. We validate this approach by remaking the control group from the RCT sample using the same approach. Thus, we can replicate the RCT results, which lends our method more validity.

We show that the parenting training program has positive well-being effects for both groups of mothers, with similar effect sizes, but some differences in when the effects occur. We show this for mothers in Germany, a country without a long tradition of parental programs. Germany provides a considerably different context from other countries where parenting programs exist and are analyzed more frequently, such as the US (see Sanders et al. 2014 for an overview) or low- and middle-income countries (see, e.g. Nahar et al. 2015).

2 Previous research on early childhood interventions focusing on maternal well-being

The literature on parental and, in particular, maternal well-being as an outcome of early childhood interventions is relatively small.[5] There is one strand of studies on the effects of center based childcare programs on maternal well-being, which are considered as early childhood interventions directly addressing children. For instance, Baker et al. (2008 and 2019) show that a reform in Quebec, Canada, which led to an expansion of child care provision, adversely affected various child and family well-being outcomes, including parental life-satisfaction, maternal depression, and work-family conflicts. For similar evidence, see Brodeur & Connolly (2013), Kottelenberg & Lehrer (2013), and Herbst & Tekin (2014). One proposed reason for that is that these expansions were made at the cost of childcare quality[6]. Based on Australian data, Yamauchi (2010) shows a slightly positive correlation of local day care availability with maternal life-satisfaction. For Germany, studies by Schober & Stahl (2016), Schober & Schmitt (2017), and Schmitz (2019) similarly point to a modest positive association between childcare provision and maternal well-being. Using a quasi-experimental approach, Schmitz (2019) shows that an expansion of childcare provision for children aged three or older significantly increased maternal life-satisfaction in Germany.

Other studies analyze childcare programs with a parental component. Examples of these include evaluations of the Incredible Years Program (Webster-Stratton et al. 2004), Early Head Start (Love et al. 2005; for an overview of earlier research see Barnett 1995), the parental and home

[5] If parental outcomes are included in studies outside economics, these most often refer to parenting practices or specific mental health aspects (Brooks-Gunn et al., 2000, gives an overview of earlier research on the topic).

[6] Another interpretation is related to the so-called "second-shift effect", which means that mothers whose employment increase due to an expansion of childcare provision, still bear the brunt of housework and childrearing after formal child care ends (e.g. Schmitz 2019).

components of the Perry Preschool Program (Belfield et al. 2006 and Heckman et al. 2010), and the Abecedarian Program (Masse & Barnett 2007). Using data from the Head Start Impact Study, Ansari et al. (2016) and Ansari & Gershoff (2016) examine whether one year of Head Start differentially benefited parents as a function of their initial parenting behaviors. The outcomes they consider include maternal depressive symptoms, which only showed improvements among parents most at risk.[7] The Perry Preschool Program had home visits once a week on average. The Abecedarian program did not have home visits, but interacted with parents at the Abecedarian day care center. The evidence generally supports positive (complementary) responses of parents to interventions. However, these evaluation studies are targeted at high-risk groups and do not take maternal well-being explicitly into account.[8] Moreover, as these studies direct the interventions on children *and* parents, it is difficult to distinguish the effects of the interventions through the focus on children *or* parents (see also Elango et al. 2015 and Kim et al. 2018).

Another strand of the literature covers studies on the effects of home visitation programs and exclusive parenting programs. This literature on home visiting programs mainly focus on maternal depression, as a very specific and extreme well-being measure. In these studies, the effects are modest, as summarized in the review by Ammerman et al. (2010).[9] In contrast to mental health, only a small number of studies investigate the effects of parenting programs on maternal well-being, as a more general measure of well-being: Notable examples of such studies analyze the Infant Health and Development Program (IHDP; Brooks-Gunn et al. 1994, Klebanov et al. 2001) and the Nurse Family Partnership Program (NFP; Olds 2006). Klebanov et al. (2001) analyze a randomized controlled trial of the IHDP with respect to maternal well-being. They provide an overall measure for well-being that includes aspects of mental health as well as general life satisfaction. Using US-data, they find that overall the program improved maternal well-being, especially for those mothers without a high school degree. Doyle et al. (2017), evaluating an adaption of the NFP program in Ireland, find improvements in experienced positive well-being.[10]

[7] Although maternal depression correlates with maternal well-being, it is more a measure of a well-being that clinically matters and that could be considered as an extreme bad well-being.
[8] The Incredible Years Program (Webster-Stratton et al., 2004) focuses on positive parenting for parents of children younger than eight. It specifically aims to reduce parent-child violence - a very specific goal.
[9] Johnson et al. (2000) evaluate the effect of the Community Mothers Program over the long term and find that parents reported greater self-esteem apart from positive effects on child development. Moreover, there are various programs offering, for example, text messages to parents to improve their parental behavior. However, they do not analyze the effects on parental well-being (for a recent study see Cortes et al. 2018).
[10] The evaluation of the NFP program by Olds (2006), using a randomized controlled trial, finds improvements for mothers through reduced dependence on public assistance. However, he does not analyze well-being measures explicitly.

The two closest studies to ours are Lindsay et al. (2011) and Sandner (2019). Lindsay et al. (2011) is similar in that it evaluates parenting programs (Incredible Years, Triple P, and "Strengthening families, strengthening communities") specifically for a European country (the United Kingdom), but it differs in that it looks at overall well-being. Nevertheless, in their measure for well-being, they include maternal life satisfaction, but they do not separately report the effects of the program on this measure. Furthermore, their study is not a randomized controlled trial; rather, it describes differences of parenting programs. Their analyses shows large improvements in their well-being measure from before to after participation in both Triple P and Incredible Years, but without comparisons to a control group. Moreover, they only focus on short-term outcomes. Sandner (2019) is similar to our study in that is looks at Germany, and uses a RCT as well. However, he evaluates a broader home visitation program, and the focus is on disadvantaged parents only. Using a randomized controlled trial, data from an own survey, and administrative data, his study finds evidence of improved maternal well-being. He uses various satisfaction measures to analyze well-being as mediator to fertility changes induced by the analyzed intervention. Thus, well-being is only focused as a mediator to other outcomes.[11]

The parental program we analyze is the Triple P program, which focuses on providing information that improves parenting skills with respect to handling child behavior and reducing parental stress. In this sense, it directly aims at improving parental well-being. Existing psychological studies analyze the effects of Triple P. Meta-analyses by Nowak and Heinrichs (2000), Sanders et al. (2014), and Wilson et al. (2012) report positive effects of Triple P on outcomes such as parenting style, parenting efficacy, as well as parental adjustment and relationships. Although, these are related to parental well-being, the outcomes they address could be results and drivers of it (e.g. Belsky 1984). None of these studies explicitly focuses on maternal well-being using an econometric approach. There are more recent Triple P-studies not covered in the mentioned summaries, such as Heinrichs et al. (2014), Kim et al. (2018), and Hahlweg & Schulz (2019). However, they all focus on child outcomes. Further, almost all studies – with the exception of the most recent ones – do not analyze outcomes beyond one year after the intervention and, thus, cannot measure mid-term effects. We focus on parental well-being measures up to three years after, which allows us to observe potential fade-out-effects, or sleeper-effects, that might occur due to changes in child behavior that might take some time to develop.

[11] There are other intervention studies taking place in developing countries, see section 1.

3 The Intervention

Aimed at teaching parents how to engage with their children, the Triple P-Program can be administered at five different intensities, referred to as levels (Sanders, 1999 and 2012). The first two levels consist of broadly spreading information through videos and leaflets. Level three includes group training, but with only a narrow focus on parenting skills, whereas level four broadens the focus to include general improvements to the home environment. Level five is an intensive family intervention resembling, and going beyond the aforementioned home-visitation programs.

The intervention we analyze covers level four of the Triple P program. It lasted 4 weeks, covering one weekly training sessions, 2 hours each. Moreover, the parent training took place in group settings with an average of six parents and with telephone follow-ups. It included video lectures as well as role-playing to learn how to handle difficult situations under the supervision of a trainer. Our treatment consisted of four sessions, each lasting two hours, conducted in the rooms of childcare centers that the child of the participating parents attended, and four optional telephone follow-ups.

4 Data and descriptive evidence

To analyze the effect heterogeneity of our Triple P-intervention different data sets for more and less advantaged families are used.

4.1 Sample of more advantaged families

For the first analysis, we use data from a unique intervention study (the so-called *Projekt Zukunft Familie 1*- ZF1 from here on,[12] Heinrichs et al. 2006) of Triple P in Germany. In this study, a randomized control trial (RCT) was carried out in one German city in 2001.[13] The participants were recruited from day care centers in a middle class neighborhood. Due to resource constraints, 17 of 23 daycare centers showing interest in the study were selected randomly to participate. Randomization was at the day care center level with a 66% chance of a center being in the intervention group.[14] Of 915 eligible families in the 17 selected day care centers, 282 agreed to participate in the study. After two families moved away, parents were told in which group they were, either the treatment group or the control group. After this, 42 families from the treatment group decided to skip the intervention, but they agreed to participate

[12] This study was funded by the German Science Foundation (DFG). All procedures were approved by the Human Subjects Protection Board of the German Association of Psychology.
[13] This is the city of Braunschweig, is a medium sized city in the state of Lower Saxony.
[14] Only one child was targeted for the parenting intervention.

in all follow-up surveys. The data includes 280 families of children between 2.6 and 6 years of age. All children attended day care centers at the beginning of the study and all parents spoke German. The field experiment was conducted with 186 families in the treatment and 94 families in the control group. In most cases, only the mother attended the program (only 6% of fathers in two-parent-households participated directly in the program). Single parents are excluded from the analyses, as differential effects of parenting programs are expected for them (Heinrichs et al. 2009) and because the sample size of single parents is too small for a separate analysis. Although the sample was relatively large compared to other intervention studies, the interpretation of the results have to keep in mind that the later analysis lacks statistical power due to the clustering of the standard errors on the center level. The later allows us to take into account that the randomization took place at the day care center level.

Panel attrition of the remaining sample was low: Of 219 coupled mothers, who participated in the study, 206 still answered the questionnaire three years later; an attrition rate of 6%. All subsequent analyses use the sample of mothers for whom information is available for the whole period.

