Comparative Population Studies
Vol. 46 (2021): 453-486 (Date of release: 24.11.2021)

What You Need to Know When Estimating Impact Functions with Panel Data for Demographic Research*

Volker Ludwig, Josef Brüderl

Abstract: The estimation of impact functions – that is the time-varying causal effect of a dichotomous treatment (e.g., marriage, divorce, parenthood) on outcomes (e.g., earnings, well-being, health) – has become a standard procedure in demographic applications. The basic methodology of estimating impact functions with panel data and fixed-effects regressions is now widely known. However, many researchers may not be fully aware of the methodological subtleties of the approach, which may lead to biased estimates of the impact function. In this paper, we highlight potential pitfalls and provide guidance on how to avoid these in practice. We demonstrate these issues with exemplary analyses, using data from the German Family Panel (pairfam) study and estimating the effect of motherhood on life satisfaction.

Keywords: Impact functions · Fixed effects regression · Negative weighting bias · Motherhood · pairfam · Panel data analysis

1 Introduction

In demography, many research questions examine the total causal effect of an event (e.g., marriage, divorce, parenthood) on certain outcomes (e.g., earnings, well-being, health). *Allison* (1994) described this approach as asking for "the effects of events". Several papers in this Special Issue provide examples for such research questions and discuss research designs for the identification of such treatment effects. Experiments are in most cases not viable for analysing the effects of events in the social sciences, because manipulating them is unethical or not practically feasible. Therefore, cross-sectional survey data have traditionally been used as an alternative. With cross-sectional comparisons, the unbiased identification of a

* This article belongs to a special issue on "Identification of causal mechanisms in demographic research: The contribution of panel data".

URL: www.comparativepopulationstudies.de
DOI: https://doi.org/10.12765/CPoS-2021-16
URN: urn:nbn:de:bib-cpos-2021-16en3

treatment effect hinges on a very strong assumption: *all time-constant and time-varying* variables that affect both the treatment and the outcome (confounders) must be controlled for in the analysis. If some confounders are unobserved and cannot be controlled for, causal inferences will be biased.

Another research design for identifying the effect of events uses panel data. Panel data offer the possibility of implementing within-person designs for identifying causal effects. The main advantage of within-person designs is that time-constant confounders will not bias the estimation of the treatment effect. Thus, the unbiased identification of a treatment effect hinges on the much weaker assumption (compared to cross-sectional designs) that *all time-varying* confounders are controlled for in the analysis. In most cases, demographers implement a within-person design by using fixed-effects (FE) regressions. In fact, all papers in this Special Issue refer to this method and its advantages for causal analysis. Therefore, it is no surprise that with the advent of more and more panel data (see e.g., the most recent addition of "The Comparative Panel File, CPF", *Turek et al.* 2021), the estimation of treatment effects by using FE regression is a growing business in demographic research.

With long-running panel data, as they are meanwhile widely available, one observes many treated individuals for more than one time period after treatment. Given this set-up, it is even possible to estimate time-varying treatment effects, i.e., one can estimate the time-path of a causal effect. We term such a causal time-path "impact function", a term we borrow from *Andreß et al.* (2013). Impact functions obviously provide more insight, since we learn more than when estimating a time-constant treatment effect. Therefore, the recent demographic literature using FE methodology increasingly reports impact functions.[1]

The basic methodology of estimating impact functions with panel data and fixed-effects regressions is now widely known. Yet it is our impression that many researchers are not fully aware of the methodological subtleties of the approach, which may lead to biased estimates of the impact function. In this paper, we highlight potential pitfalls and provide guidance on how to avoid these in practice. We demonstrate these issues with exemplary analyses, using data from the German Family Panel (pairfam) study and estimating the effect of motherhood on life satisfaction.

The paper proceeds as follows: In the first section, we summarize the rationale for estimating impact functions. Section 2 provides the basic framework for estimating

[1] Another popular research question is to investigate "the effect of an event on (the transition rate of) another event". Event history methods with time-varying covariates are useful in this case (see the contribution by *Michaela Kreyenfeld* to this Special Issue). One could even estimate an analog to an impact function by modeling an interaction between the coefficient of the time-varying event variable and process time. For instance, one might be interested in the effect of the birth of a child on the divorce rate. If one would allow the effect of a birth to vary over marriage duration, this would be an event history analog to an impact function.
It must be noted, however, that (standard) event history analysis does not implement a within-person design. Therefore, estimates are unbiased only if both time-constant and time-varying confounders are controlled for. Only event history analysis with repeated events implements a within-person design (see *Allison* 2009).

impact functions. Section 3 provides an illustration using pairfam data and estimating the effect of the birth of a first (biological) child on happiness. Section 4 discusses pitfalls and remedies when estimating impact functions. Finally, Section 5 offers a discussion of the negative weighting bias with staggered treatment events.

2 Why Do We Estimate Impact Functions?

The estimation of impact functions has become a standard practice in demography, as well as in the social sciences and economics more generally. In fact, when setting up studies using a panel design, we should always care about the impact function for both substantive and methodological reasons.

Substantively, knowledge of the precise shape of the impact function is often important for theoretical reasons. In demographic research, for example, we would not only like to know to what extent childbirth raises well-being, but also whether this is a short-lived or a persistent – and perhaps even increasing – effect. Theoretically, the psychological adaptation/treadmill theory of happiness would lead us to expect a temporary effect that diminishes rapidly, a similar pattern as found for many other critical life events (e.g., *Diener et al.* 2006). Family economic models, however, would predict a permanent effect due to the benefits of household specialization (*Stutzer/Frey* 2006). This effect may even increase over time if specialization grows stronger after childbirth. The estimation of the impact function may thus help us discriminate between theories.

Methodologically, in case of time-varying treatment effects, different statistical models will yield different results. In the case of time-constant treatment effects, FE and first-differences (FD) regressions will provide the same results (without controls). This, however, is not true for time-varying treatment effects. As *Laporte* and *Windmeijer* (2005) show with time-varying treatment effects, FE and FD models produce estimates that differ tremendously.

Furthermore, we need to address a potentially time-varying treatment effect, because assuming a time-constant effect may bias our causal inferences. Recent papers demonstrated a "negative weighting bias" of the constant impact function (*Borusyak/Jaravel* 2017; *de Chaisemartin/D'Holtfoeuille* 2020; *Goodman-Bacon* 2018). This bias appears generally in the situation of staggered treatments (when the treatment appears at different time points) simply due to the mechanics of the FE estimator: FE estimation of a time-constant impact induces a bias due to down-weighting of treated observations well after treatment (for details, see Section 5). Since demographic panel data typically use staggered treatment designs, where treatment assignment occurs at different points in time, the issue is clearly of relevance. The problem can be solved, however, by simply modelling a flexible impact function (such as a dummy impact function, see below).

3 The Basic Framework for Estimating Impact Functions

Although impact functions could in principle be estimated by any basic statistical model for panel data (generally, clustered data), including pooled OLS and random effects regression, we focus in this paper on fixed effects modelling. The power of the FE approach for modelling impact functions stems from its property of unbiasedness under weak assumptions compared to other estimators. Although the crucial condition for the consistency of FE is strict exogeneity, the estimator is consistent and unbiased even if stable unit-specific characteristics are related to the treatment variable or other covariates. The FE model therefore allows, first, for the estimation of a causal treatment effect even when treatment assignment is not random, but rather the result of a conscious choice or otherwise systematic pattern (e.g., self-selection). Second, FE methodology helps with the age-period-cohort problem: It allows for the consistent estimation of age and period effects (while implicitly also controlling for cohort effects), which is crucial for the estimation of treatment effects with panel data, as we will argue below.

A simple FE model for estimation of a time-constant treatment effect with panel data for $i = 1, \ldots, N$ persons observed at $t = 1, \ldots, T$ points in time would be:

$$Y_{it} = \beta D_{it} + \boldsymbol{X}'_{it}\boldsymbol{\delta} + \alpha_i + \varepsilon_{it} \tag{1}$$

where Y_{it} is the outcome, D_{it} the treatment variable, X_{it} are confounders, and α_i are unit (person) fixed effects. ϵ_{it} denotes a time-varying error term that is assumed to be (strictly) exogeneous. Hence, for unbiasedness it must hold that $E(\epsilon_{it}|D_{it}, X_{it}, \alpha_i)=0$. From a design-based perspective, the strict exogeneity condition implies that the parallel trends assumption must hold. Furthermore, perfect collinearity is ruled out. Even strong imperfect collinearity may cause serious problems for the estimation of time paths. More details on the FE model can be found in *Brüderl* and *Ludwig* (2015).

There are three primary ways to model an impact function. Above, we assumed that D_{it} is a binary treatment variable, and that treatment assignment is an absorbing state, that is, once a unit switches to the treated status, the treatment variable turns 1 and subsequently does not return to 0. The treatment effect β in the model above thus is the single parameter that characterizes a *step impact function*: whenever a unit is treated, the model predicts an immediate and permanent change in the outcome (i.e., a time-constant impact).

To estimate a time-varying impact function, one parameter is not sufficient. Hence, we might add a further variable K_{it} to the model that contains the time passed since treatment assignment (technically this is an interaction variable between D_{it} and time since treatment). This would be the simplest way to specify a *continuous impact function*. The coefficient for D_{it} would tell us by how much the outcome changes immediately after treatment, and the coefficient for K_{it} indicates how much the outcome additionally changes with each time unit passing subsequently. Of course, continuous impact functions can be specified with polynomials to approximate more complex time paths. Alternatively, linear splines may be fit to model changes in the treatment effect in order to relax functional form assumptions.

The most often used type of impact function, however, is the *dummy impact function*, where D_{it} is replaced by a set of dummy variables for K_{it} (e.g., a separate dummy for every month or year since treatment):

$$Y_{it} = \sum_{k=0}^{K} \beta_k D_{it}^k + \boldsymbol{X}'_{it}\boldsymbol{\delta} + \alpha_i + \varepsilon_{it}, \tag{2}$$

where $k = 0$ is the first observation after treatment, and K is the last one.[2] (In the following, we denote D_{it}^k as the "k-year dummy".) Obviously, this specification allows us to fit the time path of a treatment effect in a very flexible way.

4 An illustration: The effect of the birth of a first (biological) child on happiness

In an influential paper, *Myrskylä* and *Margolis* (2014) studied parental happiness trajectories before and after the birth of a child (using data from the SOEP and the BHPS). This paper became a model for applying impact function methodology to demographic research questions. Therefore, to illustrate the estimation of impact functions, we replicate the study of *Myrskylä* and *Margolis* (2014) with data from the German Family Panel (pairfam). Thus, our research question is: How does the event of the birth of a first (biological) child impact the outcome "life satisfaction" (synonymously: "happiness"). We restrict our analysis to mothers. pairfam is particularly well-suited for the task at hand because it prospectively measures the birth of children and mothers' life satisfaction.

