MAX-PLANCK-INSTITUT FÜR SOZIALRECHT UND SOZIALPOLITIK
MAX PLANCK INSTITUTE FOR SOCIAL LAW AND SOCIAL POLICY

mea Munich Center for the Economics of Aging

Dangerous Flexible Retirement Reforms – A Supplementary Placebo Analysis across Time

Nicolas Goll

09-2020

MEA DISCUSSION PAPERS

mea – Amalienstr. 33_D-80799 Munich_Phone +49 89 38602-355_Fax +49 89 38602-390_**www.mpisoc.mpg.de**

Dangerous Flexible Retirement Reforms – A Supplementary Placebo Analysis across Time

Nicolas Goll*

Abstract:

In the last decades, many governments have enacted flexible retirement reforms as a seemingly elegant way to increase older workers' labor supply. Börsch-Supan et al. (2018) use the synthetic control method to evaluate the effects of flexibility reforms from nine OECD countries that came into effect between 1992 and 2006. To evaluate the significance of the treatment effects, the authors apply in-space placebo studies. This paper scrutinizes these results by applying in-time placebo studies. Using the time dimension means an artificial reassignment of the flexibility reforms to placebo reform dates other than the actual reform year. The supplementary analysis reveals that the results found in Börsch-Supan et al. (2018) are valid to this robustness check. Overall, the supplementary analysis sustains the result that the reforms have produced zero to negative effects on total labor supply.

Zusammenfassung:

Bestandteil des Renten-Reformprozesses der vergangenen Jahrzehnte war in vielen Ländern die Flexibilisierung des Renteneintritts. Diese Reformmaßnahme wird von einigen Regierungen als scheinbar elegante Alternative zur politisch unpopulären Anhebung der gesetzlichen Rentenalter angesehen und soll dazu dienen das Arbeitskräfteangebot älterer ArbeitnehmerInnen zu erhöhen. Börsch-Supan et al. (2018) verwenden die synthetische Kontrollmethode, um die zwischen 1992 und 2006 in Kraft getretenen Reformen für einen flexiblen Renteneintritt in neun OECD-Ländern zu untersuchen. Um die Signifikanz der berechneten Effekte zu evaluieren, wenden die Autoren *in-space* Placebotests an. Dieses Papier überprüft deren Ergebnisse durch die Anwendung von *in-time* Placebotests. Der zusätzliche Test bringt hervor, dass die ursprünglichen Ergebnisse robust sind. Insgesamt kann also die Schlussfolgerung aufrecht erhalten bleiben, dass die bisherigen Reformen für einen flexiblen Renteneintritt entweder keinen oder einen negativen Effekt auf das Arbeitskräfteangebot hatten.

Keywords: Demographic change, retirement policies, flexible retirement, synthetic control, placebo tests

JEL Classifications: C32, H55, J2

---

* *Contact details:* Munich Center for the Economics of Ageing (MEA) at the Max-Planck-Institute for Social Law and Social Policy, Amalienstr. 33, 80799 Munich. Email: goll@mea.mpisoc.mpg.de.
I am grateful to Tabea Bucher-Koenen and Felizia Hanemann for useful comments. I additionally thank Elisabeth Gruber, Hannah Horton and Christina Meyer for their excellent research assistance.

1. INTRODUCTION

Declining birth rates paired with an increase in life expectancy lead to population aging in many countries around the world. These developments together with the looming retirement of the baby boomers put enormous pressure on pension systems within the next years. To ease the burden of these developments, a common aim of governments in many developed countries has been to strengthen the pool of older workers. However, increasing eligibility ages for drawing pension benefits as one option of keeping older workers in the labor market is not a very popular policy. Therefore, many governments have enacted flexible retirement options that allow workers to gradually reduce work effort with increasing age. The aim of those reforms is to provide a seemingly elegant way to increase older workers' labor supply.

From a theoretical point of view, however, the model of a stylized flexibility reform in Börsch-Supan et al. (2018) shows that flexibility reforms can have ambiguous effects on total labor supply. While flexibilization is likely to increase labor force participation among older workers, it may decrease their working hours. Consequently, the effect on total labor supply is *ex ante* unclear and remains an empirical question. The analysis in Börsch-Supan et al. (2018), therefore, evaluates flexibility reforms from different OECD countries enacted between 1992 and 2006 and finds that they produced zero to negative effects on total labor supply. The conclusion is that flexibility reforms can be regarded as dangerous instruments for two reasons: First, if the aim was to increase older worker's labor supply these reforms have failed to reach their objective. Second, the reforms may have additionally postponed or even replaced the introduction of more effective policies.

In Börsch-Supan et al. (2018), the synthetic control method – proposed by Abadie and Gardeazabal (2003) and extended in Abadie et al. (2010, 2015) – is applied to study the effects of different flexibility reforms. Aim of Börsch-Supan et al. (2018) is to investigate the effect of flexibility reforms on the labor force participation rate (extensive margin), the average number of weekly working hours (intensive margin), and on the total labor volume of men aged 55-64. The latter is the product of intensive and extensive margin. According to Athey and Imbens (2017), the synthetic control method "*is arguably the most important innovation in the policy evaluation literature in the last 15 years*". It builds on the idea of difference-in-differences estimations which have been an important tool for empirical research since the early 1990s (see, for instance, the classic difference-in-differences study by Card 1990). The advantage of the synthetic control method lies in the acknowledgement of the premise that, when the unit(s) of interest are only a few aggregate entities, such as countries, states, regions, or cities, a combination of comparison units (called “synthetic control group”) usually better reproduces the characteristics of the unit(s) of interest than any single comparison unit alone. Hence, the idea of the synthetic control method consists in constructing the synthetic control group as a weighted average of all potential comparison units that best resembles the pre-treatment characteristics of the case of interest (Abadie et al. 2015). The post-treatment outcomes of the treated unit are then compared to the post-treatment outcomes of the untreated synthetic control group. With that, it becomes possible to estimate what would have happened to the treated unit of interest in the absence of a specific treatment or intervention (such as, e.g., event, shock, law, reform).

Abadie et al. (2010) state that the potential applicability of the synthetic control method to comparative case studies is very large since many policy interventions or other events of interest take place at an aggregate level (e.g., countries, states, regions, cities) and affect only a small number of aggregate units. This holds especially in situations where traditional regression methods are not appropriate. In line with this prediction, the synthetic control method has been applied across various fields over the last years.

As Abadie et al. (2010) further note, large sample inference techniques are, however, problematic for comparative case studies when the number of comparison units is small. Already in Abadie and Gardeazabal (2003), the authors therefore proposed placebo studies to perform inference when applying the synthetic control method. The basic idea of placebo studies is to reassign the treatment to members of the control group that were not actually exposed to it. Abadie et al. (2015) refer to this procedure as “in-space placebos”. If many of the placebo effects are as large as the actual effect, it is likely that the actual effect is observed by mere chance. An alternative dimension that can be used for placebo studies is the timing of the treatment: Instead of reassigning the treatment to actually untreated countries, it can also be assigned to points in time when the intervention did not occur. Abadie et al. (2010) refer to this exercise as “in-time placebos”.