4.2 Sample of the less advantaged families

In addition to the randomized controlled trial evaluated above, Triple P was independently offered to mothers in a deprived neighborhood of the same Germany city. This was part of another intervention study, the *Projekt Zukunft Familie 2*[15] (ZF2 from now on, Heinrichs & Jensen-Doss 2010). The analysis of this sample shows how the parenting training affects less advantaged mothers. This group is of particular interest with respect to avoiding inequalities in parental behavior early on. This intervention study includes an additional 197 treated families from 15 day care centers where the program was presented in various languages, including, among others, Turkish and Russian. They also took part in the Triple P-intervention as described in section 3. According to the ZF1 sample, the families had to have children between 2.6 and 6.5 years of age and they had to speak at least basic German. However, no randomization of the treatment took place and so no control group exists for this study, meaning that all families in the ZF2 group were actually treated. In order to generate a control group for ZF2 group, we use survey data from the German Socio-Economic Panel Study (SOEP) and administrative data.

The SOEP is the largest household panel study in Germany (Goebel et al. 2019). We use SOEP waves from 2010 through 2014, as these are the first to include a question on family life

[15] This study was funded by the Jacobs Foundation.

satisfaction that is identical to that in the ZF2 (and ZF1) data. Overall, 2,548 individuals with children under the age of seven answered the personal questionnaire for all four waves in the SOEP. We restrict the sample to mothers whose child is between 2.6 and 6 years old and attended a day care center at the time of the interview. This reduces the sample size to 2,120 observations.[16] In addition, we make use of data on the regional level, provided by *Federal Statistical Office* and *Statistics Braunschweig* (see below for details).

An advantage of creating a control group using SOEP data is that, in addition to the well-being measure we analyze, it includes a rich set of regional, socioeconomic, and sociodemographic characteristics. One challenge of the dataset is that the outcome is measured on slightly different scales in the two samples we use. While the ZF2 study uses a five point Likert-type scale, SOEP uses an eleven-point scale. Thus, we transform the eleven-point scale into a five-point scale. Another challenge is that the well-being information in the SOEP is measured some years later than in ZF1- and ZF2-samples. This question might be crucial with respect to the outcome variable. Additionally, there might be differential trends stemming from unobservable shocks to one of the datasets. Using related well-being measures, which are available in the SOEP for the same time-period in which the ZF1- and ZF2-data collection took place, we show that these are unjustified concerns. We show empirically that it is valid to assume that the trajectory of satisfaction with family life in our SOEP based control group is a good proxy for that in the treatment group given all observables. For the presentation of this evidence and further discussion of concerns related to this, see Appendix A.

4.3 Outcomes and descriptives

Outcomes. We use the satisfaction of the mother with her family life as our measure for maternal well-being.[17] More specifically, we analyze the differences between pre-treatment satisfaction with family life and measurements of it at four different points in time following the treatment. We use this satisfaction variable for two reasons: First, Triple P aims at improving parent-child-interactions and, thus, changes in family life satisfaction are reasonable due to various mechanisms, among them improvements in child behavior (see the discussion on mechanisms). Second, this measure of satisfaction is also available in the representative survey dataset that we use for additional analyses (see above). Furthermore, Schober & Stahl (2016) show that satisfaction with family life is the well-being measure that is most affected by early childhood programs compared to satisfaction measures related to other areas of life. We follow the practice in the well-being literature and assume the satisfaction variable is on a

[16] There are no SOEP observations that were excluded due to insufficient knowledge of German.
[17] For more information on the scale, see Henrich & Herschbach (2000).

cardinal scale.[18] We use standardized measures of the well-being measure: Differences of well-being are Z-standardized for each point in time such that coefficients of regression models correspond to differences in terms of standard deviations.

In order to obtain a more comprehensive view on the mechanisms of the influence of Triple P on maternal well-being, we analyze two additional outcome measures: First, we consider a variable that measures parenting skills (see Arnold et al. 1993). The ZF1 study measures parenting skills as strategies and actions concerning the upbringing of children.[19] Second, we consider a child related outcome measure, the Child Behavior Checklist (CBCL, Achenbach and Rescorla 2000, Döpfner et al. 2014). The CBCL is a widely used and validated scale that measures child behavior. Parents answer items on noticeable behavioral and emotional problems, indicating how frequently these occur. The CBCL can be grouped into two subscales, one on internalizing behavior (for example displayed through depressive symptoms) and one on externalizing behavior (for example aggressive behavior toward others). We use this scale as another outcome measure to learn more about a potential indirect channel of parenting via child behavior on maternal well-being.

Covariates. Our estimations control for various covariates, such as child-, household- and, mainly, mother-related characteristics, such as education and employment status. Table 1 reports the descriptive statistics of the covariates, which are measured pre-treatment. Children are on average four years old, the gender ratio is roughly balanced, and children have an average of one sibling. About 65% of the mothers live in households with a monthly net income above 1,500 Euro,[20] which can be considered as higher income households. Additionally, more participants in the ZF1 study have earned a university entry degree (*Abitur* in German) than the average population and the sample of less advantaged mothers. Overall, the ZF1 sample covers more advantaged and, thus, higher SES families (see also Kim et al. 2018).

[18] Although life satisfaction is principally a latent variable, in many surveys respondents are asked to grade it on an ordinal scale (Schröder & Yitzhaki 2017 for a discussion.). Researchers then usually treat these answers as cardinal variables, which assumes that all respondents interpret the question in a similar way (that is, assuming the distances between items in terms of the latent underlying variable are equal).

[19] The German version (Naumann et al. 2010) of the parenting scale (for the English version, see Arnold et al. 1993) is administered to assess parenting skills. The scale is a 35-item questionnaire that measures dysfunctional discipline styles in parents. It yields a total score based on three factors: *laxness* (permissive discipline), *over-reactivity* (authoritarian discipline, displays of anger, meanness, and irritability) and *verbosity* (overly long reprimands or reliance on talking). Higher scores indicate dysfunctional parenting behaviour. The total score has adequate internal consistency, good test-retest reliability (r = 0.84), and reliably discriminates between parents of clinic and non-clinic children.

[20] Income is only measured on an ordinal scale with seven steps of around 500€. Data alignment to make additional analyses feasible (see below) required us to generate a binary variable on income to have sufficient large overlap with the ZF2 sample.

In the ZF2 sample, pre-treatment satisfaction with family life is lower when compared to the ZF1 sample and the SOEP (Table 1, column 2). With respect to covariates, a large difference between the ZF2 data and the other two datasets emerges. Less than a quarter of families fall into the income group with an income over 1,500 Euros - compared to around 60% for the two other subsamples. This is what we would expect, as the sampling of ZF2 families focuses on deprived neighborhoods. The education variables also show a similar picture: Participants in the ZF2 study have, on average, lower school degrees, with a smaller share of mothers having earned a college degree. Thus, this sample reflects a group of lower SES families.

5 Empirical strategy

5.1 Strategy for the more advantaged families

Since the ZF1 data come from a RCT study, identification of causal effects is straightforward. The randomized assignment of individuals to either treatment- or control-groups provides exogenous variation that makes it possible to interpret mean differences as causal. However, we know from the implementation of the RCT that some mothers chose not to take part in the treatment even though they were assigned to the treatment group. Thus, the usual estimations will deliver the intention-to-treat-effect. However, as we are also interested in the treatment effect on compliers only, we further estimate a LATE-effect using a two-stage least squares estimation. In the two-stage least squares estimations the random assignment to the treatment group is used as an instrumental variable (IV). The assignment to the treatment group constitutes a valid instrument, as it is uncorrelated with unobserved characteristics of the mothers via the randomization. Therefore, the exogeneity assumption holds for this instrument. At the same time, it is highly correlated with taking up the treatment and, thus, also a relevant instrument (see Bloom 1984, Angrist 2014). Thus, we correct the ITT effects for dilution by non-compliers. In summary, we first measure the effect of providing the intervention and then the effect of participating.

To learn more about control and treatment group differences and the differences between actually treated (compliers) and those not treated (non-compliers), we analyze if covariates are balanced between the groups. Naturally, this only works for observable characteristics. Table 2 shows the mean differences of observable characteristics between groups after randomization and between those who subsequently took up the treatment and the remaining control group. The table shows that there are two variables for which the randomization did not seem to result in similar means for the treatment and control group: Mothers in the treatment group are statistically significantly more likely to have a college degree and less likely to only have a vocational degree. Therefore, we control for these differences. Furthermore, if we consider the

actual treated and untreated, mothers in the complier group are still less likely to have vocational training, but also more likely to not have a tertiary degree. This indicates that there was likely no significant selection into actual treatment participation based on observable characteristics apart from a higher likelihood of mothers without a tertiary degree to opt out of the treatment. Moreover, we use the others covariates in our estimations to further increase the precision of our estimates (see above).

In addition to looking at the effects of Triple P on maternal well-being at different points in time, we follow the procedure laid out by Anderson (2008) and also consider a summation index of the outcome over all four points in time when the well-being was measured as a robustness check. The summation index is a weighted average generated with predictions from a generalized least squares (GLS) model including only a constant. In this way, the weights are set according to the covariance of the outcomes per individual in order to maximize the amount of information. The summation index has an advantage for interpretation as it gives an indication of *overall* effectiveness of the parenting program. This procedure reduces the number of tests concerning the overall effectiveness of the treatment to one, making the analyses additionally more robust to multiple testing.

As we cluster our standard errors (see above) and only have a small number of clusters (11 clusters in the treatment group and 6 in the control group), conventional parametric tests based on an asymptotic distribution to calculate the standard errors would be inappropriate. Thus we apply a bootstrap *t* -test using the wild-cluster bootstrap procedure (Cameron & Miller 2015), which allows precise estimation of *p*-values, even with a small number of clusters. We use the wild cluster bootstrap method with 999 replications, which maintains the cluster structure in each bootstrap sample (see Cameron & Miller 2015, Davidson & MacKinnon 2010). As we are not aware of evidence suggesting that the Triple P has negative effects on parents (see also chapter 2), we conduct a one-tailed test. However, estimations are very similar using a two-tailed test, sometimes leading to slightly higher p-values.