The German Family Panel is a nationwide longitudinal study of initially more than 12,000 randomly sampled individuals from the birth cohorts 1971-73, 1981-83, and 1991-93. pairfam started in 2008/09 with roughly one-hour face-to-face interviews. Respondents were approached annually in subsequent waves. For a detailed description of the study, see *Huinink et al.* (2011). Release 11.0 with Waves 1-11 is used for this analysis, which covers the observation period 2008-19 (*Brüder/ et al.* 2020). We use only the pairfam base sample and do not use the refreshment samples added to the survey continuously.

We construct the estimation sample according to the recommendations laid out in Section 5. Only never-treated females, i.e., those who never gave birth to a child before pairfam Wave 1 are included. We include only females with at least two observations in pairfam. For first-time mothers, the observation window is censored with the second pregnancy/birth. After these exclusions, our estimation sample comprises 2,982 women, of whom 505 gave birth to a first (biological) child.

---

2 Starting the numbering with 0 may intuitively seem confusing. But it makes sense, because the first time period after treatment might be approximately thought as the "period of treatment", the second time period as "one period after treatment", and so on. In addition, pre-treatment dummies can then be taken into account with -1, -2, and so on (see below).

These women provide 19,996 person-years. They were observed in the age range of 14-48. Due to low number of cases, we recoded 14 to 15 years, and 48 to 47 years. Thus, models include age dummies for ages 16 to 47 (15 being the reference category).

The outcome variable, life satisfaction, is measured by the question: "Now I would like to ask about your general satisfaction with life. All in all, how satisfied are you with your life at the moment?" Answers are recorded on an 11-point scale ranging from "very dissatisfied" (0) to "very satisfied" (10).

To estimate the total causal effect of a first birth on happiness, we include treatment variables as explained above (child dummy, time since birth). We derive these from the generated variable "nkidsbio", as it is provided by the pairfam team. (This has the implication that timing is not exact, since the exact birth date is not taken into account.) Time since birth starts at 0 (less than one year after birth) and reaches a maximum of 9 (nine years after birth). The maximum of 9 is reached if a woman is childless in the first wave, gives birth before the second wave, and does not drop out of the survey. For reasons of parsimony, we group the time dummies in an upper category of 4+ years.

According to the rules of modern causal analysis (*VanderWeele* 2019) one must control for potential confounders to identify a total causal effect.[3] Potential confounders are variables that are assumed to influence both the treatment and the outcome, first birth and life satisfaction. The most important confounder in our context is age (see below). As further controls, we include relationship status (dummies for living-apart-together, cohabitation, marriage; using single as the reference category), subjective health in the past four weeks (dummies on a five-point scale), and a dummy for pregnancy.

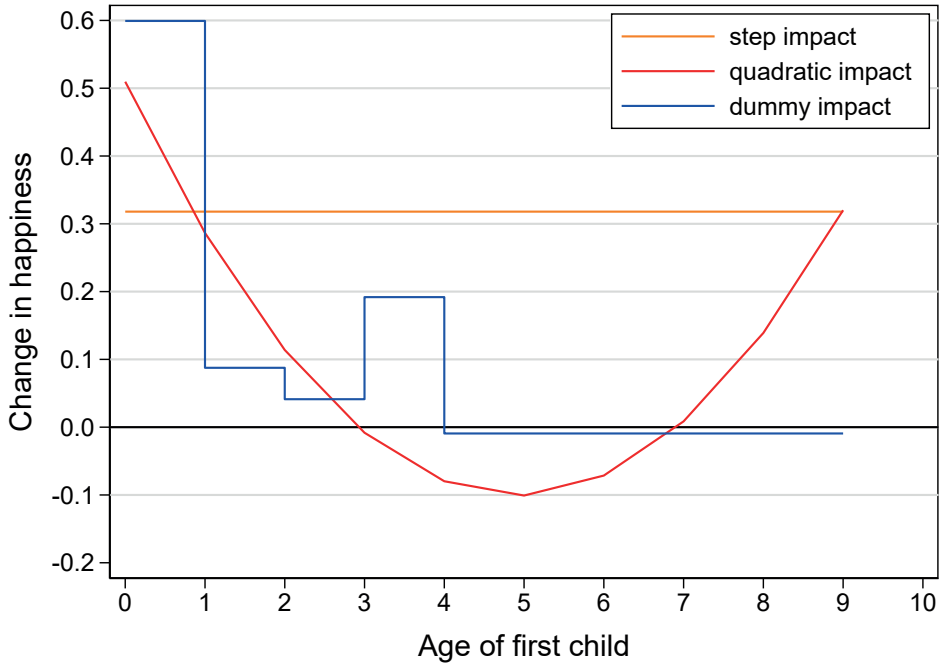
Fixed-effects regression with cluster-robust standard errors is applied (for details of fixed-effects estimation, see *Brüderl* and *Ludwig* 2015). Estimation is conducted with Stata 16.1. Below we provide only the graphical representations of the estimation results. Numerical results can be found in the Appendix.

First, we compare the results obtained with the three impact functions detailed in the last section. Figure 1 gives a graphic representation of the results. The step impact function tells us that after birth, mothers' happiness is on average higher by .32 scale points than before birth (the average happiness of all person-years before the event is the reference period). Interpreted as a causal effect, this means that giving birth to a child makes mothers happier by .32 scale points.

The quadratic impact function provides a completely different answer: Here, happiness increases by .51 scale points right after birth, followed by a steep decline and negative values three years after birth, with a low point after 5 years. This is followed by a steep increase in happiness.

[3] Mediators (intervening variables) should not be controlled for (overcontrol bias). Mediators must be controlled for if the target estimand is an indirect effect. Furthermore, colliders (variables that are affected by both the treatment and by the outcome) must not be included in the regression.

Fig. 1: The effect of the birth of a first child on mothers' happiness: comparing different impact functions



Notes: Results from three FE regression models. Controls are age dummies, relationship status dummies, subjective health dummies, and a dummy for pregnancy. See Appendix, Table A1 for numerical results.

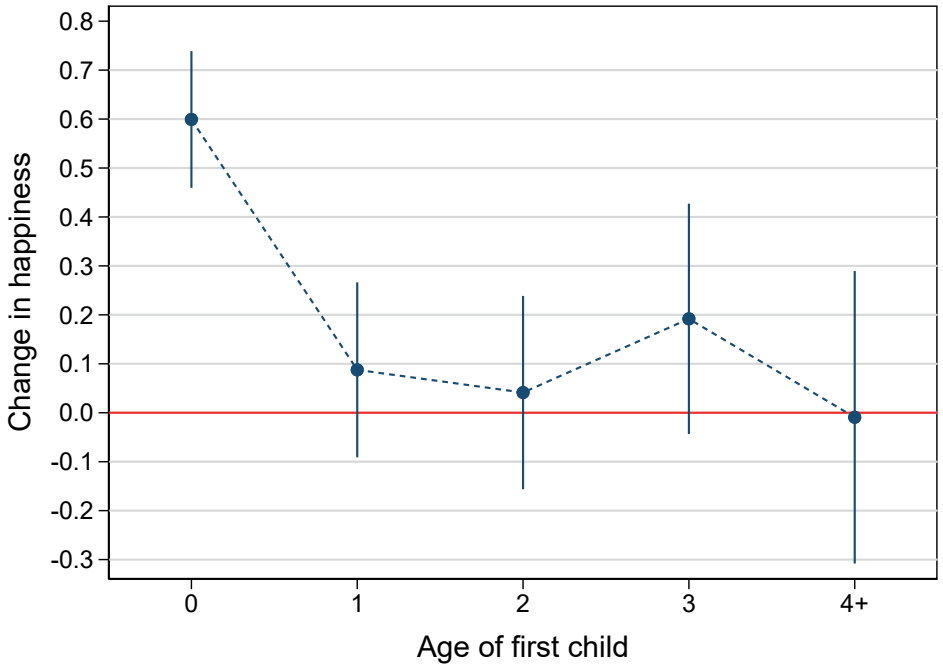
Source: pairfam release 11.0, own calculations

However, both step and quadratic impact functions provide grossly misleading answers, as the results of the dummy impact function show. We see that the birth of a child immediately increases happiness by .60 scale points (this is the coefficient of the "0-year dummy", see Table A1). However, this "baby effect" is very short-lived: it exists only in the first year after birth. The effect is much lower in the following years, and generally close to zero. This means that already one year after birth, mothers' happiness essentially returns to the level before birth, on average.

Thus, the lesson learned here is that one always should start with a flexible specification of the impact function, i.e., a dummy impact function. If the dummy impact function provides a clear pattern, then one could use a more parsimonious specification.

Figure 1 provides also an illustration of the negative weighting bias (we simplify here somewhat; exact details are provided in Section 5). Averaging the ten treatment effects estimated by the dummy impact function (see Table A1) gives: (0.60 + 0.09 + 0.04 + 0.19 – 0.01 * 6)/10 = 0.086. However, the step impact function provides an estimate of 0.32, i.e., grossly overestimates the true average of

Fig. 2: The effect of the birth of a first child on mothers' happiness: dummy impact function



Notes: Regression coefficients for "years since birth dummies" including 95 percent confidence intervals from a FE regression model. Controls are age dummies, relationship status dummies, subjective health dummies, and a dummy for pregnancy. This will be our reference model in the following. See Appendix Table A1, column 3, for numerical results.
Source: pairfam release 11.0, own calculations

the post-treatment effects. This is, because the mechanics of FE estimation down-weights the treatment effects that occur in later post-treatment periods. Thus, the FE estimate of the step impact is biased towards the early treatment effects. Note that the negative weighting bias does not always provide a negative (i.e., downward) bias of the treatment effect. In our case the bias obviously is upwards. "Negative" refers to the down-weighting of later treatment effects.

Further insights can be gained if one adds the confidence intervals to the plot of the dummy impact function. Figure 2 does this by plotting the regression coefficients of the dummies for the time since treatment along with 95 percent CIs. This figure also shows that the impact of treatment on outcome is indistinguishable from zero already one year after birth. Furthermore, it can be seen that the SEs increase with time since birth, simply because fewer and fewer observations contribute to estimation.

5 Pitfalls and Remedies When Estimating Impact Functions

In the following, we highlight important issues arising with the estimation of impact functions that are often overlooked in applied research.

Specification of the outcome trajectory

Specifying the average trajectory of the outcome over process time is a crucial step for any panel data analysis. Misspecification of the outcome trajectory may easily result in biased estimates of the treatment effect. This is true regardless of the type of impact function we assume. Thus, if we work with panel surveys of persons or households, it is essential to model the life course appropriately. Our specification must include age effects (and/or period effects). Otherwise, we run into problems of confounding: effects that are driven by ageing and maturation might produce a spurious treatment effect or suppress the true effect.