To perform inference, Börsch-Supan et al. (2018) applies in-space placebos by reassigning the intervention (i.e. the coming into force of a flexible retirement reform) to other members of the control group which did not experience a reform in the same year. The research aim of this paper is to investigate the same reforms and to scrutinize the former results by making use of the time dimension in the placebo studies. This is achieved by reassigning the flexibility reforms to dates when they did not actually happen (i.e. in-time placebos). The rationale behind this exercise is the following: If I found estimated effects in the analysis of reforms assigned to dates where they did not actually happen that are of similar or larger magnitude than the effects estimated for the actual reform years in Börsch-Supan et al. (2018), the confidence in these results would diminish. In this case, the estimated effects for the actual flexibility reforms would hardly be attributable to these reforms. Put in the opposite way: It increases the confidence of the results found in Börsch-Supan et al. (2018), if the application of the in-time placebo studies does not yield significant effects.

The remainder of the paper is as follows: Section 2 discusses the synthetic control method, highlights requirements and limitations of the method and reviews the application of placebo studies in the field of the synthetic control method in more detail. A description of the data used in this paper is given in Section 3. Section 4 presents the results from the in-time placebo study and recalls the results of Börsch-Supan et al. (2018) to ease a direct comparison of all results. Section 5 summarizes the findings and concludes that the results found in Börsch-Supan et al. (2018) are stable to the robustness check in the form of the in-time placebo studies.

2. THE SYNTHETIC CONTROL METHOD

The synthetic control method is a data-driven approach to estimate treatment effects of policy interventions in comparative case studies (Abadie and Gardeazabal 2003, Abadie et al. 2010, 2015). Similar to a difference-in-differences design, the synthetic control method exploits the difference in treated and untreated units across an event of interest. In the environment of Börsch-Supan et al. (2018) and this paper, the events of interest are the flexibility reforms in the respective OECD countries. However, in contrast to a difference-in-differences approach, the synthetic control method does not give the same weight to all untreated units. Instead, the procedure induces a weighted average of the untreated units that closely matches the pretreatment-trend of the treated unit. Outcomes for this synthetic control are then projected into the post-treatment period using the weights emerging from the pre-treatment trend matching. This projection serves as a counterfactual for the treated unit approximating the outcome that would have been observed in the treated country without the intervention. In the following, the synthetic control model is described in more formal detail following Abadie et al. (2010, 2015).

2.1 The model

The model supposes that there is a sample of $J+1$ units indexed by j which will be countries in the context of this paper. Only the first country $j=1$ is exposed to the treatment or intervention of interest. Here, the event of interest is the adaption of the flexibility reform. The remaining countries $j>1$ constitute the synthetic control group, a reservoir of comparison countries (in the usage of Abadie and co-authors called “donor pool”).

Let D_{jt} further be an indicator that takes the value 1 if the treatment occurred for country j at time t. Then the observed outcome variable Y_{jt} can be defined as the sum of a time-varying treatment effect $\alpha_{jt}D_{jt}$ and the outcome that would have been observed for country j at time t if the reform had not taken place. The latter is the counterfactual expressed as Y_{jt}^N :

$$Y_{jt} = \alpha_{jt}D_{jt} + Y_{jt}^N \tag{1}$$

Suppose that the counterfactual Y_{jt}^{N} is given by

$$Y_{jt}^{N} = \delta_t + \boldsymbol{\theta_t Z_j} + \boldsymbol{\lambda_t \mu_j} + \varepsilon_{jt} \tag{2}$$

where δ_t is an unknown time factor, $\boldsymbol{Z_j}$ a vector of observed covariates (not affected by the treatment) which can be either time-invariant or time-varying, $\boldsymbol{\theta_t}$ a vector of unknown parameters, $\boldsymbol{\lambda_t}$ a vector of unobserved common factors, $\boldsymbol{\mu_j}$ a vector of unknown factor loadings and the error terms ε_{jt} which are unobserved transitory shocks at the country level with zero mean.

The treatment effect a_{jt} is estimated by approximating the counterfactual Y_{1t}^{N} with a weighted combination of untreated countries:

$$\hat{a}_{1t} = Y_{1t} - \sum_{j\geq 2}^{J+1} w_j Y_{jt} \tag{3}$$

for $t \in \{T_0 + 1, \ldots, T\}$, with a (J × 1) vector of weights $W = (w_2, \ldots, w_{J+1})$' with $0 \leq w_j \leq 1$ for $j = 2, \ldots, J + 1$ and $w_2 + \cdots + w_{J+1} = 1$. $T_0 + 1$ is the year of the treatment, and T is the total number of years. Note that choosing a specific value of W is equivalent to choosing a particular synthetic control.

The weights are chosen such that pre-treatment characteristics of the treated country are best resembled by the characteristics of the synthetic control. More formally, suppose that X_1 is a (k × 1) vector with the values of the pre-treatment characteristics of the treated units which should be matched as closely as possible, and let X_0 be the (k × 1) vector collecting the same variables for the units in a synthetic control group. Note that pre-treatments characteristics in X_1 and X_0 may include pre-treatment values of the outcome variable. Consequently, the vector $(X_1 - X_0 W)$ gives the difference between the pre-treatment characteristics of the treated unit and a synthetic control. The weights, $W*$, are selected in a way which minimizes the distance

$$\|X_1 - X_0 W\|\text{v} = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)} \tag{4}$$

with V is some (k × k) positive semidefinite matrix indicating the importance of each predictor.[1]

[1] There are different techniques of obtaining V (see Abadie et al. 2010). Following Abadie et al. 2010 and Abadie and Gardeazabal 2003, in the empirical section in this paper the weights are chosen such that the root mean squared error (RMSE) of the outcome variable is minimized for the pre-treatment periods (see also Abadie and Gardeazabal 2003, Appendix, for more details). This is the same approach as in Börsch-Supan et al. (2018).

Notice that Equation 1 and Equation 2 put together generalize a traditional fixed effect model that is often applied in empirical studies. The traditional fixed effects model can be obtained when imposing that $\lambda_t u_j = \phi_j$ in Equation 2 by assuming that unobserved heterogeneity is time-invariant. The advantage of the synthetic control method over the fixed effect estimation is that it deals with endogeneity stemming from omitted variable bias as it allows unobserved variables in the estimation that vary with time. Moreover, the synthetic control method allows for the presence of a common time trend across countries as well.