5.2 Strategy for the less advantaged families

The empirical approach with the sample of the less advantaged families, is twofold: First, we use the ZF2 sample and the SOEP based control group to estimate the effects of the treatment via entropy balancing (Hainmüller 2012). We do this to make the treatment and control group comparable. Entropy balancing is a weighting method to preprocess the data such that the means and potentially higher moments of the generated control group match the treatment data.

It can be regarded as a matching method.[21] Thus, the conditional independence assumption (selection on observables) is necessary for the identification of causal effects. The corresponding balancing graphs, based on our samples, are shown in Appendix B. They all demonstrate that we use comparable groups.

Second, to further control for differences between the treatment group and the generated control group, we add regional administrative data. We do this, as we would like to use all relevant information that is available to us to generate a valid control group. However, this can make the optimization algorithm used for entropy balancing infeasible for two reasons (Hainmüller 2012): (1) Too many covariates compared to the number of observations, and (2) too large differences between the moments of the two groups for specific characteristics. Thus to resolve these issues, we conduct the following additional procedure: Based on the regional administrative data we selected, we generate subgroups of the SOEP control group: Each subgroup covers a certain share of SOEP respondents. Which individuals are in which subgroup depends on a measure of similarity between the SOEP individuals' home municipality and the one municipality of the ZF2 sample. Given the set of regional characteristics for this municipality, we calculate a similarity index for the regions where the SOEP individuals come from. As a measure of similarity, we calculate the Mahalanobis distance[22] based on regional indicators. For each region, the measure is given by

$$d_i(x,y) = \sqrt{\sum_{j=1}^{k} \left(\frac{x_j - y_{ij}}{sd(y_j)}\right)^2}$$

where *i* is the indicator for the region and *j* denotes the different characteristics, as shown in the data section. Observations are then ordered according to this similarity measure.

The regional indicators for the SOEP are given on the county (*Kreise* in German) or the municipality level (*Gemeinde* in German).[23] The regional sample unit of the ZF2 data set is a lower regional level, then a county or municipality. For this lower level, *Statistics Braunschweig* provided us with the same indicators given for the upper SOEP levels, namely the counties and municipalities. The selected indicators are the following: The unemployment rate, the percentage of families on welfare, the percentage of foreigners, and the average living

[21] For a discussion of the differences of regression and matching estimators with special regard to entropy balancing, see e.g. the empirical strategy section of Anger et al. (2017).
[22] The Mahalanobis distance is a generalization of the Euclidean distance including a z-standardization (subtracting the mean and dividing by the standard deviation) of the distance for each characteristic.
[23] We used data from the 2011 German census, accessed from https://ergebnisse.zensus2011.de/ (download November 2017).

space per dwelling in square meters. In addition, we include the socioeconomic background of families. See Table 6 for descriptive statistics of these regional indicators. Overall, means of the ZF1 sample are relatively similar to those in our SOEP sample. By definition, regional characteristics for ZF2 show that these families come from a deprived neighborhood. Note that Table 6 describes the sample before we apply our above-mentioned procedure to make the samples comparable.

As previously noted, we then use the SOEP and ZF1 and ZF2 samples and assign each observation a similarity measure based on regional characteristics. Then we create different control groups based on the quantiles in which an observation lies on the similarity distribution. We define ten cutoffs, starting with the first decile of the distribution, continuing in ten-percentage-point steps. In our main specification, we take the 20%-cutoff (for the reason of this see next paragraph). Results for other specifications are presented in Appendix B. A way to validate this entire procedure is to apply it to the ZF1 sample.

Through this procedure, we created artificial control groups. When we rerun the models using the ZF1 treatment group and the new artificial ZF1 control group, we find that the specification that comes closest to the original estimation is the one using the 20%-cutoff. This is the reason for choosing the 20%-cutoff for our main specification with the ZF2 treatment group. Results for other cutoffs are presented in Appendix B, as well.

6 Results

6.1 Results for the more advantaged families

Table 3 shows the treatment effect of Triple P participation on maternal well-being based on the ZF1 sample – the sample of more advantaged families. In principle, there are four different outcomes, each given by the difference between the well-being measurement at a given point in time and the pre-treatment measurement. First, we present the ITT effects. We regress on the pre-post difference in the well-being including only the pre-treatment value of well-being (column 1) to control for potential level effects and, in a second step, all covariates (column 2). These estimations show medium sized positive effects (Cohen 1988), which are only statistically significant for the third year following treatment and only when covariates are included (column 2). The effect of the treatment on the summation index is also positive, but, as expected, statistically insignificant. As explained above, these effects have to be interpreted as the effects of Triple P provision. Our two-stage least squares estimations (using the instrumental variable approach with control variables) for the ZF1 sample are presented in column 3 of Table 3. These are the LATE effects for the compliers. Again, we only observe an

effect three years after the treatment. Thus the positive effect of the treatment seems to become most apparent after a few years instead of directly after the treatment.

Overall, the increases in well-being of between 20 and 30% of a standard deviation are quite sizeable when compared results in the literature. For example, Yamauchi (2010) finds an increase of satisfaction with free time by 16% of a standard deviation when day care is available but no increase in satisfaction with family life. Schmitz (2019) finds an increase of general life satisfaction of 30% of a standard deviation if the child attended day care. However, day care is comparatively more costly than the Triple P intervention. Sandner (2019) finds a 15% increase of general life satisfaction after participating in the Pro-Kind program, a slightly lower effect.[24]

6.1.1 Robustness

In the main specifications for the ZF1 sample above, we use changes in the well-being measures as outcomes in a linear model. Since the original measures come from a five-point ordinal scale, it is *a priori* unclear if this is a valid approach. Thus, we also estimate ordered probit models as a robustness check for the upper satisfaction levels (Table 4).[25] Doing this, we can show that the effects are mainly driven by an increase to the highest satisfaction level. Much smaller negative effects occur if the satisfaction level of four is used as the outcome measure. Apart from this, it is confirmed that significant effects only occur three years after the treatment.

6.1.2 Mechanisms

As described above, we further analyze several mechanisms for the change in maternal well-being. Table 5 gives the results of the two-stage least squares estimations to discuss potential mechanisms, such as changes in parenting skills and changes in child behavior. Skills and behavior are recoded such that higher values correspond to better parenting skills and less problematic behavior by the child. Column 1 shows a strong effect of the treatment on parenting skills. This change in the home environment induced by changing parental skills seem to be permanent, even the summation index of this variable is highly significant. Positive effects for child behavior direct after the treatment are smaller and driven by the CBCL internalizing behavior subscale (columns 3): These are most pronounced and strongly statistically significant in the one-year follow-up. For the other points in time, effects on the internalizing subscale are

[24] To put these results into perspective, these effects are similar in size to the effects of involuntary job loss, which is considered to be one of the most detrimental shocks with respect to well-being. Kassenboehmer & Haisken-DeNew (2009) report effects of between -0.230 and -0.531 points on an 11-point Likert-scale for women in Germany, which translate into effect sizes of -13% to -31% of their reported standard deviations. The effects for men in this study are between .0.702 and -0.724 (-41% to -42%).

[25] Thus we estimate effects on the so-called "endline outcome". The downside of a more complex model is, of course, a loss of statistical power, so given that both models provide similar estimates, we prefer the simpler model.

also relatively high, but smaller and still significant in all measurement points. Thus, our results show that both - improved parenting skills and less problematic child behavior - could be channels through which the intervention improves maternal well-being. [26] However, there is no short-term effect on overall child behavior. This might be one explanation for the mid-term effects on well-being. Furthermore, it might take some time until the improvements of parenting skills (and in child behavior) materialize in well-being increases. This might be another explanation why the effects on well-being get statistically significant not earlier than three years after the intervention. Another potential mechanism, namely the actual social support mothers receive during the training, might not be the main mechanisms, as we would have expected the well-being to change in the short- and not medium-run, which is not the case.

6.2 Results for the less advantaged families

The results for the group of the less advantaged are shown in Table 7. Before we discuss these, we briefly refer to columns 1 and 2 of the table. We do this to validate the procedure for generating a control group for the ZF2 sample. Column 1 of Table 1 is identical to column 3 of Table 3. Column 2 shows the estimates for the ZF1 sample with the artificial ZF1 control group.[27] A comparison of both columns show that the effects using the artificial control group are, in general, somewhat smaller than those from the evaluation of the RCT, except for one case, the measurements after one year. However, they show a similar pattern with the largest effects emerging after three years.

Next, we consider the effect of the treatment on our less advantaged sample. We also find positive effects of the treatment for mothers living in deprived neighborhood. Moreover, we find a relatively large effect of the treatment direct after it is over. The effect size is almost the same as three years after the treatment. However, the effects are smaller and not significant in-between. For the positive effect three years after the treatment, we could assume similar mechanisms as discussed above for the more advantaged sample. However, the short-term effect directly after the treatment might indicate an additional mechanism. Less advantaged mothers' well-being might have increased due to the social support of the program. However, this is only a short run effect, fading out over time. Such an effect could not be measured for the more advantaged mothers. Here, we also check the assumption necessary to treat well-being

[26] We test if the confidence intervals of the coefficients overlap, which is not the case right after the treatment.
[27] In the Appendix, the full results for different subgroups of the artificial control groups are shown. Here we include only estimations using 20%-cut off for our similarity index to show that, in this case, the results come closest to the results derived from the analyses of the RCT.

as a metric variable and test for a linear relationship between treatment and outcome as a robustness check. The test indicates that linearity seems to hold.[28]

7 Conclusion

This study analyzes the effect of a parenting program on maternal well-being using both experimental (RCT) and non-experimental methods. The study allows focusing on samples of more and less advantaged families, which allows for identifying effect differences for different SES groups. In general, there is relative little literature on the effects of interventions aimed at improving child outcomes on parental well-being. The literature on the evaluation of parenting programs mainly focuses on child outcomes or on extreme maternal outcomes like maternal depressions. To our knowledge, this is the first paper evaluating the Triple P program with respect to maternal well-being in the medium-term using econometric measures to identify the causal effect of this program on the actual treated. Only Lindsay et al. (2011) analyze short-term associations for Triple P on maternal satisfaction.