The FE model allows the researcher to consistently estimate age (A) effects, while simultaneously holding cohort effects (C) constant. C is constant across units, so the cohort effects are subsumed in the unit fixed effects. Controlling for A is crucial for estimating treatment effects on person or household panel data. Intuitively, age is very likely a relevant confounder in many demographic studies. In settings with all units initially untreated and treatment being an absorbing state, A and D are related by definition: as time passes, the treatment probability strictly increases (although possibly in a nonlinear way; first childbirth, for instance). Whenever the outcome changes over the life course, so that we have a correlation of A and Y as well, leaving out A from the regression would bias the effect of D.

An important issue concerns the inclusion of an appropriate control group in the estimation sample to make use of the advantages of panel data when studying life courses. The popular event study design that uses only those units who eventually get treatment while observed does not allow the researcher to identify a time-varying treatment effect in general. *Borusyak* and *Jaravel* (2017) discuss the under-identification problem of a FE model specifying a dummy impact function with a full set of leads and lags of the treatment variable (a "fully dynamic specification"). They show that it is not possible to disentangle the effects of calendar time (or age) and time elapsed after treatment due to collinearity. It certainly would help to restrict the number of dummy variables for K_{it}, notably to include only post-treatment dummies. However, with all persons eventually experiencing the event, we inevitably run into collinearity problems towards the end of the observation window (especially if all persons in the sample are treated for a longer period of time). Therefore, we should always include a control group of never-treated units in our FE estimation sample to be able to identify continuous or dummy impact functions.

Another important issue is the correct specification of the functional form of the outcome trajectory. Misspecification of the age effects would imply misrepresented potential outcome trajectories, which in turn biases our estimates of the treatment effect (the so-called "bias transfer", see *Ranjbar* and *Sperlich* (2020) for formal proofs). Therefore, we should always take care regarding the underlying functional form assumptions of age effects. It is seldomly a good idea to simply specify a linear

age effect without further checks for linearity. Generally, we recommend including age as dummies in the model, since outcome trajectories often tend to be highly non-linear. A more parsimonious specification should only be chosen if the resulting dummy coefficients show a regular pattern.

Besides controlling for age effects, it may also be necessary to control for period (P) effects. For example, happiness in a given population might change not only due to health deterioration (biological ageing), but also due to economic/political cycles. Including (unrestricted) period effects, however, produces the so-called age-period-cohort (APC) problem, because the three variables are linearly dependent (see *Kratz* and *Brüderl* (2021) for more details). To separate A and P effects, we need to introduce parameter restrictions. A minimal restriction needed for identification would be, for instance, that two period effects are equal. One should always think carefully about these restrictions, as they often exert strong impacts on the results (as well as on the treatment effect). It is not a good idea to simply let one's software decide. For example, Stata and R would exclude the first and the last period dummy, thus assuming the period effects are the same in the first and last year of the observation window. This assumption clearly is not met if there is some (positive or negative) trend over calendar time. A better idea might be to exclude the first two period dummies, or to group several years into a smaller set of period dummies (assuming piecewise constant effects over calendar time). Since all these restrictions are not testable, the more recent literature favors a "proxy approach", whereby the period effects are proxied by variables that measure the economic/political cycle directly (e.g., GDP growth rates; see *Kratz* and *Brüderl* (2021) for more details).

Finally, we want to stress that in most demographic applications, age effects are more important than period effects. Maturation is ubiquitous in demographic outcomes and age is inherently and systematically related to treatment. It is unlikely that period effects are related in a similar systematic fashion to treatment. Therefore, one should include a full control set for the age effects (i.e., age dummies). Period effects may be included in a restricted/proxied version only.

The widely used two-way FE model does not follow this recommendation: it prioritizes period/wave effects over age effects. This model includes a full set of period/wave dummies (i.e., period fixed effects alongside the unit fixed effects; hence the name), but no age controls:

$$Y_{it} = \beta D_{it} + \boldsymbol{X}'_{it}\boldsymbol{\delta} + \gamma_t + \alpha_i + \varepsilon_{it}, \tag{3}$$

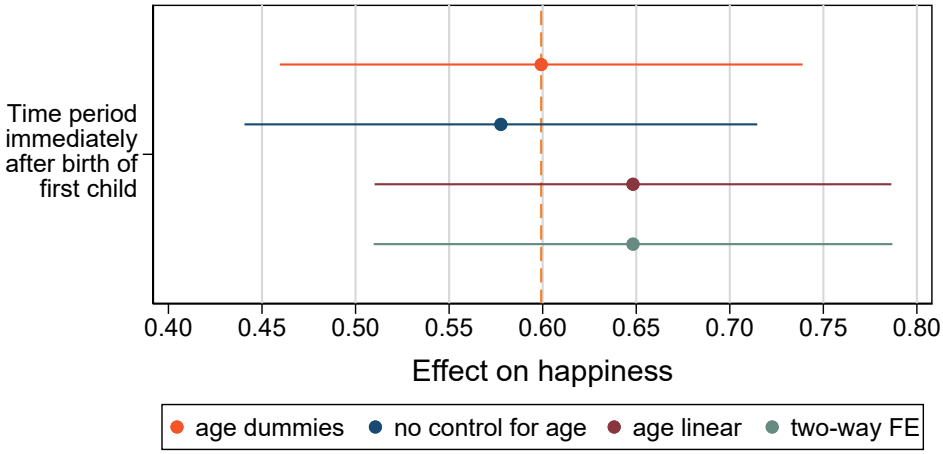
where γ_t represents the period fixed effects and X_{it} does not include age. Given this specification, the period fixed effects will capture a mixture of period and age effects. This very likely will result in a misspecification of the outcome trajectory and consequently in a biased treatment effect estimation. Thus, we warn against the "default use" of the two-way FE model.[4]

[4] Researchers often inadvertently specify a two-way FE model by including a full set of wave dummies in their (single-way) FE model.

Continuing with our childbirth example, we will illustrate some of the points made in this section. First, we want to note that our estimation sample includes an appropriate control group of never-treated women: 2,477 women did not give birth to a first child while observed in the pairfam study. These observations do not contribute to the estimation of the birth effect, but they contribute to the estimation of the age effect (and the effects of the other control variables[5]).

Second, we address the importance of specifying the functional form of the age effect correctly. As *Kratz* and *Brüderl* (2021) show, the age-happiness trajectory in Germany declines with age. The decline is slow between ages 18 and 65, and becomes steep afterwards. Thus, not controlling for age generally would lead to downward biased treatment effects. Since the decline is slow between ages 18 and 65 (about .02 scale points per year), the bias is expected to be moderate in our application. Figure 3 presents the results for the 0-year dummy for differently specified models. The first one ("age dummies") is the reference model. It includes a full set of age dummies (ages 16-47, 15 being the reference age). This is the

Fig. 3: The effect of the birth of a first child on mothers' happiness: varying the specification of the age effect



Notes: Regression coefficients for the 0-year dummy including 95 percent confidence intervals from four FE regression models with varying specifications of the age effect. Further controls are relationship status dummies, subjective health dummies, and a dummy for pregnancy. The first model is our reference model. See Appendix, Table A2 for numerical results.

Source: pairfam release 11.0, own calculations

5 Age is not the only important control variable. In our case, relationship status is also important. The effect of the 0-year dummy increases to .64 if we drop the relationship dummies from the model. The reason is that birth commonly is related to a partnership, and being partnered increases happiness strongly. This demonstrates that impact functions – as FE estimators generally – may be severely biased if important time-varying confounders are not controlled for.

specification used in the previous section, since age-happiness trajectories tend to be highly non-linear (see *Kratz/Brüderl* 2021). Thus, the reference birth effect immediately after birth is .60 scale points. The second model shows what happens if we do not control for age at all. Figure 3 shows that the effect is biased downwards to .58. As expected, this a moderate downward bias. The third model adds age as a linear control. We see that this is not a good idea in our case, as now there is an upward bias to .65. This is most likely some kind of over-control bias, where the linear outcome trajectory over-estimates the happiness decline over the main childbearing years. This demonstrates that the correct modelling of the functional form of the outcome trajectory is important.

Third, we will show how large the bias is by using the two-way FE model. The fourth coefficient in Figure 3 gives the result, if we include a full set of wave dummies instead of age dummies. Once again, we observe an upward bias to .65. In addition, the two-way FE model finds further significant positive birth effects in the second and fourth years after birth. This indicates that the wave dummies are not able to control for the full age effect.[6]

Pre-treatment and late post-treatment dummies

As argued above, it is important to model the impact function as flexibly as theory demands. We have explained how to model dummy impact functions by including indicators for discrete time after the exposition to a treatment. Theory often suggests that demographic events are anticipated by individuals, and thereby already exert a causal effect before treatment (anticipation effect). For instance, in the childbirth example, it is reasonable to assume that the mere expectation of a baby may already boost happiness. So, we might include a pre-treatment dummy indicating a person-year within the period of nine months prior to birth to capture anticipation during pregnancy. In fact, it is even advisable to do so: Otherwise, a positive anticipation effect would be incorrectly attributed to the life course of the reference period (comprising all person-years before birth), and any positive FE impact function would be biased downwards.

Note that by including pre-treatment dummies, the reference period for the treatment effect changes. It is no longer all person-years before treatment, but only person-years before the first pre-treatment dummy. For instance, if we include a -1-year dummy, the reference period is $k < -1$.

However, we must be particularly careful when specifying further "leads" of the treatment dummy. We should not mindlessly include pre-treatment dummies as regressors. The problem is that pre-treatment dummies may capture not only an anticipation effect, which is usually interpreted causally, but also various other effects. First, there may be time-varying confounders that are omitted from our specification. For example, the positive effects of forming a new relationship or

6 Period effects do not seem to be very relevant in our observation window. Including a dummy for Wave 1 (to capture a potential effect due to the financial crisis 2008/09) did not change results. Therefore, we do not include any period effect in our preferred specification.

household which usually occur prior to pregnancy may be confused with the anticipation of the child. (We should of course control for such confounders directly whenever it is possible.) Second, pre-treatment dummies might capture reverse causality, and feedback effects in particular: a positive "shock" to life satisfaction may lead people to decide to have a child. Third, self-selection into treatment may be related to growth of the outcome (see below). For instance, people who are on a steeper happiness trajectory throughout their lives might be more likely to eventually have children.

Thus, the effects of pre-treatment dummies usually are not interpretable. It is often unclear whether they are driven by anticipation, omitted variable bias, reverse causality, or selection on growth. Only if we have a clear theoretical argument for anticipation (and can also exclude the other sources mentioned) and can pin the effect down by a precise measure (e.g., respondents answering that they are pregnant) should we include a pre-treatment dummy. Otherwise, pre-treatment dummies may capture spurious effects and thereby also bias the estimated treatment effects.