After estimating the treatment effect, statistical significance can be determined by running placebo tests and calculating pseudo p-values. Estimating the same model on each untreated unit of the synthetic control yields a distribution of placebo effects. If many of the placebo effects are as large as the actual effect, it is likely that the actual effect is observed by mere chance. Following Galiani and Quistorff (2016), the pseudo p-value can be written as

$$p - value = \frac{\sum_{j\geq 2}^{J+1} I(|\hat{a}_{jt}| \geq |\hat{a}_{1t}|)}{J} \quad (5)$$

with t as given above and $I(|\hat{a}_{jt}| \geq |\hat{a}_{1t}|)$ being an indicator that takes the value one if the inequality in parentheses is fulfilled.[2]

2.2 Requirements and limitations of the synthetic control method

Constructing a synthetic control group requires a careful selection of comparison units: First, the units in the control group should have similar characteristics as the treated unit to avoid interpolation biases. To address this issue, both Börsch-Supan et al. (2018) and this paper use OECD member countries as potential comparison countries only. Moreover, units affected by the same or similar treatments, or units that may have suffered large idiosyncratic shocks to the outcome variable should not be included in the synthetic control group. In the sense of this paper, this means that only those OECD member countries, which have not adopted flexible retirement reforms during the observation period, are potential comparison countries.

Key challenge of the synthetic control method is that it requires a substantial amount of data. This in particular holds true for the number of pre-intervention periods. The reason is that the pre-

[2] Following Galiani and Quistorff (2016), the pseudo p-values in the empirical section are adjusted with the pre-treatment match qualities. Otherwise, p-values could get too conservative. See Footnote 18 in Börsch-Supan et al. (2018) for more details.

treatment match of the synthetic control unit with the treated unit determines the credibility of the synthetic control. Time series data in particular on working hours for older workers are, however, hard to obtain. This especially holds true for data from before the turn of the millennium. Only in recent years, data availability on working hours in the older age groups has increased. Abadie et al. (2015) do not recommend applying the synthetic control method when the number of pre-treatment periods is small unless the pre-treatment fit is good.[3]

2.3 Inference with the synthetic control method: in-space placebos and in-time placebos

Abadie et al. (2015) recall that the use of statistical inference in comparative case studies is problematic due to, for instance, the small-sample nature of the data or the absence of randomization. These difficulties complicate traditional approaches to statistical inference. The synthetic control method, however, allows conducting different falsification exercises akin to permutation tests – termed placebo studies. The idea of these placebo studies is to artificially reassign the intervention either to units that were not exposed to the treatment or to dates prior to the actual intervention. The advantage of these inferential methods is that it can be used for both individual (micro) and aggregate (macro) data, and can even be applied if the number of comparison units in the synthetic control group is small (Abadie et al. 2010).

In their seminal paper on the economic effects of the terrorist conflict in the Basque Country, Abadie and Gardeazabal (2003) already performed a placebo study to assess whether the terrorist conflict truly had an effect on the economic performance or whether it was rather an artifact of a poorly measured synthetic control. They find that after the outbreak of terrorism in the late 1960s GDP per capita declined about ten percentage points compared to a synthetic control region without terrorism. To assess whether the gap in GDP is actually driven by terrorism, they also applied the method to a similar region (Catalonia in this case) which was not exposed to terrorism during the observation period and compared the resulting estimates to the actual ones.

Abadie et al. (2010) extend the idea of placebo studies by applying the placebo treatments to every comparison country in their control group. The authors study the effect of Proposition 99, a large-scale tobacco control program which was implemented in California in 1988. They demonstrate that tobacco consumption fell remarkably in California relative to a comparable synthetic control region following Proposition 99. Their estimates suggest a decline of per-capita cigarette sales of

[3] Abadie et al. (2010) state that time series data for at least ten years before the treatment would be ideal.

about 26 packs that can be attributed to Proposition 99. The control group in their study contains 38 US states that did not introduce formal statewide tobacco control programs or substantially raised the states' cigarette taxes during the observation period. To assess the significance of their estimates, the authors conduct a series of placebo studies by iteratively applying the synthetic control method to every other comparison state in the control group. They show that the estimated effect for California during the study period 1989-2000 is unusually large relative to the distribution of the effects for the states in the synthetic control.

In their most recent work, Abadie et al. (2015) apply the synthetic control method to the 1990 German reunification and investigate the economic impact on West Germany. Austria, the United States, Japan, Switzerland, and The Netherlands constitute the control group. They find that reunification had a negative effect on West German income, with a reduction of per-capita gross domestic product (GDP) of about $1,600 per year on average over the 1990-2003 period. This amounts to approximately 8% of the 1990 baseline level. To evaluate the credibility of their estimates, they for one thing conduct in-space placebos by artificially reassigning the intervention to each member of the control group. In addition, they apply in-time placebos by reassigning the reform to dates when the actual intervention did not occur. As placebo reform year, they use 1975 instead of the actual reunification year 1990.[4] The interpretation of this approach is similar to the in-space placebo approach: If similarly large effects could be obtained when applying the treatment to dates at which it did not occur, the confidence about the existence of an effect would dissipate. They show that their results are robust across both placebo dimensions and a further sensitivity check.[5]

The results of Börsch-Supan et al. (2018) are based on in-space placebos. In other words, it is under investigation if the treatment effects of the flexibility reforms are driven by chance by estimating the same model on each country in the synthetic control group, assuming it was treated

[4] The authors show that their results are similar when reassigning the placebo reform year to 1970 and 1980, respectively.

[5] Though sometimes described as "in-place" or "cross-sectional placebo tests" by other authors, the main methodology of in-space placebos remains the same. I follow the nomenclature of Abadie et al. (2015) regarding "in-space placebos" and "in-time placebos", respectively, in the rest of the paper.

Another sensitivity check proposed by Abadie et al. (2015) is reducing the number of units in the synthetic control to analyze whether the results are sensitive to single units in the synthetic control. The exclusion of Luxembourg later in this study represents such a further sensitivity check (see Section 4 and Appendix A.5).

at the same time. In doing so, we obtain a distribution of placebo effects against which we can evaluate the effect estimated for the reform-treated country.

In contrast to the application of in-space placebos in Börsch-Supan et al. (2018), this paper applies in-time placebos by reassigning the flexibility reforms to placebo reform dates other than the actual reform year. The placebo reform years are set three years earlier than the actual reform years. The choice of three years is the same for all countries and is the result of a compromise: As historical time series data, in particular on working hours for older age groups are hard to obtain, the choice of three years still allows a pre-treatment period with some years. Simultaneously, the determination of three years still ensures a post-placebo-treatment period consisting of three treatment effects until the actual reform effect would influence the effects. Abadie et al. (2015) emphasize that the number of post-treatment years should not be too small in case the treatment effect only emerges gradually after the intervention.