Moreover, unlike other studies, we do not only measure the effects of program provision, but also actual treatment, using an instrumental variable approach and, thus, focusing on program compliers. From a methodological perspective, the study is also innovative because it constructs an artificial control group using survey and regional administrative data alongside various statistical methods, such as entropy balancing, and the creation of a similarity index using special distance measures. Moreover, with this method we could also replicate our RCT results.

Unlike most parenting program evaluations, this paper uses data from a context, namely Germany, with a universal day care system and an underdeveloped infrastructure of parenting programs. Thus, our results might be transferable to similar environments. However, we are cautious in claiming external validity of the RCT results to the entire population because the intervention excluded non-German-speaking parents.

The focus on both a more and a less advantaged sample allows us to evaluate if a more targeted approach of such a program or a universal approach would be more effective. We find that the program has a positive medium-run effect on maternal well-being for both groups of mothers, while e.g. the study by Klebanov et al. (2001) only finds effects of the parenting program they analyze for less educated mothers. Our medium-term effect sizes for both groups are quite similar, which means that we do not find any indication of targeting the program to one group only. However, if we assume the program is aimed at reducing inequality among children, a

[28] This is analogous to the robustness check in section 6.1.2. These results are available from the authors upon request.

universal approach is ineffective: in this case, a targeted approach toward more disadvantaged parents would be more effective. This is especially the case as we measure a small effect one year after the treatment for the group of less advantaged mothers that might be related to changes in gained social support due to the program. Nevertheless, in the medium-run, effects on maternal well-being are larger compared to the short-run. A potential mechanism for this is the improvement in parenting skills and child outcomes, which take time to materialize.

Overall, we find improvements of maternal well-being noteworthy, as Triple P is a low cost intervention yielding comparatively large effects on maternal well-being. These effects are comparable to the effects of much more costly day care expansion programs (e.g. Schmitz 2019). Moreover, our analysis shows that evaluations of parenting programs that only take child outcomes into account (e.g. Kim et al. 2018) may underestimate the benefits of the program. Improved maternal well-being has positive consequences for many people: the mothers, the children, the family, and society. It might even further increase the positive effect on children in future periods, as maternal well-being has been shown to improve child outcomes (e.g. Berger & Spiess 2011). Thus, an increase in well-being is not only a result of changes in child behavior but also a driver of further improvements in child behavioral outcomes. For policy makers who are interested in increasing the well-being of mothers with young children and their children, early childhood interventions addressing parents are an effective tool.

References

Achenbach, T. M. and L. A. Rescorla (2000): *Manual for the ASEBA preschool forms & profiles: An integrated system of multi-informant assessment; Child behavior checklist for ages 1 1/2-5; Language development survey; Caregiver-teacher report form.* University of Vermont.

Anger, S., G. Camehl and F. Peter (2017): Involuntary Job Loss and Changes in Personality Traits, *Journal of Economic Psychology* 60: 71-91.

Angrist, J. (2014): Instrumental Variables (Take 2): Causal Effects in a Heterogeneous World, https://ocw.mit.edu/courses/economics/14-387-applied-econometrics-mostly-harmless-big-data-fall-2014/lecture-and-recitation-notes/MIT14_387F14_Causaleffects.pdf, (Download May 2019).

Ansari, A., K. M. Purtell and E. T. Gershoff (2016): Parenting Gains in Head Start as a Function of Initial Parenting Skill, *Journal of Marriage and the Family* 78 (5): 1195–1207.

Ansari, A. and E. T. Gershoff (2016): Parent Involvement in Head Start and Children's Development: Indirect Effects Through Parenting, *Journal of Marriage and Family* 78 (2): 562–579.

Anderson, M. L. (2008): Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects, *Journal of the American Statistical Association* 103 (484): 1481-1495.

Ammerman, R. T., F. W. Putnam, N. R. Bosse, A. R. Teeters and J. B. van Ginkel (2010): Maternal depression in home visitation: A systematic review, *Aggression and Violent Behavior* 15 (3): 191–200.

Arnold, D. S., S. G. O'Leary, L. S. Wolff and M. M. Acker (1993): The Parenting Scale: A measure of dysfunctional parenting in discipline situations, *Psychological Assessment* 5 (2): 131-144.

Baker, M., J. Gruber and K. Milligan (2008): Universal child care, maternal labor supply, and family well-being, *Journal of Political Economy* 116 (4): 709-745.

Baker, M., J. Gruber and K. Milligan (2019): The Long-Run Impacts of a Universal Child Care Program, *American Economic Journal: Economic Policy* 11 (3): 1-26.

Barnett, W. S. (1995): Long-term effects of early childhood programs on cognitive and school outcomes, *The Future of Children* 5 (3): 25-50.

Barnett, W. S. (2011): Effectiveness of Early Educational Intervention, *Science* 333 (6045): 975-978.

Belfield, C. R., M. Nores, S. Barnett and L. Schweinhart (2006): The High/Scope Perry Preschool Program: Cost-Benefit Analysis Using Data from the Age-40 Follow up, *Journal of Human Resources* 41 (1): 162–190.

Berger, E. M. and C.K. Spiess (2011): Maternal Life Satisfaction and Child Outcomes: Are they related?, *Journal of Economic Psychology* 32 (1): 142-158.

Belsky, J. (1984): The determinants of parenting: A process model, *Child Development* 55: 83-96.

Bloom, H. S. (1984): Estimating the Effect of Job-Training Programs, Using Longitudinal Data: Ashenfelters Findings Reconsidered, *The Journal of Human Resources* 19: 544-556.

Brodeur, A. and M. Connolly (2013): Do higher child care subsidies improve parental well-being? Evidence from Quebec's family policies, *Journal of Economic Behavior & Organization* 93: 1-16.

Brooks-Gunn, J., L.J. Berlin and A.S. Fuligni (2000): Early childhood intervention programs: What about the family? In: J. P. Shonkoff and S. J. Meisels (eds.): *Handbook of Early Childhood Intervention.* Cambridge University Press.

Brooks-Gunn, J., M.C. McCormick, S. Shapiro, A.A. Benasich and G.W. Black (1994): The Effects of Early Education Intervention on Maternal Employment, Public Assistance, and Health Insurance: The Infant Health and Development Program, *American Journal of Public Health* 84 (6): 924-931.

Browning, M. and M. Gørtz (2012): Spending Time and Money within the Household, *The Scandinavian Journal of Economics* 114 (3): 681-704.

Cahuc, P., S. Carcillo and A. Zylberberg (2014): Labor Economics. 2nd edition. MIT Press.

Cameron, A. C. and D. L. Miller (2015): A practitioner's guide to cluster-robust inference, *Journal of Human Resources* 50 (2): 317-372.

Caucutt. E. M., L. Lochner and Y. Park (2017): Correlation, Consumption, Confusion, or Constraints: Why Do Poor Children Perform so Poorly?, *The Scandinavian Journal of Economics* 119 (1): 102-147.

Cobb-Clark, D., N. Salamanca and A. Zhu (2016): Parenting Style as an Investment in Human Development, *IZA Discussion Paper No. 9686.*

Cohen, J. (1988): *Statistical power analysis for the behavioral sciences.* Hillsdale. Lawrence Earlbaum Associates (eds.), 2nd edition.

Kalena E. C., H. Fricke, S. Loeb and D. S. Song (2018): Too Little or Too Much? Actionable Advice in an Early-Childhood Text Messaging Experiment, *IZA Discussion Paper No.11669.*

Cunha, F., J. J. Heckman, L. Lochner and D. V. Masterov (2006): *Interpreting the Evidence on Life Cycle Skill Formation*. In: Hanushek, E. A. und F. Welch (Eds.): Handbook of the Economics of Education, Vol. 1, Chapter 12.

Cunha, F. and J. J. Heckman (2007): The Technology of skill formation, *American Economic Review* 97 (2): 31-47

Cunha, F. and J. J. Heckman (2009): The Economics and Psychology of Inequality and Human Development, *Journal of the European Economic Association* 7 (2-3): 320–364.

Dahlen, H. M. (2016): The impact of maternal depression on child academic and socioemotional outcomes, *Economics of Education Review* 52: 77-90.

Darling, N. and L. Steinberg (1993): Parenting style as context: An integrative model, *Psychological Bulletin* 113 (3): 487-496.

Davidson, R. and J. G. MacKinnon (2010): Wild bootstrap tests for IV regression, *Journal of Business & Economic Statistics* 28 (1): 128-144.

Del Boca, D., C. Flinn and M. Wiswall (2014): Household choice and child development, *The Review of Economic Studies* 81 (1): 137-185.

Diener, E. and M.Y. Chan (2011): Happy People Live Longer: Subjective Well-Being Contributes to Health and Longevity, *Applied Psychology: Health and Well-Being* 3 (1): 1-43.

Döpfner, M., J. Plück, C. Kinnen and Arbeitsgruppe Deutsche Child Behavior Checklist (2014): *CBCL/6-18R, TRF/6-18R, YSR/11-18R. Deutsche Schulalter-Formen der Child Behavior Checklist von Thomas M. Achenbach. Elternfragebogen über das Verhalten von Kindern und Jugendlichen (CBCL/6-18R), Lehrerfragebogen über das Verhalten von Kindern und Jugendlichen (TRF/6-18R), Fragebogen für Jugendliche (YSR/11-18R).* Göttingen: Hogrefe.

Doyle, O., L. Delaney, C. O'Farrelly, N. Fitzpatrick, M. Daly (2017): Can early intervention improve maternal well-being? Evidence from a randomized controlled trial, *PLoS ONE* 12 (1).

Elango, S., J. L. García, J. J. Heckman and A. Hojman (2015): Early childhood education, NBER Working Paper No. 21766.

Ferrer-i-Carbonell, A. and P. Frijters (2004): How important is methodology for the estimates of the determinants of happiness?, *The Economic Journal* 114 (497): 641-659.