A further important decision when modelling dummy impact functions in particular, concerns how to handle late post-treatment dummies. With staggered treatments, early-treated persons potentially could still be observed long after treatment. However, due to panel attrition, we usually will have only few observations for late post-treatment dummies. Thus, the question of what we should do with these late post-treatment dummies arises. It is common practice in the demographic literature to group all post-treatment dummies into a residual dummy after a certain time point. Although this seems like it may be an innocuous decision, it can cause a bias of the treatment effects.

As we noted earlier, assuming a step impact function can cause a negative weighting bias. This reasoning also applies to the dummy impact function with a cap time point. After the cap time point, we assume a step function. This assumption will be violated if there are time-varying treatment effects after the cap time-point. Basically, the functional form of the impact function is mis-specified. Then, negative weighting might bias the estimate for the residual dummy, and due to bias transfer, all coefficients of the dummy impact function may be biased (*Borusyak/Jaravel* 2017).[7] It would thus be wrong to expect that grouping late treatment periods leaves the estimates of earlier post-treatment dummies unaffected in general. Rather than grouping late treatment periods in one dummy, we therefore recommend truncating the data, i.e., excluding the late periods from the sample. Clearly, paying attention to late post-treatment periods is particularly important if results of a flexible impact function suggest strong treatment effects for these periods.

Once again using our childbirth example, we will illustrate some of the points made in this section. In our application, an anticipation effect during pregnancy is highly likely: most pregnant women will be much happier in anticipation of the child

[7] A toy example with fictitious data demonstrating this bias is given in Figure A1 in the Appendix. In this example, the true impact function is increasing. Grouping post-treatment dummies biases the impact function downwards. The earlier the cap time point is, the larger the bias.

(especially if it is an intended pregnancy). In addition, a self-reported pregnancy indicator is available in pairfam. Therefore, we controlled for a pregnancy dummy in all models. The effect of a pregnancy in our reference model is large, at .61 scale points (see Table A1). This means that the pregnancy effect is as strong as the baby effect (which is .60 in the reference model)! Note that by including a pregnancy dummy, the reference period changes: The baby effect is in reference to all person-years before birth, except person-years in pregnancy.

What happens with the baby effect if we choose different specifications? Figure 4 provides answers to this question. The first model is our reference model controlling for pregnancy. The second model does not control for pregnancy. As expected, the baby effect is strongly biased downwards to .45, because the anticipation effect now is in the reference period before birth. In our application, it would thus be a specification error not to include the pregnancy dummy.

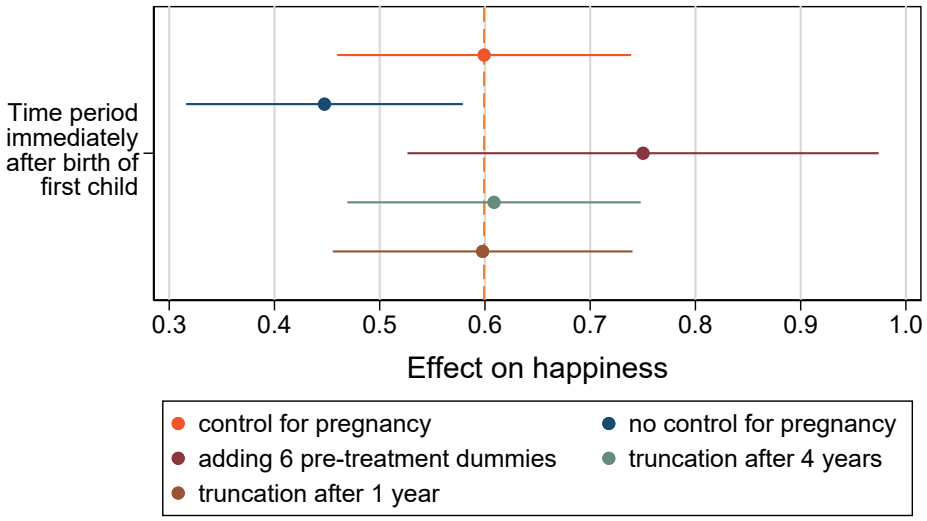
However, in most applications, we will not have a direct measure of anticipation (i.e., pregnancy). In such cases, researchers often simply include pre-treatment dummies. We apply this procedure in the third model, where we include six pre-treatment dummies up to $k = -6$, as is often done in the literature. The -1-year dummy may then capture the pregnancy effect (though imprecisely). The other pre-treatment dummies may capture other mechanisms, as argued above. To demonstrate this, we do not control for relationship status and health in this model. As can be seen in Figure 4, the result is a strong upward bias of the baby effect to .75. This is expected, since now the reference period is $k < -6$, long before birth, and when women may have had no partner and perhaps were in bad health. In addition, the confidence interval is now much larger, because there are much fewer reference observations. Thus, we strongly advise against using pre-treatment dummies without good reason.

Finally, the last two models address the grouping issue. So far, we grouped all late post-treatment observations into a single group of $k >= 4$. As argued above, this is not optimal. However, there might be no problem in our case, because all treatment effects after the first year are zero. So, the grouping has, in fact, no consequence, as shown with the estimate of the fourth model that excludes all person-years observed later than 4 years after childbirth. Generally, a sensible strategy to identify a short-lived treatment effect might be to truncate the post-treatment observation period after one year. As can be seen in Model 5 in Figure 4, we obtain basically the same result as in our reference model containing more post-treatment dummies.

Biased Impact Functions due to Heterogeneous Trajectories

Estimates of impact functions may be biased in the case of heterogeneous outcome trajectories that are related to the treatment process. In practice, it may therefore be necessary to allow for heterogeneous life courses when modelling an impact function. One way to do so would be to interact process time or age with a time-constant indicator of the treatment group (FE with group-specific slopes, FEGS). An even more flexible way would be to allow for individual-specific trajectories (FE with individual-specific slopes, FEIS) (*Brüderl/Ludwig* 2015; *Rüttenauer/Ludwig* 2020; *Wooldridge* 2010).

Fig. 4: The effect of the birth of a first child on mothers' happiness: varying the specification of pre- or post-treatment effects



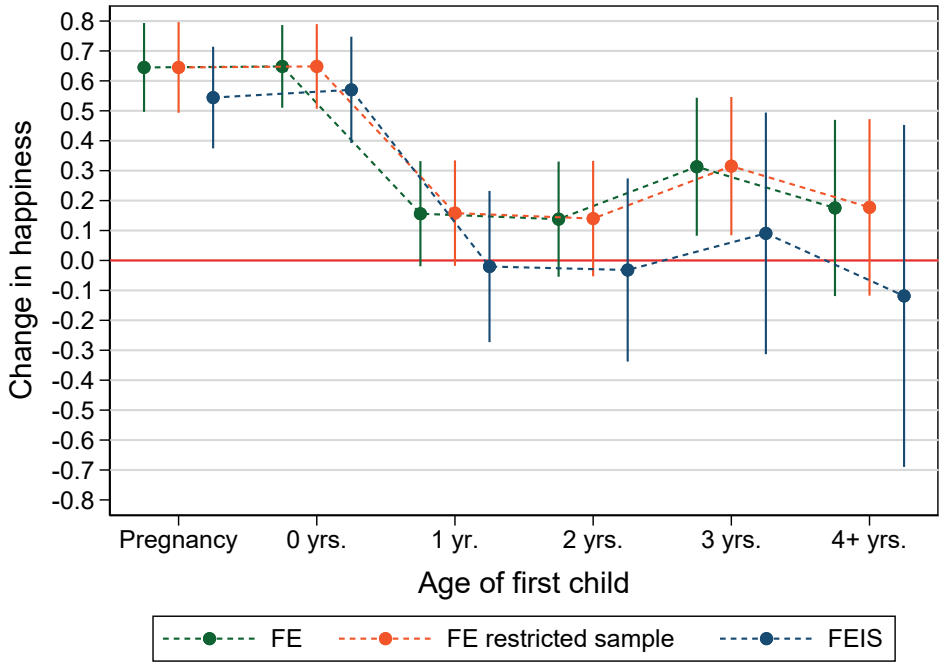
Notes: Regression coefficients for the 0-year dummy including 95 percent confidence intervals from five FE regression models with varying specifications of pre- or post-treatment effects. The first model is our reference model. See Appendix, Table A3 for numerical results.

Source: pairfam release 11.0, own calculations

Including individual-specific slopes for process time, age, or some other variable is a straightforward extension of the standard FE model. Modelling individual slopes explicitly relaxes the parallel trends assumption that is crucial for the consistency of FE estimates. Hence, FEIS guards against bias due to time-constant confounders related to differences in the outcome trajectories between treatment groups. An extended Hausman test can be used to decide whether FE estimates are biased (*Rüttenauer/Ludwig* 2020).

Next, let us extend our empirical example on childbirth and happiness in that we allow for heterogeneous age-happiness trajectories for the women included in our pairfam sample. First, we must decide on a functional form for the individual slopes. We cannot use a dummy specification for the age effect. The reason for this is simple: The FEIS model builds on variation in the data that is left over after controlling for the individual age effect. Technically, we therefore estimate the age effect for each person (by OLS), and to do so we need more observations than parameters for each person. Hence, we must use a parsimonious model specifying few parameters to estimate the individual slopes. Here, we simply allow for an individual constant and specify a linear effect of age. Thus, we estimate two individual-specific parameters (a constant and a slope). Therefore, FEIS estimation requires at least three observations per individual. For this reason, we are left with a restricted sample including 2,525 women, i.e., 84 percent of our previous sample. Note that these

Fig. 5: Effects of childbirth on happiness in models specifying homogeneous or heterogeneous age effects



Notes: Regression coefficients for "pregnancy" and "years since birth dummies" including 95 percent confidence intervals from FE and FEIS regression models. Controls are age (linear), relationship status dummies and subjective health dummies. FE models are run on a full estimation sample and a restricted sample (only women with 3+ person-years). The FEIS model uses the restricted sample. The FEIS model specifies age as a variable with individual-specific slopes. See Appendix, Table A4 for numerical results.

Source: pairfam release 11.0, own calculations

women nevertheless provide 95 percent of the observations of the previous sample, because we excluded only women with very few observations. Although we usually would hesitate to exclude such a large number of units from our estimation sample, it may actually not do much harm in terms of "representativeness" if viewed from the perspective of sampling life courses.

Figure 5 depicts estimated coefficients for the pregnancy and age of first child dummies from three models: a FE model using the full estimation sample, a FE model using the restricted sample, and a FEIS model (on the restricted sample). The restricted FE estimates indeed reproduce the unrestricted FE estimates almost perfectly. The FEIS effects are somewhat smaller, but still very close to the full FE

model.[8] This is also the result of formal testing: The extended Hausman test tells us that the coefficients of the restricted FE and the FEIS model are not significantly different on the 5 percent significance level (χ^2 (13) = 21.80; p = 0.0586). The conclusion here would thus be that there is no substantial bias of the impact function due to heterogeneous age profiles (see *Gattig* and *Minkus* 2021 in this Special Issue for an analysis of the impact of marriage on happiness arriving at a similar conclusion). However, in other empirical applications, a bias may show up. We therefore recommend to at least conduct robustness analyses based on the FEIS model. Routines for Stata and R (xtfeis and feisr) are available for estimation and specification tests (*Ludwig* 2019; *Rüttenauer/Ludwig* 2021).