Since Abadie and Gardeazabal (2003), numerous authors have utilized the synthetic control method across many fields over the last years.[6] Among the literature, it appears that in-space placebos are the most widely used base to performing inference (Bilgel and Galle 2015, Bohn et al. 2014, Cavallo et al. 2013, Kleven et al. 2013, Liu 2015, Stearns 2015). A few papers utilize in-time placebos to prove the validity of their estimates (Freire 2018, Saia 2017). However, in-time placebos are clearly in the minority. While other authors have analyzed the effectiveness of the synthetic control method using Monte Carlo simulations (see, e.g., Ferman et al. 2020, Hahn and Shi 2017, Kaul et al. 2015, O'Neill et al 2016, Gobillon and Magnac 2016), it appears that the majority of synthetic control literature focusses on choosing one placebo test, either in-space or in-time. This paper adds to the very few to date who compare both (see beside, Abadie et al. 2015).

[6] For instance in Acemoglu et al. (2016) on political connections, Abadie et al. (2010) (tobacco control program), Abadie et al. (2015) (Germany's reunification), Bilgel and Galle (2015) (organ donations), Bohn et al. (2014) (2007 Legal Arizona Workers Act), Cavallo et al. (2013) (natural disasters), Gobillon and Magnac (2016) (enterprise zones), Hinrichs (2012) (affirmative action bans on college enrolment), Kleven et al. (2013) (taxation of athletes), Kreif et al. (2016) and O'Neill (2016) (health improvement), Liu (2015) (spillover from universities), Nannicini and Billmeier (2011) and Billmeier and Nannicini (2013) (economic growth).

3. DATA AND VARIABLES

The empirical analysis with the synthetic control method requires a large amount of data. In particular, the fact that time series data are not only necessary for the treated countries but also for all control countries increases the required amount of data.[7]

Dependent variables. The main dependent variables are **labor force participation** (extensive margin) and **working hours** (intensive margin) of males for the age groups 55-64. The outcome variable **total labor supply** is obtained by multiplying labor force participation rates and working hours at the country and year level. Annual time series data on labor force participation and working hours are obtained from different sources: the OECD's Employment database, Eurostat, Eurofound, the International Labour Organization (ILO) and from several national statistical agencies (Australian Bureau of Statistics, Statistics Canada, Statistics Finland, Statistics Japan, Central Bureau of Statistics of Norway, Statistics Portugal, Statistics Sweden, UK Data Service, US Bureau of Labor Statistics).

Control variables. From the same sources as the dependent variables, I use the labor force participation rate and the average number of weekly working hours of the young (age group 25-54) as control variables. These variables capture country-specific labor market trends over time. Variables describing the pension system serve as control variables as well: these are the statutory eligibility age at which workers become eligible for full pension benefits regardless of any other qualification and the earliest eligibility age. The latter is defined as the age at which early retirement is possible, mostly with reduced benefits.[8] The data describing the pension system are obtained from the Social Security Administration's *Social Security Programs throughout the World* (1985-2014), OECD's *Pensions at a Glance* (OECD 2011, 2013) and Duval (2003). Data on years of total schooling are from Barro and Lee (2013). GDP per capita and life expectancy at birth are taken from the OECD's database (OECD 2016a, OECD 2016b).

Treated countries and placebo reform years. The countries under investigation in this paper are the ones part of the empirical analysis with the synthetic control method in Börsch-Supan et al. (2018) to enable a meaningful comparison, namely: Australia, Austria, Belgium, Denmark, France,

[7] The data used in this paper are the same as in Börsch-Supan et al. (2018).
[8] See the glossary of Börsch-Supan and Coile (2019).

Germany, and Sweden.[9] For the application of in-time placebo studies, the placebo reform years (PRY) are chosen to be three years earlier compared to the years where the flexibility reforms actually were adapted. This means: Australia: PRY=2002, Austria: PRY=1997, Belgium: PRY=1999, Denmark: PRY=1992, France: PRY=1990, Germany: PRY=1989, and Sweden: PRY=1997.[10]

Control countries. The pool of potential comparison countries comprises the OECD member countries that have not adopted a flexibility reform during the observation period of each country.[11] The actual selection of comparison countries for the synthetic controls for the treated countries is the same as in Börsch-Supan et al. (2018) and differs for each treated country since the (placebo) reform years are different in each country. In addition, data availability determines observation periods. The data set for each treated country and its comparison countries in this paper constitute a balanced panel, meaning a longitudinal data set where all units are observed at the same time periods. Table A.1.1 reports the time series included before and after the placebo reform year by country and by outcome variable. Table A.2.1 and Table A.2.2 show the comparison countries, which constitute the synthetic controls for each treated country, and present the weight of each comparison country in the control group. The two tables in addition show the time periods included in the estimation.

[9] The analysis in Börsch-Supan et al. (2018) started by additionally looking at flexibility reforms in Finland, The Netherlands, and Sweden. However, with the data at hand we were not able to find robust synthetic controls for Finland and The Netherlands for all outcome variables. Regarding Sweden, we only found a proper synthetic control for the outcome variable labor force participation, but not for the outcome variable weekly working hours. Therefore, the analysis on Sweden is restricted to labor force participation.

[10] Table A.1.1 (Appendix A.1) comprises comprehensive details of the flexibility reforms. The table comprises country-specific information on gender-specific statutory eligibility ages for public pension, the age at which the flexible retirement window starts if flexible retirement schemes are available through systems other than the public pension scheme, information on the extent to which the working time must be reduced within the flexible retirement option, information on the income loss compensating financial sources, whether earnings tests apply, and mandatory retirement regulations.

[11] Belgium and Sweden, which had flexibility reforms rather late, have also been included among the untreated countries for the construction of the synthetic control group of countries that were treated early.

4. RESULTS

Finding a good synthetic control country is crucial for the estimation quality. The synthetic country should as closely as possible approximate how the outcome variable of the treated country would have developed without the flexibility reform. This is the case if the synthetic control country provides a counterfactual pre-treatment value of the outcome variable that comes close to the corresponding value of the treated country.

The approach for constructing the synthetic control in this paper is the same as in Börsch-Supan et al. (2018) to allow a clear comparison: For the outcome variables, labor force participation of males aged 55-64 and working hours of males aged 55-64, I use the average of the pre-treatment values of the outcome variables for constructing the synthetic control.[12] In addition, I control for a set of covariates which explain the outcome variables. These covariates are the labor force participation rate of the younger age group (age 25-54) when the outcome variable is labor force participation, the number of working hours of the younger age group (age 25-54) when the outcome variable is working hours, the multiplication of extensive and intensive margin of the younger age group (age 25-54) when the outcome variable is total labor volume, the statutory eligibility age or the earliest eligibility age. Moreover, gross domestic product (GDP, per capita), years of schooling, and life expectancy are control variables.[13] Table A.4.1 and Table A.4.2 show the quality of pre-treatment characteristics by comparing the pre-treatment characteristics of the treated country with that of the synthetic control country. Total labor supply is measured as the product of labor force participation and working hours for those who participated.