Fryer Jr, R.G., S. D. Levitt and J. A. List (2015): Parental incentives and early childhood achievement: A field experiment in Chicago heights (No. w21477). National Bureau of Economic Research.

Goebel, J., M. M. Grabka, S. Liebig, M. Kroh, D. Richter, C. Schröder, and J. Schupp (2019): The German Socio-Economic Panel Study (SOEP). *Jahrbücher für Nationalökonomie und Statistik / Journal of Economics and Statistics* 239 (2): 345-360. doi: https://doi.org/10.1515/jbnst-2018-0022.

Hainmueller, J. (2012): Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies, *Political Analysis* 20: 25-46.

Hahlweg, K. and W. Schulz (2019): Universal Prevention of Child Behavior Disorders by the TRIPLE P-Parent Training: 10 -Year Effectiveness from Mothers', Fathers' and Adolescents' Perspectives,
https://doi.org/10.24355/dbbs.084-201901071038-0.

Havnes, T. and M. Mogstad (2011): No Child Left Behind: Subsidized Child Care and Children's Lon-Rung Outcomes, *American Economic Journal: Economic Policy* 3 (2): 97-129.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev and A. Yavitz (2010): "The rate of return to the HighScope Perry Preschool Program", *Journal of Public Economics* 94: 128.

Heckman J. J. and Mosso S. (2014): The economics of human development and social mobility, *Annual Review of Economics* 6: 689–733.

Heinrichs, N., K. Hahlweg, H. Bertram, A. Kuschel, S. Naumann and S. Harstick (2006): Die langfristige Wirksamkeit eines Elterntrainings zur universellen Prävention kindlicher Verhaltensstörungen, *Zeitschrift für Klinische Psychologie und Psychotherapie* 35 (2): 82-96.

Heinrichs, N., K. Hahlweg, S. Naumann, A. Kuschel, H. Bertram and D. Ständer (2009): Universelle Prävention kindlicher Verhaltensstörungen mit Hilfe einer elternzentrierten Maßnahme:

Ergebnisse drei Jahre nach Teilnahme, *Zeitschrift für klinische Psychologie und Psychotherapie* 38: 79-88.

Heinrichs, N. and A. Jensen-Doss (2010): The effects of incentives on families' long-term outcome in a parenting program, *Journal of Clinical Child & Adolescent Psychology* 39 (5): 705-712.

Heinrichs, N., Krüger, S., & Guse, U. (2006). Eine experimentelle Studie zum Einfluss von Anreizen auf die Rekrutierung und Effektivität eines präventiven Elterntrainings, *Zeitschrift für Klinische Psychologie und Psychotherapie 35*: 72–86.

Heinrichs N., S. Kliem and K. Hahlweg (2014): Four-year follow-up of a randomized controlled trial of triple p group for parent and child outcomes, *Prevention Science* 15 (2): 233-245.

Helliwell, J. F., R. Layard and J. D. Sachs (2016): *World Happiness Report 2016, Update (Vol. I)*. New York: Sustainable Development Solutions Network.

Henrich, G., and P. Herschbach (2000): Fragen zur Lebenszufriedenheit (FLZ-M). In: U. Ravens-Sieberer, A. Cieza, M. Bullinger, N. von Steinbuechel & E. Poeppel (Hrsg.), *Lebensqualität und Gesundheitsökonomie in der Medizin. Konzepte, Methoden, Anwendung*. Landsberg: ecomed.

Herbst, C. M. and E. Tekin, (2014): Child Care Subsidies, Maternal Health, and Child–Parent Interactions: Evidence from Three Nationally Representative Datasets, *Health Economics* 23 (8): 894-916.

Johnson, Z., B. Molloy, E. Scallan, P. Fitzpatrick, B. Rooney, T. Keegan and P. Byrne (2000): Community mothers programme-seven year follow-up of a randomized controlled trial of non-professional intervention in parenting, *Journal of Public Health* 22 (3), 337–342 .

Klebanov, P. K., J. Brooks-Gunn, and M. C. McCormick (2001): Maternal coping strategies and emotional distress: Results of an early intervention program for low birth weight young children, *Developmental Psychology* 37 (5): 654.

Kassenboehmer, S. C. and J. P. Haisken-De New (2009): Your're fired! The negative effect of entry unemployment on life satisfaction, *The Economic Journal* 119 (546): 448-462.

Kim, J. H., W. Schulz, T. Zimmermann and K. Hahlweg (2018): Parent–child interactions and child outcomes: Evidence from randomized intervention, *Labour Economics* 54: 152–171.

Kottelenberg, M. J. and S. F. Lehrer (2013): New evidence on the impacts of access to and attending universal child-care in Canada, *Canadian Public Policy* 39 (2): 263-286.

LaLonde, R. (1986): Evaluating the Econometric Evaluations of Training Programs, *American Economic Review* 76 (4): 604-620.

Lindsay, G., S. Strand and H. Davis (2011): A comparison of the effectiveness of three parenting programmes in improving parenting skills, parent mental-well being and children's behavior when implemented on a large scale in community settings in 18 English local authorities: The parenting early intervention pathfinder (PEIP), *BMC Public Health* 11 (1): 962-975.

List, John A., A. M. Shaikh and Y. Xu (2016): Multiple hypothesis testing in experimental economics, *Experimental Economics*: 1-21

Love, J. M., E. E. Kisker, C. Ross, H. Raikes, J. Constantine, K. Boller, J. Brooks-Gunn, R. Chazan-Cohen, L. B. Tarullo, C. Brady-Smith, A. S. Fuligni, P. Z. Schochet, D. Paulsell and C. Vogel (2005): The effectiveness of early head start for 3-year-old children and their parents: Lessons for policy and programs, *Developmental Psychology* 41 (6): 885.

Masse, L.N. and W. St. Barnett (2007): Comparative benefit-cost analysis of the Abecedarian program and its policy implications, *Economics of Education Review* 26 (1): 113-125.

Nahar B, Hossain I, Hamadani J, Ahmed T, Grantham-McGregor S, Persson L.(2015): Effect of a food supplementation and psychosocial stimulation trial for severely malnourished children on the level of maternal depressive symptoms in Bangladesh, *Child: Care, Health and Development* 41 (3): 483–93.

Naumann, S., H. Bertram, A. Kuschel., N. Heinrichs, K. Hahlweg and M. Döpfner (2010): Der Erziehungsfragebogen (EFB). Ein Fragebogen zur Erfassung elterlicher Verhaltenstendenzen in schwierigen Erziehungssituationen, *Diagnostica* 56 (3): 144-157.

Nowak, C. and N. Heinrichs (2008): A Comprehensive Meta-Analysis of Triple P-Positive Parenting Program Using Hierarchical Linear Modeling: Effectiveness and Moderating Variables, *Clinical Child and Family Psychology Review* 11 (3): 114.

OECD (2017): *How's Life? 2017: Measuring Well-being.* Paris: OECD Publishing.

Olds, D.L. (2006): The Nurse-Family Partnership: An Evidence-Based Preventive Intervention, *Infant Mental Health Journal* 27 (1): 5-25.

Oswald, A.J., E. Proto and D. Sgroi (2015): Happiness and Productivity, *Journal of Labor Economics* 33 (4): 789-822.

Peter, F., P. S. Schober and C. K. Spiess (2016): Early Birds in Day Care: The Social Gradient in Starting Day Care and Children's Non-Cognitive Skills, *CESifo Economic Studies 62 (*4): 725–751.

Sanders, M. R. (1999): Triple P-Positive Parenting Program: Towards an Empirically Validated Multilevel Parenting and Family Support Strategy for the Prevention of Behavior and Emotional Problems in Children, *Clinical Child and Family Psychology Review* 2 (2): 71-90.

Sanders, M. R. (2012): Development, Evaluation, and Multinational Dissemination of the Triple P-Positive Parenting Program. *Annual Review of Clinical Psychology* 8: 1-35.

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

Sandner, M. (2019): Effects of Early Childhood Intervention on Fertility and Maternal Employment: Evidence from a Randomized Controlled Trial, *Journal of Health Economics* 63: 159–18.

Schmitz, Sophia (2019): The Impact of Publicly Funded Childcare on Parental Well-Being: Evidence from Cut-Off Rules, *European Journal of Population*, https://doi.org/10.1007/s10680-019-09526-z (online first).

Schober, P. and C. Schmitt (2017): Day-care availability, maternal employment and satisfaction of parents: Evidence from cultural and policy variations in Germany, *Journal of European Social Policy* 27 (5): 433-446.

Schober, P. and J. Stahl (2016): Expansion of full-day childcare and subjective well-being of mothers: Interdependencies with culture and resources, *European Sociological Review* 32 (5): 593-606.

Schroeder, C. and S. Yitzhaki (2017): Revisiting the evidence for cardinal treatment of ordinal variables, *European Economic Review* 92: 337-358.

Statistisches Bundesamt (2002): *Statistisches Jahrbuch 2002 für die Bundesrepublik Deutschland.* Wiesbaden: Metzler-Poeschl.

Webster-Stratton, C., M. J. Reid and M. Hammond (2004): Treating children with early-onset conduct problems: Intervention outcomes for parent, child, and teacher training, *Journal of Clinical Child and Adolescent Psychology* 33 (1): 105-124.

Wilson, P., R. Rush, S. Hussey, C. Puckering, F. Sim, C. S. Allely, P. Doku, A. Mc-Connachie and C. Gillberg (2012): How evidence-based is an 'evidence-based parenting program'? a PRISMA systematic review and meta-analysis of Triple P, *BMC Medicine* 10 (1): 130.

Yamauchi, C. (2010): The availability of child care centers, perceived search costs and parental life satisfaction, *Review of Economics of the Household* 8 (2): 231-253.