Consecutive Life Events

Another problem for the estimation of treatment effects may result from consecutive life events with heterogeneous effects. To give an example: Causal effects of childbirth may depend on parity, such that the effect of the first child differs from the effect of a second child. At first sight, this setting does not introduce any difficulties. It is common to use a sample of initially childless persons (as we recommended above) and proceed with a panel regression, including two separate dummy variables for person-years after the birth of a first and second child (see, for an example, *Abendroth et al*. 2014). We call this the "single estimation sample strategy". But this standard procedure may be problematic if the effects of early treatments are themselves heterogeneous, and selection into later treatments depends on the magnitude of earlier treatment effects. In our example, it might be that people who get higher returns for first childbirth are more likely to have a second child than people without positive effects after their first child. In this case, estimates of both treatment effects will be biased.

To clarify the issue: From a counterfactual perspective, we might distinguish between three different treatment effects in this situation (each built from differences between two out of three potential outcome trajectories):

E1: Effect of the 1st child for persons having only a 1st child compared to those always childless

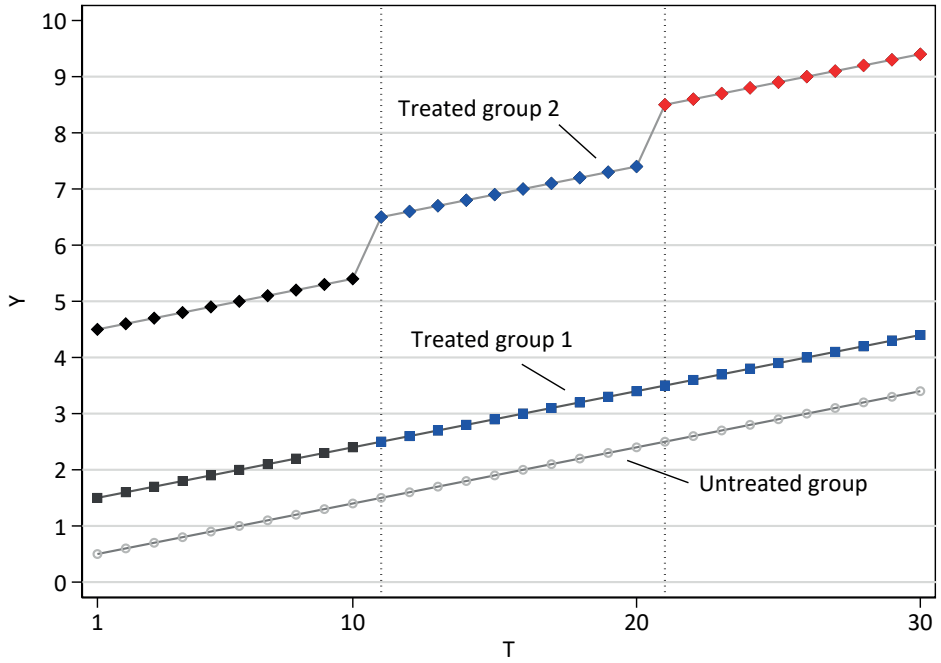
E2: Effect of the 1st child for persons having a 1st and a 2nd child compared to those always childless

E3: Effect of 2nd child for persons having a 1st and a 2nd child compared to persons having only a 1st child (Here, we assume homogeneous effects.)

Using the single estimation sample strategy assumes that effects E1 and E2 are identical. This assumption often remains tacit and untested, but it may easily be wrong in practice. Effect E1 might be zero, which might be the reason that these

[8] Note that the baby effect in the FE models is overestimated here, because we use only a linear age control. In fact, the FE model with the full sample is identical to the model "Age linear" in Table A2/Figure 2. Note also that the coefficient of the three-year dummy is significantly larger than 0. The FEIS model seems to "correct" these overestimations.

Fig. 6: A toy example with fictitious data of consecutive life events



Notes: Treated Group 2 receives a first treatment effect of 1 at $t = 11$ and a second treatment effect of 1 at $t = 21$. Treated Group 1 receives a first treatment effect of 0 at $t = 11$ and no second treatment.

Source: fictitious data, own calculations

particular people decide against having a second child, while effect E2 is positive. In this situation, the estimate of the dummy for the second child will not provide E3. It will be biased (upwards) due to the heterogeneity in E1 and E2. As a consequence, the average treatment effect (ATE, the average of E1 and E2) for the first child will be biased (downwards).

We constructed a toy example to illustrate these biases. In the data shown in Figure 6, we have two treatment groups (of equal size) with different effects for a first treatment: Group 1 has a first treatment effect of 0 (E1), Group 2 of 1 (E2). Thus, the true ATE is 0.5. Now, only Group 2 receives a second treatment later on ($t = 21$), the second treatment effect being 1 (E3). The single estimation sample strategy introduces a dummy for treatment one (=1 for the blue and red person-years) and a second dummy for treatment two (=1 for the red person-years only). What are the results if we apply this standard strategy to the toy data from Figure 6? The FE estimate (two-way FE model) for treatment dummy one is 0.4. This is close to the true ATE, but with a small downward bias. This is due to a negative weighting bias (group one enters FE estimation with a larger weight). The FE estimate for treatment dummy two is 1.4. This is an overestimation, because the first treatment effect was underestimated.

We see that the single estimation sample strategy may provide biased treatment estimates. To recover the true ATE for treatment one, we would have to exclude all observations after the second treatment, i.e., truncate the data at $t = 20$ for all panels in the sample. In practice, however, this will not be possible with staggered treatments. Nevertheless, it would be an improvement to truncate at least those panels that receive a second treatment. Then one could include interaction effects to allow for effect heterogeneity. In our toy example, the true effects for both treated groups can be recovered by including an interaction of treated group and the dummy for first treatment.

Our advice for dealing with consecutive life events (complex life-courses) is to focus on specific transitions and to construct estimation samples accordingly. It is not advisable to estimate all treatment effects from a single estimation sample. Instead, we suggest using multiple, specially tailored estimation samples. We recommend first defining the causal effect of interest based on comparing potential outcomes. Next, construct an appropriate sample to identify the effect. Finally, specify the correct statistical model to estimate the effect.

Our recommendation to work from tailored estimation samples assumes rich data that allow for the construction of treatment groups that are sufficiently large to be able to produce reliable estimates of the causal effects of interest. It might turn out, however, that effect heterogeneity induces only negligible biases. In that case, the main analysis could still use a more standard approach (estimation in a sample of various pooled treatment groups).[9] Nevertheless, we should at least check. Moreover, the focus on specific treatment effects with reduced samples has the major advantage that it is straightforward to allow for time-varying effects. We can easily replace the step impact function with a dummy impact function, for instance.

6 Negative weighting bias with staggered treatment events

In practice, when estimating an impact function from person or household panel data, we often base our estimates on a staggered treatment design: persons receive the treatment at different points in time. Variation in the age of first childbirth is one important example of staggered treatments in demographic research. As recent papers show, using data with staggered treatments and specifying a step impact function when the true treatment effect is time-varying may result in biased estimates (*de Chaisemartin/D'Haultfœuille* 2020; *Goodman-Bacon* 2018). This problem is known as the "negative weighting bias" of FE estimates.

Negative weighting occurs with staggered treatment designs because, towards the end of the observation window, units who get the treatment earlier serve as

[9] Additional results with pairfam data showed that there was some heterogeneity for the effect of first childbirth on happiness along the lines suggested here. However, this did not bias our result for the baby effect in any relevant way. Thus, we might have added person-years with a second child to our estimation sample.

controls for units who get the treatment later. Treatment effects of early treated units that occur well after treatment are thus down-weighted by the construction of the FE estimator. As a consequence, FE estimates of the treatment effect will be biased towards the early treatment effects.

To illustrate negative weighting, suppose we observe a fully balanced panel consisting of three groups of persons $g = u, k, l$, where one group remains untreated (u), one receives treatment early in time (k) and one is assigned to the treatment later in time (l). Let $t = t_k$ and $t = t_l$ denote the timing of treatment of groups k and l respectively. We then specify a two-way FE model to estimate the treatment effect, i.e., we allow for person and time fixed-effects as in equation (3), but we do not include any further control variables. In this setting with staggered treatment adoption, whenever the true treatment effect varies over time, the standard FE model with a step impact function estimates a (downward) biased average treatment effect on the treated (ATT) due to negative weighting of the treatment effect during later post-treatment periods.[10]

A toy example with fictitious data for three treatment groups (of equal size) observed at $t = 1, \ldots, 30$ is shown in Figure 7. For all three groups, there is a linear time trend, such that Y increases by 0.1 each period. For the two treated groups, there is also a treatment effect: starting at $t_k = 11$ and $t_l = 21$, respectively, their trend slope is 0.2 (i.e., an additional 0.1).

When estimating a step impact function with these data, most researchers expect that they would get a sample weighted average of the time-varying treatment effects. In our example, we know that the mean of the true treatment effect is $\frac{0 + 0.1 + 0.2 + \cdots + 0.9}{10} = \frac{9 \cdot 0.1}{2} = 0.45$ for group k at times $t = t_k, \ldots, t_{l-1}$, and $\frac{1 + 1.1 + 1.2 + \cdots + 1.9}{10} = 1 + \frac{9 \cdot 0.1}{2} = 1.45$ at $t = t_l, \ldots, T$. It is 0.45 for group l at times $t = t_l, \ldots, T$. Given that groups k and l are equally large and group k provides 10 early and 10 late treated observations, while group l has only 10 late treated observations, it is natural to assume that each of the three average treatment effects should get a weight of 1/3 in the overall average effect. Thus, we might expect an FE estimate of 2/3 * 0.45 + 1/3 * 1.45 = 0.7833. However, estimation of a two-way FE model returns a downward biased coefficient of 0.45. How can this be?

The reason is that FE uses different (and incorrect) weights to build the average estimate. Let $E(\beta_{js})$ denote the true average treatment effect for group $j = k, l$ during treated times s of that group. Furthermore, let p_{js} be the proportion of treated person-years contributed by group j during s to the total number of treated person-years in the sample. Then,

[10] Assuming that a standard parallel trends assumption holds, it can be shown that the standard two-way FE estimator is identical to the weighted sum of Difference-in-Differences (DID) estimates (see *Goodman-Bacon* 2018). The results on the negative weighting bias thus apply to both the FE and the DID estimator. The problem also affects the estimation of FE models with individual-specific slopes (*Meer/West* 2016; *Goodman-Bacon* 2018; *de Chaisemartin/D'Haultfœuille* 2020). However, if we do specify the impact function correctly, FEIS is a viable alternative for dealing with correlated heterogeneous trajectories. *Rüttenauer* and *Ludwig* (2020) show this through simulations.