The application of the in-time placebo studies in this paper reduces the number of pre-treatment years because the placebo reforms years are reassigned to three years earlier that the actual reform

[12] Alternative specifications are, e.g., controlling for the last pre-treatment value of the outcome variable only or including all lagged outcome values as predictors. I report robustness checks in Appendix A.3 comparing the different possible specifications (i.e. average pre-treatment value vs. last pre-treatment value vs. all pre-treatment values of labor force participation rate). Kaul et al. (2018), however, demonstrate both theoretically and empirically that controlling for all outcome lags causes all other covariates to be irrelevant. Therefore, I do not include all lagged outcome values as predictors. In general, the criterion for selecting the inclusion of control countries and control variables is to minimize the root mean squared error (RMSE).

[13] For the construction of some of the synthetic control countries, I use the years of early retirement as control variable. This variable is measured as the difference between the statutory eligibility age and the earliest eligibility age. The data on schooling are available in five-year increments and, therefore, converted to annual frequency by means of linear interpolation.

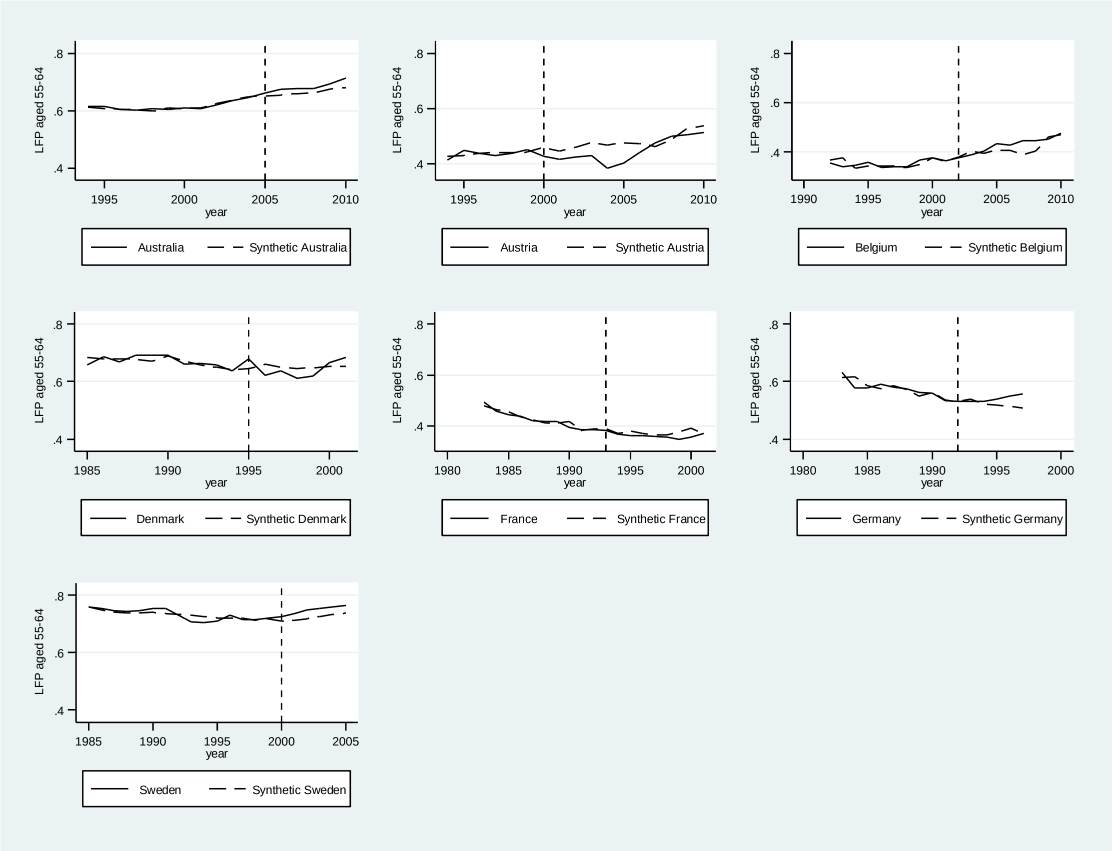
came into effect. However, even with the shorter pre-treatment periods, the estimation yields synthetic controls that come close to the corresponding values of the treated country. In particular, this holds true for the outcome variable labor force participation. The match is slightly less close for working hours and total labor supply. Despite the slightly more vague matches for the latter two outcome variables, the application of the in-time placebo studies overall seems to be a reasonable approach to validate the results of Börsch-Supan et al. (2018). The following sections show the results of the in-time placebo studies, separately for the outcome variables labor force participation, working hours, and total labor supply. Each section initializes with recalling the graphical trends in outcome variables with the actual reform year from Börsch-Supan et al. (2018) to ease the comparison of the in-space placebo studies with the in-time placebo studies.

4.1 Labor force participation

I first examine labor force participation, the extensive margin: Figure 1 shows labor force participation rates of men aged 55-64 for the treated countries and their synthetic counterparts before and after the flexibility reforms. Figure 1 displays the results from Börsch-Supan et al. (2018) with the actual reform year and shows that the labor force participation trend for the synthetic control closely matches the corresponding trend for the treated country before the reform. In Australia, for instance, the synthetic control almost exactly reproduces the actual labor force participation rates during the entire pre-treatment period.

The treatment effect is given by the difference between the outcome variable of the treated country and in its synthetic counterpart after the implementation of the reform. In Figure 1, this is the difference between solid and dashed line right of the reform year. The reform year is indicated by the vertical line. The discrepancy between solid and dashed line is positive for Australia, Belgium, Germany, and Sweden, indicating an increase in labor force participation after the reform. It is negative in France, indicating a decrease in labor force participation in the years after the reform. The picture for Austria and Denmark is mixed.