Tables and Figures

Table 1: Descriptive statistics of different samples (pre-treatment)

	(1) ZF1	(2) ZF2	(3) SOEP*
Child age (in years)	4.06	4.42	4.39
	(1.02)	(1.03)	(1.06)
Child female (%)	48.57	44.16	49.01
	(50.07)	(49.78)	(50.00)
Number of siblings	1.09	0.94	1.27
	(0.94)	(0.90)	(0.98)
Mother single (%)	10.50	12.30	10.09
	(30.73)	(32.97)	(30.13)
HH income above 1,500€ (%)	65.00	22.84	60.08
	(47.78)	(42.09)	(48.98)
Mother's age (in years)	35.23	33.49	34.55
	(4.95)	(5.54)	(5.59)
Mother German (%)	96.43	85.79	90.19
	(18.59)	(35.01)	(29.75)
Mother low school degree (%)	11.07	31.98	17.85
	(31.43)	(46.76)	(38.30)
Mother medium school degree (%)	34.64	45.69	45.81
	(47.67)	(49.94)	(49.83)
Mother high school degree (%)	54.29	22.34	36.34
	(49.91)	(41.76)	(48.11)
Mother no tertiary degree (%)	12.86	20.81	20.17
	(33.53)	(40.70)	(40.13)
Mother vocational training (%)	53.57	68.53	56.21
	(49.96)	(46.56)	(49.62)
Mother college degree (%)	33.57	10.66	23.62
	(47.31)	(30.94)	(42.48)
Mother working (%)	76.79	65.99	63.45
	(42.30)	(47.50)	(48.17)
Pre-treatment satisfaction with family life (5-point scale)	4.21	3.99	4.30
	(0.81)	(0.94)	(0.74)
N	280	197	2375

Notes: Means and Standard deviations in parentheses. * The SOEP descriptives refer to the sample before applying the described procedures to make the samples comparable.

Source: Projekt Zukunft Familie 1 - ZF1, Projekt Zukunft Familie 2 - ZF2, SOEP v31, German Census and Statistics Braunschweig.

Table 2: Observable characteristics for ZF1 treatment and control group as assigned through randomization (columns one to three) and for ZF1 participants who were eventually treated or not treated (columns four to six)

	Control group (1)	Treatment group (2)	Difference (1) – (2) (3)	Not Treated (4)	Treated (5)	Difference (4) – (5) (6)
Child age (in years)	4.11	3.97	0.14	4.03	4.00	0.03
	(0.98)	(0.97)		(1.04)	(0.93)	
Child female (%)	43.55	47.13	-3.59	45.56	46.51	-0.96
	(49.99)	(50.08)		(50.08)	(50.07)	
Number of siblings	1.15	1.08	0.07	1.08	1.11	-0.03
	(0.96)	(0.79)		(0.91)	(0.78)	
HH income above 1,500€ (%)	74.19	79.62	-5.42	77.78	78.29	-0.52
	(44.11)	(40.41)		(41.81)	(41.38)	
Mother's age (in years)	36.00	35.47	0.53	35.41	35.77	-0.36
	(4.79)	(4.43)		(5.19)	(4.02)	
Mother German (%)	98.39	96.82	1.57	97.78	96.90	0.88
	(12.70)	(17.62)		(14.82)	(17.40)	
Mother low school degree (%)	9.68	6.37	3.31	7.78	6.98	0.80
	(29.81)	(24.50)		(26.93)	(25.57)	
Mother medium school degree (%)	27.42	36.94	-9.52	32.22	35.66	-3.44
	(44.97)	(48.42)		(46.99)	(48.09)	
Mother high school degree (%)	62.90	56.69	6.22	60.00	57.36	2.64
	(48.70)	(49.71)		(49.26)	(49.65)	
Mother no tertiary degree (%)	12.90	7.01	5.90	14.44	4.65	9.79**
	(33.80)	(25.61)		(35.35)	(21.14)	
Mother vocational training (%)	38.71	59.87	-21.16***	42.22	62.02	-19.79***
	(49.11)	(49.17)		(49.67)	(48.72)	
Mother college degree (%)	48.39	33.12	15.27**	43.33	33.33	10.00
	(50.38)	(47.22)		(49.83)	(47.32)	
Mother working (%)	80.65	76.43	4.21	75.56	79.07	-3.51
	(39.83)	(42.58)		(43.22)	(40.84)	
Pre-treatment satisfaction with family life (5-point scale)	4.28	4.25	0.03	4.30	4.23	0.08
	(0.66)	(0.79)		(0.71)	(0.79)	
N			219			219

Notes: The table can be read as follows: Column one shows means of characteristics for mothers who were assigned to the control group, column two those who were assigned to the treatment group. Column three shows differences between means. Column four shows means of characteristics of mothers who did not receive the treatment (regardless of the outcome of the randomization), column five those of mothers who were threated. Column six again shows differences in means between the last two groups. Calculations exclude single mothers.

Significance levels: * $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$ of a t-test between groups.

Source: Projekt Zukunft Familie 1- ZF1.

Table 3: Effects of Triple P on difference in maternal well-being from pre-treatment to later measurements of well-being – sample of more advantaged families

Difference in well-being between pre-treatment and measurement…	(1) ITT	(2) ITT	(3) LATE+
… directly after the treatment	0.14	0.12	0.14
	(0.26)	(0.25)	(0.25)
… after 1 year	-0.01	0.04	0.05
	(0.47)	(0.38)	(0.39)
… after 2 years	0.15	0.16	0.19
	(0.24)	(0.19)	(0.18)
… after 3 years	0.13	0.23**	0.28*
	(0.29)	(0.05)	(0.06)
Summation index	0.12	0.16	0.20
	(0.33)	(0.16)	(0.18)
Covariates	No	Yes	Yes
N	206	206	206

Notes: Each cell shows effect sizes from one model including covariates as described in the data section above. + The LATE is estimated using two-stage-least-squares. P-values from single-tailed tests clustered on the center level using 999 bootstrap replications in parentheses.

Significance levels: * p≤0.1, ** p≤0.05, *** p≤0.01.

Source: Projekt Zukunft Familie 1 - ZF1.

Table 4: Estimating well-being using ordinal probit models (predicted probabilities)

Measure of well-being ...	**Ordered Probit Satisfaction Scale**	
	= 4	= 5
… directly after the treatment	-0.02	0.06
	(0.43)	(0.43)
… after 1 year	-0.01	0.03
	(0.70)	(0.69)
… after 2 years	-0.04	0.10
	(0.25)	(0.22)
… after 3 years	-0.05*	0.13**
	(0.06)	(0.05)
Summation index	-0.02	0.11
	(0.11)	(0.12)
N	206	206

Notes: Each cell shows effect sizes from one model including covariates as described in the data section above. P-values from two-tailed tests clustered on the center level using 999 bootstrap replications in parentheses.

Significance levels: *p≤0.1, ** p≤0.05, *** p≤0.01.

Source: Projekt Zukunft Familie 1 - ZF1.

Table 5: Potential mechanisms of Triple P on Well-being: Parenting skills and Child behavior

Difference in well-being between pre-treatment and measurement…	(1) Parenting skills	(2) Child behavior (CBCL)	(3) Internalizing behavior (CBCL)	(4) Externalizing behavior (CBCL)
… directly after the treatment	0.71***	0.19	0.26**	0.18
	(0.00)	(0.11)	(0.05)	(0.19)
… after 1 year	0.53***	0.31*	0.44***	0.19
	(0.01)	(0.07)	(0.01)	(0.18)
… after 2 years	0.67***	0.24*	0.24**	0.10
	(0.00)	(0.09)	(0.03)	(0.31)
… after 3 years	0.54***	0.24	0.30*	0.18
	(0.01)	(0.17)	(0.08)	(0.23)
Summation index	0.67***	0.27*	0.35***	0.18
	(0.00)	(0.09)	(0.01)	(0.20)
N	259	250	250	250

Notes: Each cell shows LATE effect sizes from one two-stage-least-squares model including covariates as described in the data section above. P-values from single-tailed tests clustered on the center level using 999 bootstrap replications in parentheses.

Significance levels: *p≤0.1, ** p≤0.05, *** p≤0.01.

Source: Projekt Zukunft Familie 1 - ZF1.

Table 6: Descriptive statistics of regional characteristics – Units are regions

	ZF1 (1)	ZF2 (2)	SOEP control group (3)
Share of foreigners (%)	7.30	11.98	7.30
	(0.00)	(0.00)	(5.45)
Living space (m^2 per person)	43.80	37.96	43.24
	(0.00)	(0.00)	(3.89)
Population density (people per km^2)	1262.20	3464.61	937.70
	(0.00)	(0.00)	(1052.10)
Mean age (in years)	43.00	44.43	43.27
	(0.00)	(0.00)	(1.98)
Strain (young + old / working age population)	55.80	74.42	59.12
	(0.00)	(0.00)	(6.10)
Share of Catholics (%)	14.00	13.66	29.75
	(0.00)	(0.00)	(24.21)
Share of Protestants (%)	39.60	42.12	28.33
	(0.00)	(0.00)	(16.86)
Share receiving welfare (%)	18.70	39.31	15.74
	(0.00)	(0.00)	(8.68)
Migration background (%)	21.40	49.95	18.49
	(0.00)	(0.00)	(9.88)
Share with high school degree (%)	53.70	n.a.	42.24
	(0.00)	n.a.	(12.44)
Unemployment rate (%)	6.10	9.41	4.87
	(0.00)	(0.00)	(2.08)
Share of women in employment (%)	74.10	69.51	75.01
	(0.00)	(0.00)	(3.63)
N	1	1	874

Notes: One observation equals one regional unit. Thus, the sample units are counties or municipalities depending on characteristics. In the case of the ZF1 and ZF2 samples, the regional unit is *Braunschweig*. Means and standard deviations in parantheses.

Sources: Projekt Zukunft Familie 1 und 2 – ZF1 und ZF2, SOEP v31, German Census and Statistics Braunschweig.

Table 7: Effects of Triple P on maternal well-being using the SOEP based control groups for the samples of more and less advantaged families

Difference in well-being between pre-treatment and measurement…	**(1) ZF1: LATE⁺ (using experimental control group)**	**(2) ZF1: Entropy balancing with SOEP control group**	**(3) ZF2: Entropy balancing with SOEP control group**
… directly after the treatment	0.14	0.10	0.22**
	(0.25)	(0.25)	(0.02)
… after 1 year	0.05	0.10	0.06
	(0.39)	(0.20)	(0.32)
… after 2 years	0.19	0.14*	0.12
	(0.18)	(0.09)	(0.21)
… after 3 years	0.28**	0.20*	0.21**
	(0.06)	(0.08)	(0.03)
Summation index	0.20	0.17**	0.19**
	(0.18)	(0.04)	(0.04)
N	206	320	316

Notes: Each cell shows effect sizes from one model including covariates as described in the data section above. + The LATE is estimated using two-stage-least-squares. For the LATE estimations on the first column, see Table 3.