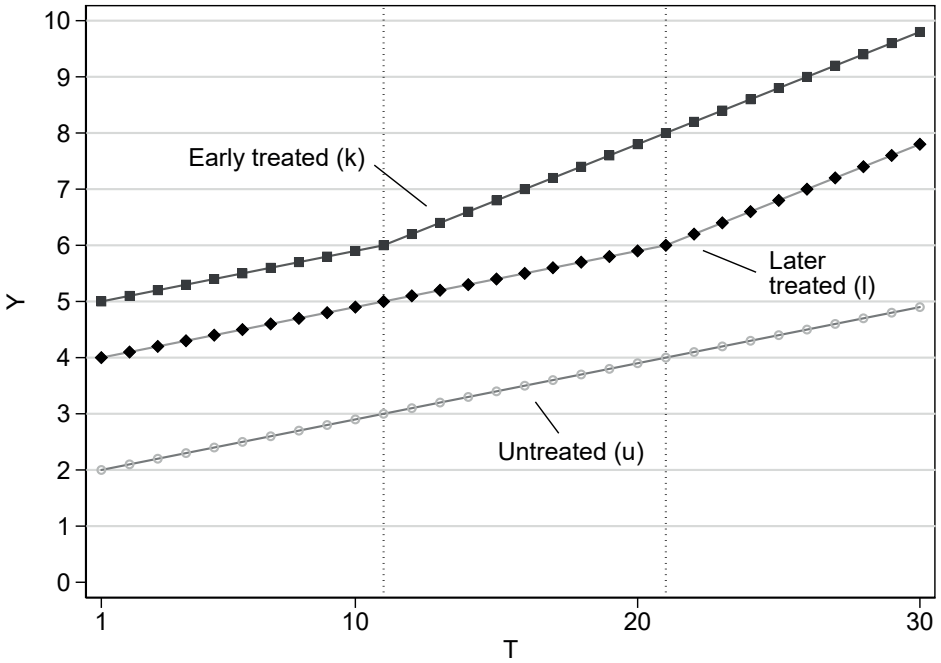
$$\beta_{FE} = \sum\sum w_{js}\, E(\beta_{js}), \tag{4}$$

with weights $w_{js} = \varphi_{js} \cdot \frac{p_{js}}{\sum\sum p_{js}\, \varphi_{js}}$ that sum up to 1.

The crucial parameter that is responsible for negative weighting is φ_{js}. The φ_{js} are the residuals from a two-way FE regression (on the full sample) including unit and period fixed effects with D_{it} as the dependent variable. The second term adjusts the weights for sample size and standardizes them so that they sum up to 1.

In our simple example, the estimated φ_{js} are 1/3 for group k during $t = t_k, \ldots, t_{l-1}$ and 1/3 for group l during $t = t_l, \ldots, T$. The φ_{js} is 0 however for group k during $t = t_l, \ldots, T$. Intuitively, the zero weight results from the fact that these persons are already treated at t_l, and FE puts them (erroneously) in the control group during the time up to T. Technically, their residuals from the regression on D_{it} are zero because, conditional on unit and period effects, there is no variation in D_{it} left over.[11]

Fig. 7: A toy example with staggered treatment design and a time-varying treatment effect



Source: fictitious data, own calculations

[11] In other cases, negative residuals may be obtained for late periods, because with an absorbing treatment indicator, the predicted treatment probability towards the end of the observation window may actually be greater than 1 in a linear model on D_{it}.

Adjusting the weights by sample size and standardizing them, we get weights w_{js} of 1/2, 1/2 and 0. Plugging that into the formula for β_{FE}, we finally get an estimated ATT of 1/2 * 0.45 + 1/2 * 0.45 + 0 * 1.45 = 0.45.

As we have argued earlier, negative weighting of time-varying treatment effects provides a strong incentive to specify a flexible impact function. Simply assuming a time-constant impact and including one treatment dummy often produces a bias in settings with staggered treatments. Contrary to its name, a negative weighting bias often biases the average effect upwards. For instance, often if we have a situation with an early positive, but transitory effect, and (close to) zero effects later on. With FE, late post-treatment effects get weights that are small, thus the overall average effect will often be similar to the short-term effect.

In fact, this occurs in our empirical example of happiness and childbirth as we discussed above on context of Figure 1. We expect an average effect of 0.086. But the step impact function provided an estimate of 0.32. Thus, in our empirical example, we observed an upward (!) bias of the step impact function due to negative weighting (see *Borusyak/Jaravel* 2017 for results in a very similar setting).

Note that the negative weighting problem carries over to heterogeneous effects more generally (*Goodman-Bacon* 2018; *de Chaisemartin/D'Haultfœuille* 2020), as we discussed one example in the context of consecutive life events. It also applies to the estimation of dummy impact functions with inappropriate grouping, as shown earlier in our discussion of grouping post-treatment dummies (*Borusyak/Jaravel* 2017).

Counterfactual Outcomes and the Selection of the Estimation Sample

An important point concerns the deliberate choice of an appropriate sample for the identification and estimation of the parameter(s) of interest. Social researchers usually try to maximize the number of observations with the objective of retaining a "representative" sample. In the framework of a cross-sectional research design, we would like to keep as many units (persons) as possible in the estimation sample. With panel data, the maximizing strategy would imply to include as many units (persons) and measurements per unit (person-years) as possible. However, there are important arguments against maximizing sample size.

In general, the within panel design calls for the restriction of the sample to units with at least two valid measurements over time. Persons with only one person-year, for example, should be dropped because we do not observe useful variation over time for them on any variable. Thus, they do not help in identifying any effect of the covariates.[12]

[12] Units with only one observation – "singletons" – should not be included for FE estimation. Including them will leave point estimates unchanged, because time-demeaning leaves them with zeros on all variables of the model. Also, conventional standard errors do not change. But singletons will affect the estimation of cluster-robust standard errors, which are usually reported in practice. Notably, the cluster-robust standard errors will be too small, because the degrees of freedom correction is incorrect (*Correia* 2015).

Moreover, maximizing the estimation sample may come at the price of biased estimates of the treatment effect. In a typical setting, in which we are interested in the causal effect of a binary absorbing treatment, we would like to restrict the estimation sample to persons who have not yet experienced the treatment event (say, the birth of a child). In fact, the estimate for a time-constant treatment effect (step impact function) must be biased if the true treatment effect is time-varying and we include persons who enter the panel only after treatment (*Sobel* 2012). For those people who do not experience treatment during the observation window, a constant treatment effect is not identified using a within-panel design.

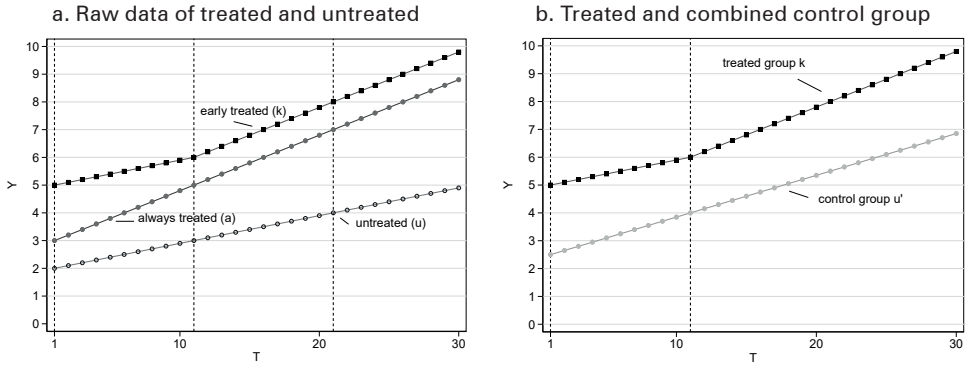
Although people who have already been assigned to treatment before they enter the panel do experience a treatment effect, their value for D_{it} does not change in the data. It is always equal to 1. Hence, the always-treated are subsumed in the control group. Unless we explicitly (and correctly) specify the time-varying treatment effect, it is confounded with the effect of age (or process time). The counterfactual outcome trajectory is misrepresented by the average trajectory of the never-treated and the always-treated. This leads to biased estimates of the age effect or time trend, which in turn biases the estimation of the treatment effect.

Note that this identification problem induces a bias that is slightly different in nature from the negative weighting bias discussed in the preceding subsection. Essentially, including always-treated units introduces two sorts of biases, because these units already serve as controls not only in the post-treatment periods, but also in the pre-treatment periods of other treated groups. We have a negative weighting bias, as before. An additional bias, however, results from misrepresenting the counterfactual trajectory of the control group. Thus, including always-treated units exacerbates the bias of an FE step function estimate with time-varying treatment effects.

To illustrate the bias due to the misrepresented counterfactual trajectory, we modify our toy data (from Fig. 7). We leave out the later-treated group *l*, but include an always-treated group *a*, which receives treatment right at entry into the panel (at $t = 1$). From then onwards, these persons experience the same time trend as groups *u*, *k* and the identical causal effect on the trend as group *k*. The modified data are plotted in Figure 8a. However, due to their earlier treatment, time trend and causal effect are inseparable for always-treated persons. Specifying a step function would assume they are never treated because D_{it} does not change for them. Thus, the FE estimate of the time trend for the control group would be biased. Since the always-treated now also serve as controls during the pre-treatment period of group *k*, a second bias to the treatment effect is introduced in addition to the negative weighting bias that results from the always-treated serving as controls during post-treatment periods of group *k*.

Since group *a* is effectively subsumed in the control group (together with group *u*), the situation is equivalent to the data shown in Figure 8b: besides the trajectory for group *k* we have an average trajectory for groups *a* and *u*, denoted with *u′*. (From the raw data in Figure 8a, we simply calculate the mean of Y at $t = 1$, which is 2.5, and the mean of the slope of the time trend in both groups, which is 0.15.) With these modifications, two-way FE estimates an effect of 0.2. This is much lower

Fig. 8: A toy example illustrating the bias due to inclusion of the always-treated



Source: fictitious data, own calculations

than the true sample weighted ATT, which equals $\frac{0 + 0.1 + 0.2 + \cdots + 1.9}{20} = 0.95$

if we exclude the always-treated and use the true counterfactual trajectory of group *u* (with a slope of 0.1). In fact, this is also what FE returns without the always-treated.

As can be seen in Figure 8b from the different pre-treatment trends of the combined control group *u'* and the treated group *k*, the bias with inclusion of an always-treated group thus effectively comes from a violation of the parallel trends assumption. Using the always-treated as part of the control group results in non-parallel pre-treatment trends, thus violating the strict exogeneity condition of FE. Confounding of the time-varying treatment effect of an always treated group with an identical overall time trend for all groups thus induces a bias of the treatment effect.