Figure 1: Trends in labor force participation: treated vs. synthetic control. Actual reform years

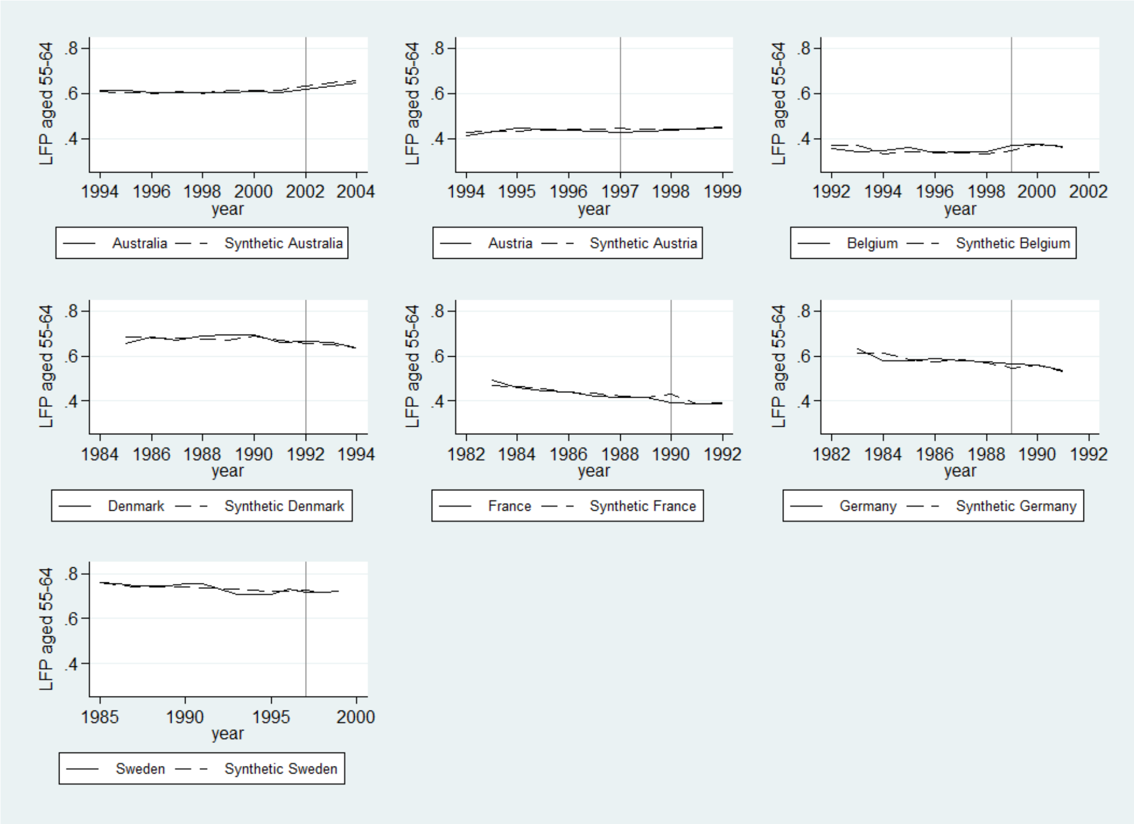


Source: Börsch-Supan et al. (2018).

In Figure 2, the reform year is now reassigned to a placebo reform year, which is three years earlier compared to the actual reform years for all countries and indicated by the vertical line. Figure 2, therefore, shows the results of the "in-time placebo" study with the placebo reform year. Despite the comparably fewer pre-treatment years, the synthetic control comes close to the corresponding value of the treated country.

The most important result, however, is that the trajectories of the actual values of labor force participation and its synthetic counterparts do not diverge considerably in the post-treatment years in the in-time placebo study. This holds true for all countries. This means, in contrast to the results in Börsch-Supan et al. (2018), the in-time placebo studies have no perceivable effects at all. This suggests that the gaps estimated in Figure 1 (with the actual reform year) actually reflect the impact of the flexibility reforms on labor force participation and not a potential lack of predictive power of the synthetic control.

Figure 2: Trends in labor force participation: treated vs. synthetic control. Placebo reform years



Source: Own calculations.

To evaluate the significance of the treatment effects, we calculated pseudo p-values according to Equation 5 (Section 2.1).[14] Table 1 presents yearly treatment effects on labor force participation together with their statistical significance. The first three columns (1) to (3) recall the results of Börsch-Supan et al. (2018) with the actual reform year. Columns (4) to (6) show the yearly treatment effects from the in-time placebo study with the placebo reform years. Pseudo p-values from the in-time placebo study also stem from estimating the model on all countries in the synthetic control group using the placebo reform years, yielding a distribution of placebo effects.

[14] Börsch-Supan et al. (2018) and the analysis in this paper follow the approach of Galiani and Quistorff (2016) in the calculation of pseudo p-values. As placebo effects could be quite large if the quality of matches in the pre-treatment period is poor, Galiani and Quistorff (2016) propose to divide the estimated treatment effects by the corresponding pre-treatment match qualities. Otherwise, p-values could get too conservative. Subsequently, inference is made based on these ratios instead of on the treatment effects only. Following the definition of Galiani and Quistorff (2016), the pseudo p-value in one period is the proportion of placebo pseudo effects (each control unit's treatment effect divided by its pre-treatment root mean square error) that are at least as large as the actual treated unit's pseudo effect.

Table 1 shows a first validation: While the results of Börsch-Supan et al. (2018) reveal significant results for some countries, shifting the reform year suspends significance of effects. This is the case for almost all countries in almost all years (except France in 1990). This means that in contrast to the actual flexibility reforms, the placebo flexibility reforms have no perceivable effect.

Table 1: Post-treatment results regarding LFP of males aged 55-64, effects and pseudo p-values

"In-space placebos" (actual reform year, Börsch-Supan et al. 2018)			**"In-time placebos"** (placebo reform year)		
(1)	(2)	(3)	(4)	(5)	(6)
year	estimates	pseudo p-values	year	estimates	pseudo p-values
		Australia			
			2002	-0.012	0.238
			2003	-0.013	0.190
			2004	-0.009	0.476
2005	**0.010****	**0.048**	2005		
2006	**0.018*****	**0**	2006		
2007	**0.017****	**0.048**	2007		
2008	**0.014***	**0.095**	2008		
2009	**0.019*****	**0**	2009		
2010	**0.032*****	**0**	2010		
		Austria			
			1997	-0.016	0.421
			1998	-0.004	1
			1999	0.006	0.947
2000	-0.03	0.263			
2001	-0.029	0.316			
2002	-0.036	0.263			
2003	-0.048	0.158			
2004	**-0.083***	**0.053**			
2005	-0.073	0.158			
2006	-0.033	0.526			
2007	0.015	0.737			
2008	0.011	0.737			
2009	-0.023	0.474			
2010	-0.023	0.368			
		Belgium			
			1999	0.021	0.235
			2000	0.001	0.941
			2001	0.002	0.882
2002	-0.005	0.824			
2003	-0.013	0.529			
2004	0.008	0.824			
2005	0.027	0.412			
2006	0.021	0.529			
2007	**0.055*****	**0**			
2008	0.041	0.118			
2009	-0.008	0.941			
2010	0.006	0.941			

Denmark					
			1992	0.006	0.824
			1993	0.010	0.647
			1994	0.001	1
1995	**0.034*** **	**0**			
1996	-0.04	0.176			
1997	-0.013	0.412			
1998	-0.035	0.235			
1999	-0.028	0.294			
2000	0.014	0.471			
2001	0.03	0.353			
France					
			1990	**-0.038***	**0.067**
			1991	-0.002	0.933
			1992	-0.008	0.733
1993	-0.006	0.667			
1994	-0.003	0.867			
1995	-0.018	0.333			
1996	-0.007	0.733			
1997	-0.008	0.667			
1998	-0.007	0.6			
1999	-0.029	0.333			
2000	-0.035	0.333			
2001	-0.004	0.867			
Germany					
			1989	0.016	0.769
			1990	0.000	1.000
			1991	0.002	0.923
1992	0.001	0.923			
1993	-0.008	0.846			
1994	0.008	0.923			
1995	0.02	0.538			
1996	0.035	0.385			
1997	0.048	0.231			
Sweden					
			1997	-0.010	0.667
			1998	-0.003	0.933
			1999	-0.002	0.867
2000	0.016	0.533			
2001	0.024	0.2			
2002	0.029	0.133			
2003	0.028	0.2			
2004	0.028	0.333			
2005	0.026	0.333			

Note: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Source: Börsch-Supan et al. (2018) and own calculations.