P-values from single-tailed tests clustered on the center level using 999 bootstrap replications in parentheses.

Significance levels: *p≤0.1, ** p≤0.05, *** p≤0.01.

Sources: Projekt Zukunft Familie 2 – ZF2, SOEP v31, German Census and Statistics Braunschweig.

Appendix A: Data alignment of well-being measures between different data sources

Maternal well-being measures. Both the ZF1- and ZF2-studies include a five-point scale on satisfaction with family life. The categories of the scale are *unsatisfied, rather unsatisfied, rather satisfied, fairly satisfied,* and *very satisfied.* The SOEP uses an eleven-point scale (with only two labels at the extreme points: *completely unsatisfied* and *completely satisfied*). A transformation is needed in order to compare scores. Multiplying the SOEP scores by 0.4 and adding 1 results in a linear transformation. We assume that the scales represent the underlying latent variable accurately enough to use these models. Any measurement error in the scores from the ZF1- and ZF2-studies due to the reduced amount of information of the scale is captured by the error term in the regression. Table A1 shows the relations between the well-being scales from the SOEP and the ZF1- respectively ZF2 - studies as well as the transformed scale.

Table A1: Relations between the different maternal well-being scales

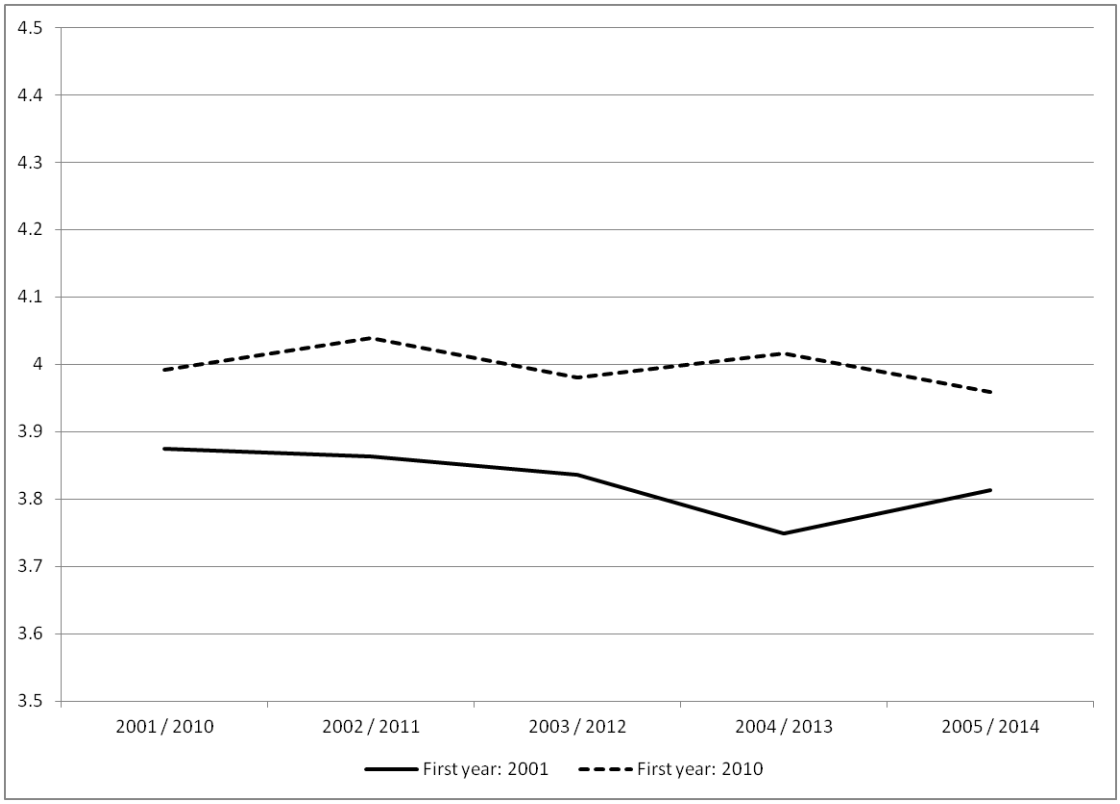
SOEP	Transformed scale	ZF1 and ZF2
0	1	1
1	1.4	
2	1.8	
-	-	2
3	2.2	
4	2.6	
5	3	3
6	3.4	
7	3.8	
-	-	4
8	4.2	
9	4.6	
10	5	5

Sources: Projekt Zukunft Familie 1 und 2 - ZF1 und ZF 2, SOEP v31

Due to data availability, SOEP data are from different years than the data from ZF1 and ZF2. This could bias estimates if systematic differences affect study participants differentially between the two studies. Figure A1 shows trajectories of general life satisfaction for the two time periods using SOEP data – this satisfaction measure is available in the ZF1 and ZF2 periods of observation and the period when the SOEP measures the satisfaction with family life. Moreover, this general life satisfaction measures correlates highly with the satisfaction measure we use. The solid line shows the trajectory of the general life satisfaction for the years 2001 to 2005, corresponding to the years when ZF1 and ZF2 were implemented. The dashed line shows the trajectory of the satisfaction variable for the years 2010 to 2014, corresponding to the years forming the basis for the generated SOEP control group. The lines run in parallel, indicating that well-being evolves similarly over time in both periods. There is only a level effect. However, as we include pre-treatment measures of our outcome variable in our models,

namely the satisfaction with family life, our estimations take potential differences in the levels into account.

Figure A1: Five-year trajectories of satisfaction with life in general for the two periods of interest (satisfaction points)



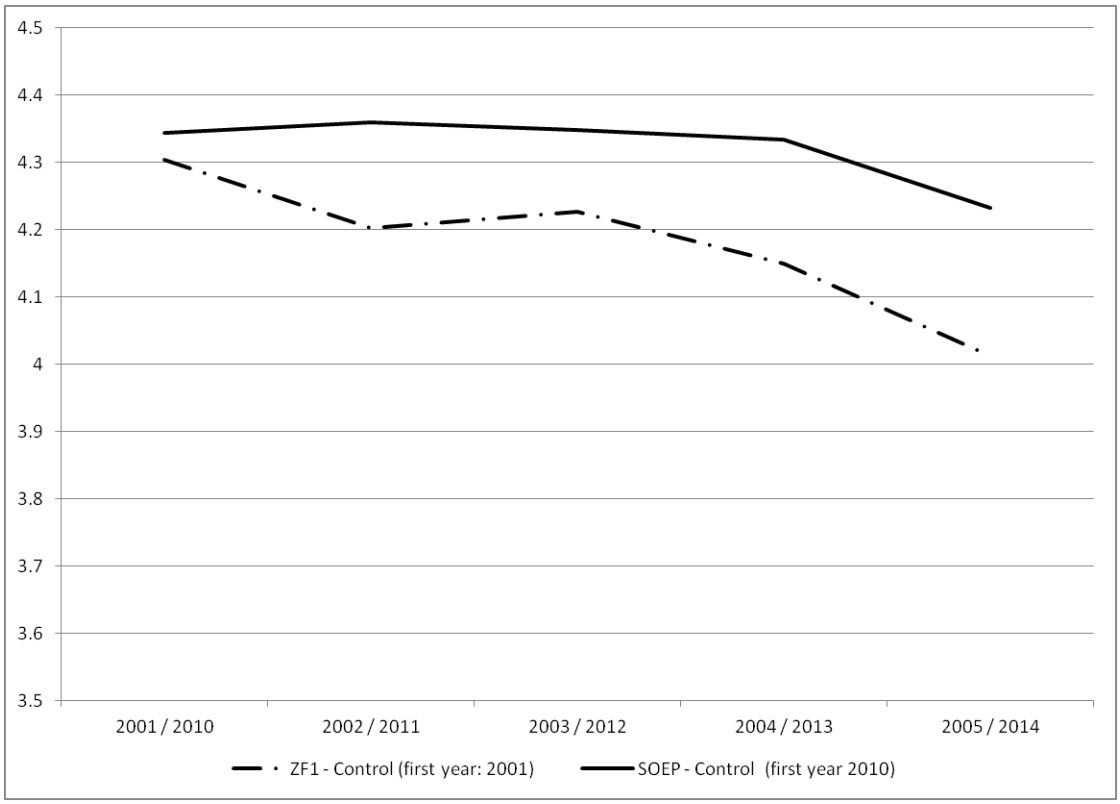
Notes: The first year always refers to the year when we have our well-being measures for the ZF1 and ZF2 samples. The second year always refers to the period were we have our well-being measures for the SOEP sample.

Source: SOEP v31.

Additionally, we check if trends for satisfaction with family life are similar in the ZF1 control group and the SOEP generated control group. Figure A2 shows that there is a decrease in satisfaction with family life for the ZF1 control group. This trend is slightly smaller for the SOEP control group. Apparently, our well-being measure, namely satisfaction with family life, decreased a bit more over time for families in our ZF1 sample in the period 2001 onwards than it did in Germany as a whole nine years later. Thus using the SOEP data without further adjustment, might lead to biased estimates regarding the effect of program participation for our estimations based on the ZF2 sample and the SOEP. Therefore we made further adjustments, as described above: When using data from regions that are similar in terms of our distance measure ("Mahalanobis distance") to the ones where the ZF1- and ZF2 -study were conducted, we get the same Triple P-effects using the SOEP generated control group as in our first main specification using only the ZF1 sample. Thus, in doing so, we can at least partly control for

factors leading to the different trajectories in figure A2. While not completely establishing external validity of the study for the entire German population, our results indicate that Triple P can be shown to have a similar effect on a population in Germany, when taking into account the variables we did.