As mentioned earlier, we should however include a control group of never-treated units in order to control for age/period effects. Leaving out the group of untreated units also increases the negative weighting bias of FE. In our toy example (Fig. 7), if we exclude group *u* and apply equation (4), we end up with an FE estimate of the ATT of 1 * 0.45 + (-1/2) * 1.45 + 1/2 * 0.45 = -0.05, which is even negative due to the negative weight of -1/2 assigned to group *k* during the final treatment period $t = t_l, \ldots, T$.

In sum, whenever we specify a time-constant impact, we should include untreated units, but we should exclude units that are not at risk of getting treatment because they have already received it earlier. If we expect a time-varying treatment effect, we should of course specify it. Even then, we must use care with the always-treated units. For example, they might belong to older cohorts (which is likely because they received treatment earlier), and the time-varying pattern of the treatment effect may have changed over cohorts (assuming that we use multi-cohort data as with a household panel of the general population). So, we might again end up with a biased age effect and a resulting bias of the treatment effect. In general, we would therefore recommend excluding the always-treated units from the estimation sample, even if we specify a flexible impact function.

7 Conclusion

The estimation of impact functions using panel data and FE models has become common practice in demographic research. In this paper, we addressed important decisions researchers face when modelling such impact functions. We identified some problems and provided hands-on advice to applied researchers for dealing with these problems. Thus, we conclude with a list of recommendations for applied research.

1. Specify the impact function as flexibly as theory demands. In most cases, this will mean that one should use a *dummy impact function*. If the dummy impact function provides a regular pattern, then one could instead use a more parsimonious specification.
2. Specify the *age-outcome trajectory* in a flexible way, i.e., by including age dummies as controls. If necessary, also control for period effects. However, do not use the two-way fixed effects model (i.e., including wave dummies without close consideration).
3. Specify *anticipation effects* only if theory predicts these (and if you have direct measures of these). Do not include *pre-treatment dummies* without close consideration. These may make your results uninterpretable and increase estimation uncertainty.
4. Do not group *late post-treatment dummies*. It is better practice to truncate the post-treatment period.
5. Do not maximize the sample size. Instead, tailor the *estimation sample* according to the research question at hand. Include only units that were not treated when first observed (i.e., drop the always-treated). Do not drop the *control group* of never-treated units.
6. In the case of *consecutive life events*, we advise defining the causal effect of interest based on comparing potential outcomes, focusing on specific transitions, and constructing estimation samples accordingly.
7. It may be necessary to allow for *heterogeneous life courses* when modelling an impact function (e.g., by allowing for individual-specific slopes).

Acknowledgment

We thank two reviewers for their helpful comments.

Data Note

This study uses data from the German Family Panel pairfam, coordinated by Josef Brüderl, Karsten Hank, Johannes Huinink, Bernhard Nauck, Franz Neyer, and Sabine Walper. pairfam is funded as long-term project by the German Research Foundation (DFG). pairfam data can be ordered at https://www.pairfam.de/en/.

Replication files (Stata do-files) are available at *Brüderl* and *Ludwig* (2021).

References

Abendroth, Anja-Kristin; Huffman, Matt L.; Treas, Judith 2014: The Parity Penalty in Life Course Perspective: Motherhood and Occupational Status in 13 European Countries. In: American Sociological Review 79,5: 993-1014. https://doi.org/10.1177/0003122414545986

Allison, Paul D. 2009: Fixed Efects Regression Models. Thousand Oaks, CA: Sage.

Allison, Paul D. 1994: Using Panel Data to Estimate the Effects of Events. In: Sociological Methods and Research 23,2: 174-199. https://doi.org/10.1177/0049124194023002002

Andreß, Hans-Jürgen; Golsch, Katrin; Schmidt, Alexander W. 2013: Applied Panel Data Analysis for Economic and Social Surveys. Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-642-32914-2

Athey, Susan; Imbens, Guido W. 2018: Design-based analysis in diference-in-diferences settings with staggered adoption. NBER Working Paper 24963. https://doi.org/10.3386/w24963

Borusyak, Kirill; Jaravel, Xavier 2017: Revisiting Event Study Designs, with an Application to the Estimation of the Marginal Propensity to Consume. https://dx.doi.org/10.2139/ssrn.2826228

Brüderl, Josef et al. 2020: The German Family Panel (pairfam). GESIS Data Archive. Cologne. ZA5678 Data file Version 11.0.0. https://doi.org/10.4232/pairfam.5678.11.0.0

Brüderl, Josef; Ludwig, Volker 2015: Fixed-Effects Panel Regression. In: *Best, Henning; Wolf, Christof* (Eds.): The Sage handbook of regression analysis and causal inference. London: SAGE: 327-359.

Brüderl, Josef; Ludwig, Volker 2021: Replication Files for 'What You Need to Know When Estimating Impact Functions.' OSF. August 10.

Correia, Sergio 2015: Singletons, Cluster-Robust Standard Errors and Fixed Effects: A Bad Mix. Unpublished Working Paper. Duke University.

de Chaisemartin, Clément; D'Haultfœuille, Xavier 2020: Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. In: American Economic Review 110,9: 2964-2996. https://doi.org/10.1257/aer.20181169

Diener, Ed; Lucas, Richard E.; Scollon, Christie N. 2006: Beyond the Hedonic Treadmill: Revising the adaptation theory of well-being. In: American Psychologist 61:4: 305-314. https://doi.org/10.1037/0003-066X.61.4.305

Gattig, Alexander; Minkus, Lara 2021: Does Marriage Increase Couples' Life Satisfaction? Evidence Using Panel Data and Fixed-effects Individual Slopes. In: Comparative Population Studies 46. https://doi.org/10.12765/CPoS-2021-05

Goodman-Bacon, Andrew 2018: Difference-in-Differences with Variation in Treatment Timing. NBER Working Paper 25018. https://doi.org/10.3386/w25018

Huinink, Johannes et al. 2011: Panel analysis of intimate relationships and family dynamics (pairfam): conceptual framework and design. In: Zeitschrift für Familienforschung – Journal of Family Research 23,1: 77-101.

Kratz, Fabian; Brüderl, Josef 2021: The Age Trajectory of Happiness. In: PsyArXiv. https://doi.org/10.31234/osf.io/d8f2z

Kreyenfeld, Michaela 2021: Causal Modelling in Fertility Research: A Review of the Literature and an Application to a Parental Leave Policy Reform. In: Comparative Population Studies 46: 269-302. https://doi.org/10.12765/CPoS-2021-10

Laporte, Audrey T.; Windmeijer, Frank 2005: Estimation of panel data models with binary indicators when treatment effects are not constant over time. In: Economics Letters 88,3: 389-396.

Ludwig, Volker 2019: XTFEIS: Stata module to estimate linear Fixed-Effects model with Individual-specific Slopes (FEIS) [https://EconPapers.repec.org/RePEc:boc:bocode:s458045, 15.11.2021].

Ludwig, Volker; Brüderl, Josef 2018: Is There a Male Marital Wage Premium? New Evidence from the United States. In: American Sociological Review 83,4: 744-770. https://doi.org/10.1177/0003122418784909

Meer, Jonathan; West, Jeremy 2016: Effects of the Minimum Wage on Employment Dynamics. In: Journal of Human Resources 51,2: 500-522. https://doi.org/10.3368/jhr.51.2.0414-6298R1

Myrskylä, Mikko; Margolis, Rachel 2014: Happiness: Before and After the Kids. In: Demography 51,5: 1843-1866. https://doi.org/10.1007/s13524-014-0321-x

Ranjbar, Setareh; Sperlich, Stefan 2020: A Note on Empirical Studies of Life-Satisfaction: Unhappy with Semiparametrics? In: Journal of Happiness Studies 21: 2193-2212. https://doi.org/10.1007/s10902-019-00165-z

Rüttenauer, Tobias; Ludwig, Volker 2020: Fixed Effects Individual Slopes: Accounting and Testing for Heterogeneous Effects in Panel Data or Other Multilevel Models. In: Sociological Methods and Research. https://doi.org/10.1177/0049124120926211

Rüttenauer, Tobias; Ludwig, Volker 2021: feisr: Estimating Fixed Effects Individual Slopes Models [https://cran.r-project.org/web/packages/feisr/, 15.11.2021].

Sobel, Michael E. 2012: Does Marriage Boost Men's Wages? Identification of Treatment Effects in Fixed Effects Regression Models for Panel Data. In: Journal of the American Statistical Association 107,498: 521-529. https://doi.org/10.1080/01621459.2011.646917

Stutzer, Alois; Frey, Bruno S. 2006: Does marriage make people happy, or do happy people get married? In: The Journal of Socio-Economics 35,2: 326-347. https://doi.org/10.1016/j.socec.2005.11.043

Turek, Konrad; Kalmijn, Matthijs; Leopold, Thomas 2021: The Comparative Panel File: Harmonized Household Panel Surveys from Seven Countries. In: European Sociological Review 37,3: 505-523. https://doi.org/10.1093/esr/jcab006

VanderWeele, Tyler J. 2019: Principles of confounder selection. In: European Journal of Epidemiology 34: 211-219. https://doi.org/10.1007/s10654-019-00494-6

Wooldridge, Jeffrey M. 2010: Econometric Analysis of Cross Section and Panel Data. 2nd ed. Cambridge, MA: MIT Press.