The significant effect for France might rather stem from the fact that in 1990 the synthetic control matches the actual value least (see Figure 2). The synthetic country exhibits a peak that leads to the largest difference between solid and dashed line in the observation period. Luxembourg and Italy have substantial weight in the construction of the synthetic control country for France. Both countries experienced a peak of labor force participation in 1990 which might drive the value of the synthetic control. For Italy, a weighting factor of 0.257 was assigned and for Luxembourg a weighting factor of 0.413 (see Table A.2.1 on synthetic control weights for the outcome variable labor force participation). In Italy, labor force participation (of the age group 55-64) in 1990 was

53.0% compared to a value of 51.8% in 1989 and 51.4% in 1991. In Luxembourg, labor force participation rate amounted to 43.2% in 1990, while the values in 1989 (38.2%) and 1991 (34.1%) were clearly lower. The significant effect in France in 1990, therefore, might more likely stem from a poor synthetic control in this specific year.

Generally, including Luxembourg in the control group might cause further objection: As Luxembourg is a small country with close labor market ties to, e.g., France, labor market developments may not be completely independent of the developments in the surrounding countries. Therefore, I repeat the analysis as a further robustness check without incorporating Luxembourg in the synthetic control (see Appendix A.5).[15] For the analysis for France, excluding Luxembourg from the control group means two things: First, the peak in the synthetic control in 1990 disappears (Appendix, see Figure A.5.1). Second, the former significant effect turns insignificant (Appendix, see Table A.5.1). The effect, therefore, actually seems to have been driven by Luxembourg's outlier in labor force participation.

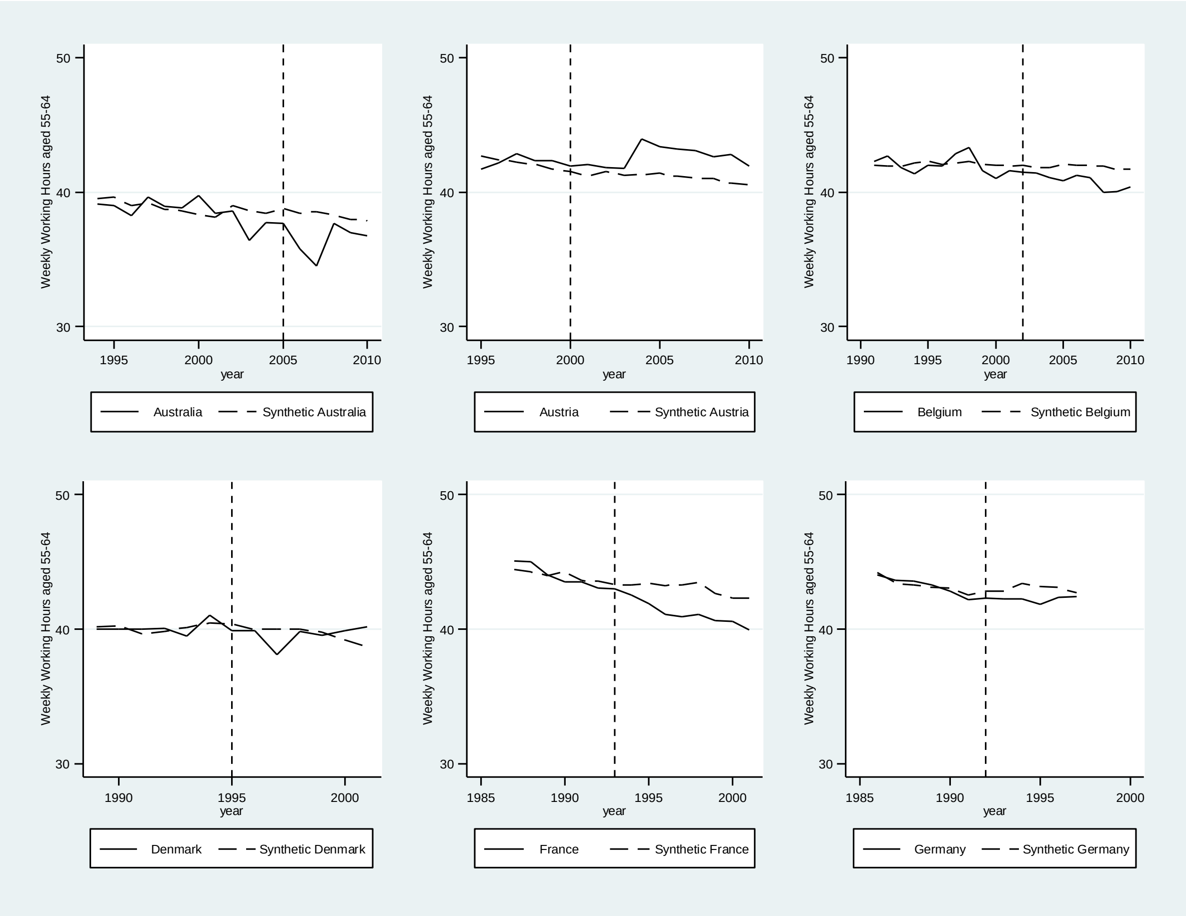
Keeping the results from the robustness check without Luxembourg in the synthetic control group in mind, the in-time placebo study for the outcome variable labor force participation reveals the overall result: reassigning the reform year to earlier dates compared to the actual reform year suggests that the results found in Börsch-Supan et al. (2018) actually reflect the impact of the flexibility reforms. Therefore, the confidence in these results is strengthened.

[15] Reducing the number of units in the synthetic control is a sensitivity check that was proposed and conducted by Abadie et al. (2015).

4.2 Working hours

The second outcome variable is the number of weekly working hours (intensive margin). Figure 3 shows the trends in weekly working hours for men aged 55 to 64 for the treated countries and their synthetic counterparts for the actual reform years (see Börsch-Supan et al. 2018).