Figure A2: Trajectories of satisfaction with family life for the ZF1 control group (2001 onwards) and the SOEP control group (2010 onwards)



Sources: Projekt Zukunft Familie 2 - ZF2, SOEP v31.

Appendix B: Entropy balancing and selection of observations based on regional data

Table B1: Using ZF1 and SOEP, including different percentiles of SOEP participants depending on distance measure

Difference in well-being between pre-treatment and measurement…	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
… directly after the treatment	0.03	0.10	0.07	0.04	0.01	0.08	0.07	0.06	0.08	0.08
	(0.12)	(0.14)	(0.13)	(0.11)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.09)
… after 1 year	0.20	0.10	0.07	0.08	0.04	0.07	0.08	0.05	0.04	0.03
	(0.17)	(0.12)	(0.11)	(0.11)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.09)
… after 2 years	0.11	0.14	0.10	0.07	0.06	0.04	0.05	0.07	0.07	0.08
	(0.13)	(0.11)	(0.10)	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
… after 3 years	0.24*	0.20*	0.14	0.12	0.13	0.11	0.12	0.13	0.14	0.12
	(0.13)	(0.11)	(0.10)	(0.10)	(0.10)	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)
Summation index	0.18	0.17*	0.12	0.10	0.07	0.09	0.10	0.10	0.10	0.09
	(0.12)	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.08)
N	222	320	419	517	614	714	810	908	1006	1103

Notes: Each cell shows effect sizes from one model including covariates as described in the data section. Robust standard errors in parentheses.

Significance levels: *p≤0.1, ** p≤0.05, *** p≤0.01.

Sources: Projekt Zukunft Familie 1 - ZF1, SOEP v31, German Census and Statistics Braunschweig.

Table B2: Using ZF2 and SOEP, including different percentiles of SOEP participants, depending on distance measure

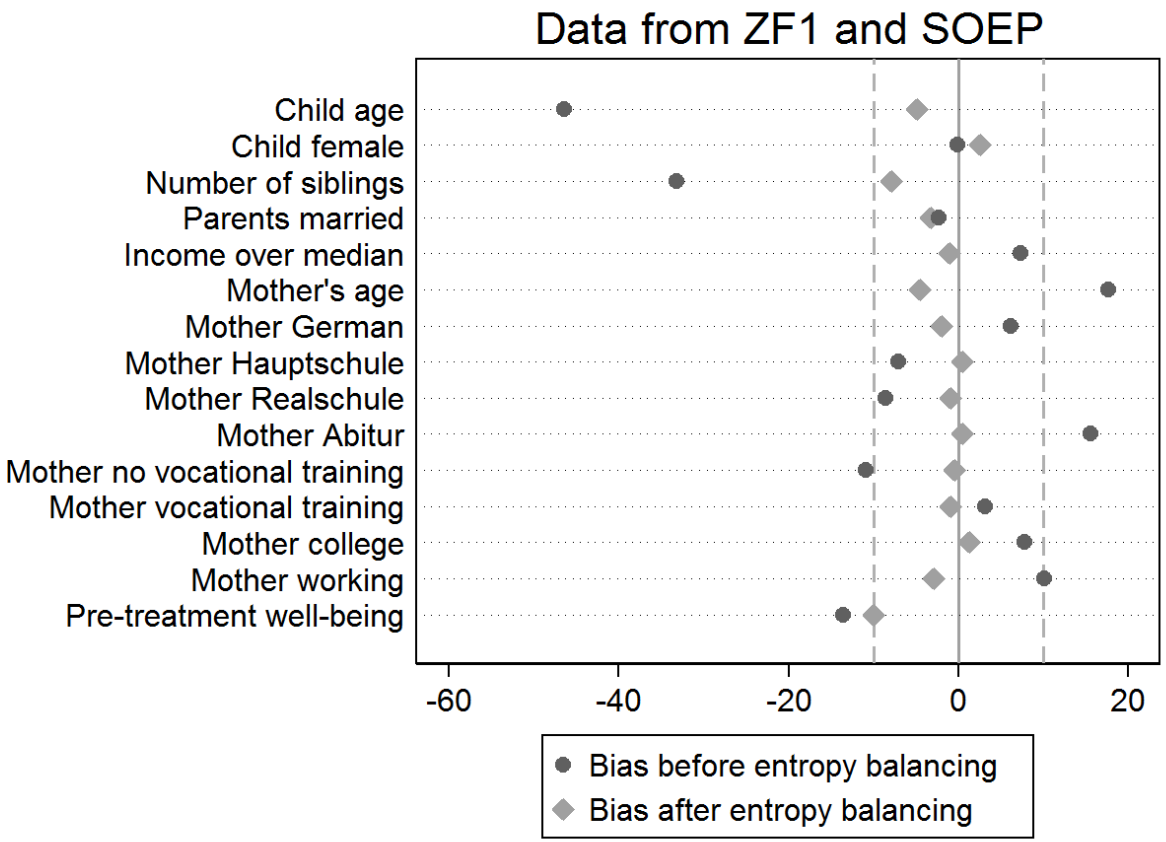
Difference in well-being between pre-treatment and measurement…	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
… directly after the treatment	0.12	0.22**	0.20*	0.20*	0.22**	0.20**	0.19*	0.20**	0.21**	0.19**
	(0.12)	(0.11)	(0.11)	(0.11)	(0.10)	(0.10)	(0.10)	(0.09)	(0.09)	(0.09)
… after 1 year	-0.02	0.06	0.04	0.01	0.03	0.01	0.03	0.06	0.06	0.04
	(0.12)	(0.11)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.09)
… after 2 years	0.40**	0.12	0.16	0.09	0.07	0.05	0.07	0.04	0.05	0.03
	(0.17)	(0.13)	(0.12)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.10)	(0.10)
… after 3 years	0.11	0.21*	0.21*	0.17*	0.13	0.12	0.14	0.12	0.11	0.10
	(0.12)	(0.11)	(0.11)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.10)	(0.09)
Summation index	0.19	0.19*	0.18*	0.14	0.14	0.12	0.13	0.13	0.13	0.11
	(0.12)	(0.10)	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
N	218	316	414	513	627	707	806	906	1002	1099

Notes: Each cell shows effect sizes from one model including covariates as described in the data section. Robust standard errors in parentheses.

Significance levels: *p≤0.1, ** p≤0.05, *** p≤0.01.

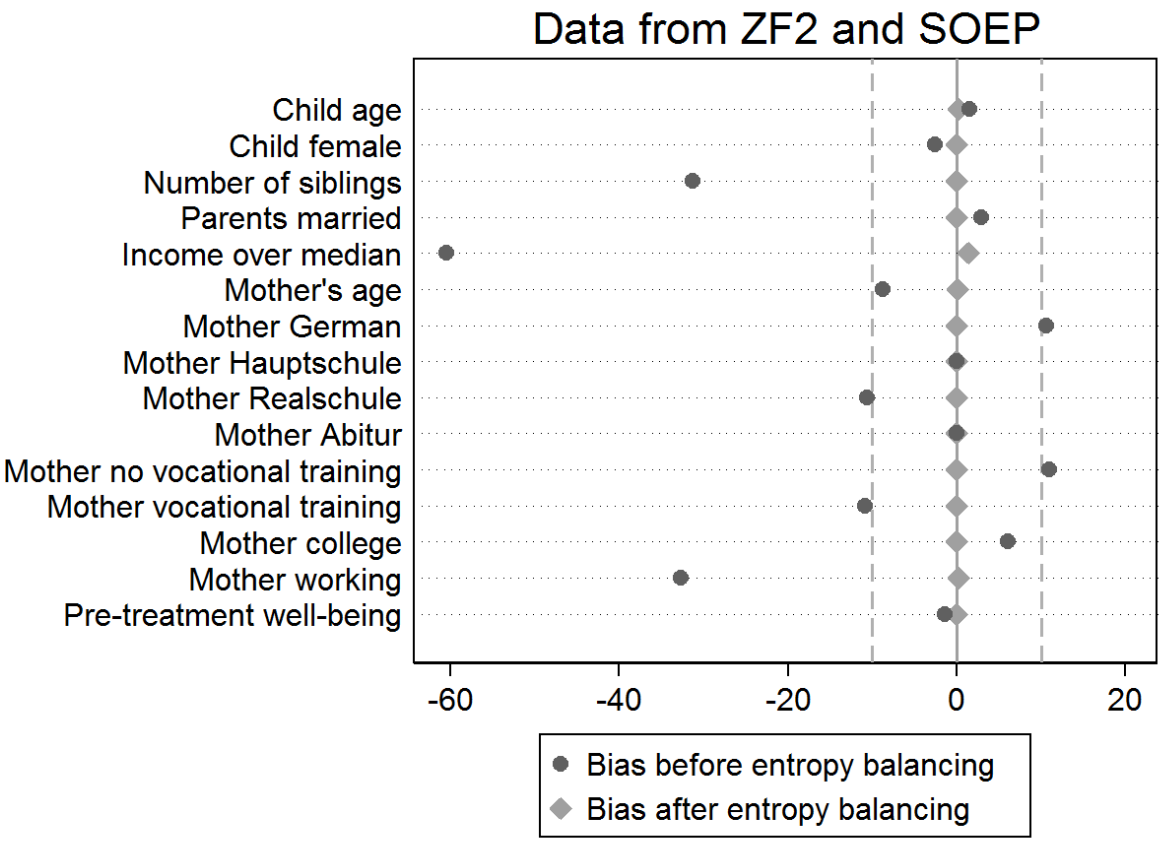
Sources: Projekt Zukunft Familie 2 – ZF2, SOEP v31, German Census and Statistics Braunschweig.

Figure B1: Covariate balances before and after entropy balancing – ZF1 (using 20% of SOEP observations)



Notes: The median income is 1,500€. Sources: Projekt Zukunft Familie 1 – ZF1, SOEP v31, German Census and Statistics Braunschweig.

Figure B2: Covariate balances before and after entropy balancing – ZF2 (using 20% of SOEP observations)



Notes: The median income is 1,500€.
Sources: Projekt Zukunft Familie 2 – ZF2, SOEP v31, German Census and Statistics Braunschweig.