Date of submission: 25.06.2021 *Date of acceptance: 11.08.2021*

Jun. Prof. Dr. Volker Ludwig (✉). TU Kaiserslautern. Department of Social Sciences. Kaiserslautern, Germany.
E-mail: ludwig@sowi.uni-kl.de
URL: https://www.sowi.uni-kl.de/en/sociology/team/ludwig

Prof. Dr. Josef Brüderl. University of Munich, Department of Sociology. Munich, Germany. E-mail: bruederl@lmu.de
URL: https://www.ls3.soziologie.uni-muenchen.de/personen/professor/bruederl_josef/index.html

Appendix

Tab. A1: The effect of the birth of a first child on mothers' happiness: comparing different impact functions

	(1) Step impact	(2) Quadratic impact	(3) Dummy impact (reference model)
Dummy 1st child	0.318***	0.510***	
	(0.066)	(0.070)	
Age of 1st child (years)		-0.248***	
		(0.051)	
Age of 1st child squared		0.025***	
		(0.007)	
Age of 1st child			
0 years			0.599***
			(0.071)
1 year			0.088
			(0.091)
2 years			0.041
			(0.101)
3 years			0.192
			(0.120)
4+ years			-0.009
			(0.152)
Relationship status: Single (ref.)			
Living apart together	0.344***	0.343***	0.342***
	(0.030)	(0.030)	(0.030)
Cohabiting	0.436***	0.423***	0.421***
	(0.043)	(0.043)	(0.043)
Married	0.484***	0.463***	0.467***
	(0.070)	(0.070)	(0.070)
Health status: Bad (ref.)			
Not so good	0.641***	0.642***	0.644***
	(0.107)	(0.107)	(0.107)
Satisfactory	0.951***	0.949***	0.950***
	(0.108)	(0.108)	(0.108)
Good	1.241***	1.239***	1.239***
	(0.109)	(0.109)	(0.109)
Very good	1.520***	1.514***	1.514***
	(0.112)	(0.112)	(0.112)
Dummy pregnancy	0.619***	0.610***	0.607***
	(0.075)	(0.076)	(0.076)
Constant	6.564***	6.555***	6.553***
	(0.129)	(0.129)	(0.129)
Age of resp. dummies included	yes	yes	yes
Within R-squared	0.0696	0.0717	0.0727
N persons	2,982	2,982	2,982
NxT person-years	19,996	19,996	19,996

Regression results from FE regression models, coefficients and panel-robust standard errors (in parentheses).
* p<0.05, ** p<0.01, *** p<0.001

Source: pairfam release 11.0, own calculations

Tab. A2: The effect of the birth of a first child on mothers' happiness: dummy impact function

	Age dummies (reference model)	No control for age	Age linear	Two-way FE
Age of 1st child				
0 years	0.599***	0.578***	0.648***	0.648***
	(0.071)	(0.070)	(0.070)	(0.071)
1 year	0.088	0.061	0.156	0.177*
	(0.091)	(0.088)	(0.090)	(0.090)
2 years	0.041	0.012	0.138	0.152
	(0.101)	(0.096)	(0.098)	(0.098)
3 years	0.192	0.162	0.313**	0.317**
	(0.120)	(0.116)	(0.118)	(0.118)
4+ years	-0.009	-0.054	0.175	0.163
	(0.152)	(0.147)	(0.150)	(0.150)
Relationship status: Single (ref.)				
Living apart together	0.342***	0.293***	0.319***	0.323***
	(0.030)	(0.030)	(0.030)	(0.030)
Cohabiting	0.421***	0.284***	0.382***	0.380***
	(0.043)	(0.040)	(0.042)	(0.042)
Married	0.467***	0.359***	0.499***	0.495***
	(0.070)	(0.065)	(0.069)	(0.069)
Health status: Bad (ref.)				
Not so good	0.644***	0.650***	0.646***	0.648***
	(0.107)	(0.108)	(0.108)	(0.107)
Satisfactory	0.950***	0.948***	0.948***	0.955***
	(0.108)	(0.109)	(0.108)	(0.108)
Good	1.239***	1.246***	1.239***	1.245***
	(0.109)	(0.109)	(0.109)	(0.109)
Very good	1.514***	1.529***	1.515***	1.514***
	(0.112)	(0.113)	(0.112)	(0.112)
Dummy pregnancy	0.607***	0.596***	0.645***	0.642***
	(0.076)	(0.075)	(0.076)	(0.076)
Age of respondent			-0.031***	
			(0.005)	
Constant	6.553***	6.253***	6.964***	6.310***
	(0.129)	(0.107)	(0.153)	(0.109)
Age of resp. dummies included	yes	no	no	no
Survey wave dummies included	no	no	no	yes
Within R-squared	0.0727	0.0632	0.0669	0.0691
N persons	2,982	2,982	2,982	2,982
NxT person-years	19,996	19,996	19,996	19,996

Regression results from FE regression models, coefficients and panel-robust standard errors (in parentheses).
* $p<0.05$, ** $p<0.01$, *** $p<0.001$

Source: pairfam release 11.0, own calculations

Tab. A3: The effect of the birth of a first child on mothers' happiness: varying the specification of pre- or post-treatment effects

	No control for pregnancy	Adding 6 pre-treatment dummies	Truncating after 4 years	Truncating after 1 year
Age of 1st child (post-treatment)				
0 years	0.448***	0.750***	0.601***	0.598***
	(0.067)	(0.114)	(0.071)	(0.073)
1 year	-0.071	0.194	0.088	
	(0.088)	(0.129)	(0.091)	
2 years	-0.121	0.099	0.043	
	(0.096)	(0.141)	(0.100)	
3 years	0.034	0.263	0.190	
	(0.116)	(0.152)	(0.120)	
4+ years	-0.187	0.021	-0.089	
	(0.147)	(0.189)	(0.193)	
Age of 1st child (pre-treatment)				
-1 year		0.460***		
		(0.110)		
-2 years		0.041		
		(0.114)		
-3 years		0.126		
		(0.110)		
-4 years		-0.004		
		(0.117)		
-5 years		0.062		
		(0.113)		
-6 years		-0.142		
		(0.118)		
Dummy pregnancy			0.612***	0.634***
			(0.075)	(0.076)
Relationship status: Single (ref.)				
Living apart together	0.343***		0.344***	0.338***
	(0.030)		(0.030)	(0.030)
Cohabiting	0.439***		0.420***	0.425***
	(0.043)		(0.043)	(0.043)
Married	0.544***		0.475***	0.528***
	(0.069)		(0.070)	(0.072)
Health status: Bad (ref.)				
Not so good	0.648***		0.619***	0.614***
	(0.107)		(0.108)	(0.109)
Satisfactory	0.960***		0.924***	0.929***
	(0.108)		(0.109)	(0.110)
Good	1.247***		1.211***	1.205***
	(0.109)		(0.109)	(0.110)
Very good	1.520***		1.489***	1.486***
	(0.112)		(0.113)	(0.114)

Tab. A3: Continuation

	No control for pregnancy	Adding 6 pre-treatment dummies	Truncating after 4 years	Truncating after 1 year
Constant	6.529***	7.798***	6.580***	6.585***
	(0.129)	(0.071)	(0.129)	(0.130)
Age of resp. dummies included	yes	yes	yes	yes
Within R-square	0.0688	0.0149	0.0724	0.0733
N persons	2,982	2,982	2,982	2,982
NT person-years	19,996	19,996	19,782	18,967

Regression results from FE regression models, coefficients and panel-robust standard errors (in parentheses).
* $p<0.05$, ** $p<0.01$, *** $p<0.001$

Source: pairfam release 11.0, own calculations

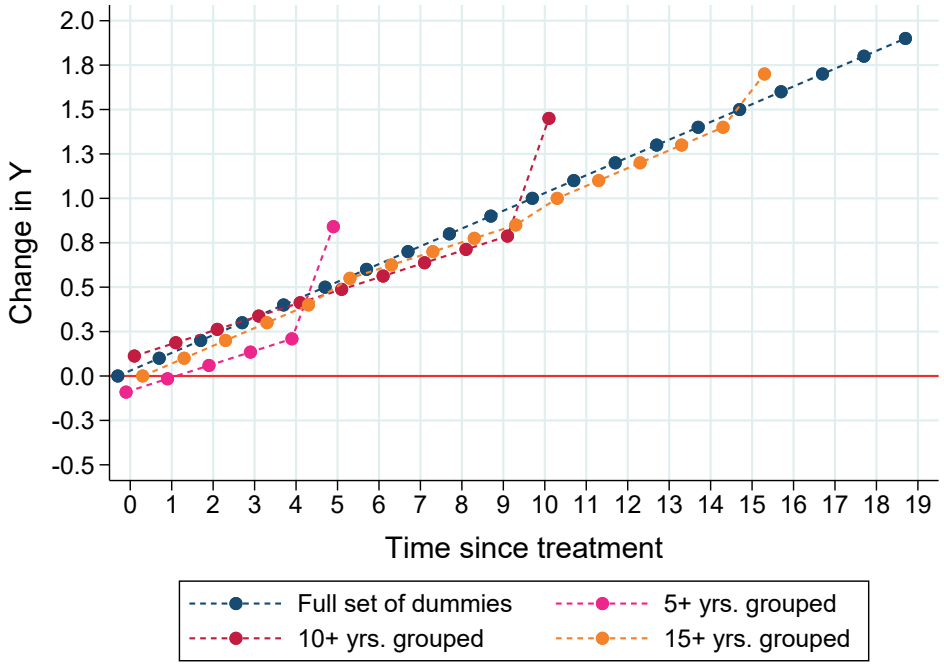
Tab. A4: Effects of childbirth on happiness in models specifying homogeneous or heterogeneous age effects

	FE full sample	FE restricted sample	FEIS restricted sample
Age of 1st child			
0 years	0.648***	0.648***	0.570***
	(0.070)	(0.072)	(0.091)
1 year	0.156	0.158	-0.020
	(0.090)	(0.090)	(0.129)
2 years	0.138	0.140	-0.032
	(0.098)	(0.098)	(0.156)
3 years	0.313**	0.315**	0.091
	(0.118)	(0.118)	(0.206)
4+ years	0.175	0.177	-0.118
	(0.150)	(0.150)	(0.291)
Dummy pregnancy	0.645***	0.645***	0.544***
	(0.076)	(0.077)	(0.087)
Relationship status: Single (ref.)			
Living apart together	0.319***	0.315***	0.348***
	(0.030)	(0.031)	(0.033)
Cohabiting	0.382***	0.383***	0.415***
	(0.042)	(0.042)	(0.049)
Married	0.499***	0.493***	0.474***
	(0.069)	(0.070)	(0.089)
Health status: Bad (ref.)			
Not so good	0.646***	0.625***	0.609***
	(0.108)	(0.109)	(0.111)
Satisfactory	0.948***	0.934***	0.928***
	(0.108)	(0.110)	(0.111)
Good	1.239***	1.224***	1.187***
	(0.109)	(0.111)	(0.112)
Very good	1.515***	1.496***	1.439***
	(0.112)	(0.114)	(0.116)
Age of respondent	-0.031***	-0.031***	Individual-specific slopes
	(0.005)	(0.005)	
Constant	6.964***	6.979***	
	(0.153)	(0.155)	
Within R-square	0.0669	0.0668	0.0615
N persons	2,982	2,526	2,526
NT person-years	19,996	19,084	19,084

Notes: Regression results from FE and FEIS regression models, coefficients and panel-robust standard errors (in parentheses); the restricted sample is restricted to women providing at least three person-years. The first model is identical to the model "Age linear" from Table A2.
* $p<0.05$, ** $p<0.01$, *** $p<0.001$

Source: pairfam release 11.0, own calculations

Fig. A1: Biased impact functions due to grouping late treatment dummies (Toy example with fictitious data)



Notes: The fictitious data are the same as in Figure 7, with a staggered treatment design (one early and one late treated group) and a time-varying treatment effect. The model with a full set of dummy variables exactly hits the true impact function. Other models that instead group treatment periods towards the end of the observation window return biased impact functions.

Source: fictitious data, own calculations

Comparative Population Studies

www.comparativepopulationstudies.de

ISSN: 1869-8980 (Print) – 1869-8999 (Internet)

Published by
Prof. Dr. Norbert F. Schneider

Federal Institute for Population Research
D-65180 Wiesbaden / Germany