Figure 3: Trends in working hours: treated vs. synthetic control. Actual reform years



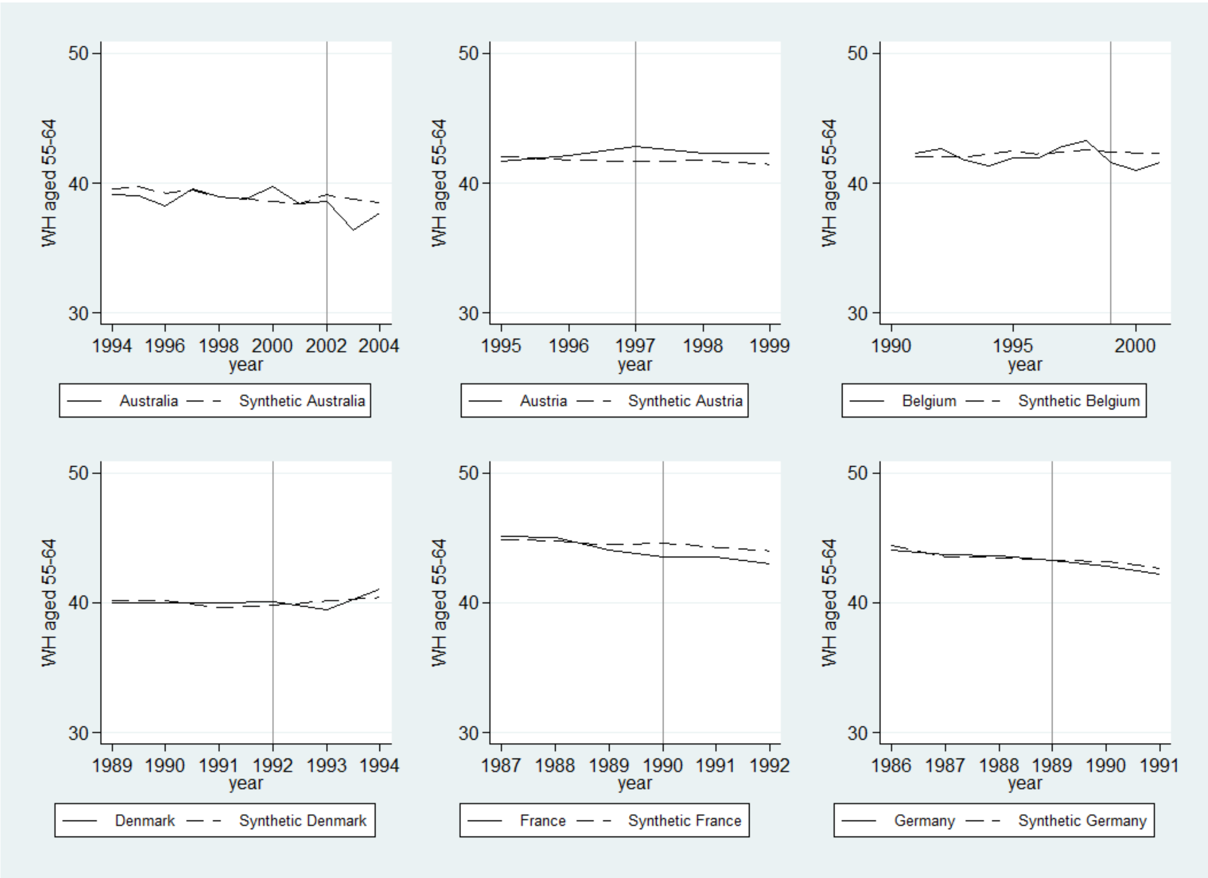
Source: Börsch-Supan et al. (2018).

The figure displays that, due to data restrictions in particular for working hours, the pre-treatment observation periods for working hours are slightly shorter compared to pre-treatment observation periods for labor force participation.[16] Figure 3 hints at negative reform effects on working hours in Australia, Belgium, France and Germany. In Austria, working hours have increased after the reform, while the picture in mixed in Denmark.

[16] Since we could not find a stable synthetic control for working hours in Sweden, we dropped Sweden from the further analysis (see Figure A.2 in Appendix A.7 of Börsch-Supan et al. 2018).

In contrast to the depiction with the actual reform years in Figure 3, Figure 4 shows the results with the placebo reform years. Likewise, the analysis of labor force participation, the placebo reform years were reassigned three years earlier compared to the actual reform year for all countries. This leads to even less pre-treatment years due to data restrictions. Overall, the synthetic controls do not as closely match the corresponding trend for the treated country as for labor force participation. However, in some countries such as Denmark and Germany, the synthetic controls fairly precisely match the actual pre-treatment trend of working hours.

Figure 4: Trends in working hours: treated vs. synthetic control. Placebo reform years



Source: Own calculations.

The size of the treatment effects on working hours and its significance are shown in Table 2. Columns (1) to (3) recall the results from Börsch-Supan et al. (2018) with the actual reform years. The results show that the effects of the reforms on working hours tends to be negative or close to zero for all post-treatment years and all countries except Austria. The positive effects in Austria in 2004, 2005, and 2006 might also stem from other pension reforms which were enacted during the same time.

Columns (4) to (6) display the yearly treatment effects and pseudo p-values from the in-time placebo study. As in the analysis of labor force participation, almost all effects disappear when artificially reassigning the treatment years to placebo reform years three years earlier. The disappearance of effects on working hours again means that the effects found in Börsch-Supan et al. (2018) actually reflect the effects of the flexibility reforms. This interpretation is based, as explained above, on the initial idea of the in-time placebo studies: The confidence about the validity of results dissipates if the estimation procedure of the synthetic control method had also produced large effects when applied to dates where the reforms did not occur. Only in Australia (2003) and France (1990), the estimation procedure yields occasional significant effects.[17] For France, this most likely stems from a poor synthetic control. Due to data availability, the pre-treatment period is only three years and therefore does not constitute a solid basis to develop a stable synthetic control. I observe for 1989 that the actual trend of working hours of the treated country and the trend of the synthetic control drift apart. Regarding Australia, the actual data show a downward spike in 2003 which may explain the significant effect. However, this effect may be due to a set of reforms of the superannuation system that happened in 2002 and 2003 (Warren 2008). The placebo reform year therefore most likely coincides with these other reforms which took place during the same time.

[17] Excluding Luxembourg from the synthetic control group in this case does qualitatively not change anything as Luxembourg's weight in the synthetic control for the outcome variable working hours for Australia is zero and for France only 0.195 (Appendix, see Table A.2.2 and Table A.5.2).

Table 2: Post-treatment results regarding working hours of males aged 55-64, effects and pseudo p-values

"In-space placebos" (actual reform year, Börsch-Supan et al. 2018)			**"In-time placebos"** (placebo reform year)		
(1)	(2)	(3)	(4)	(5)	(6)
year	estimates	pseudo p-values	year	estimates	pseudo p-values
Australia					
			2002	-0.575	0.400
			2003	**-2.352***	**0.067**
			2004	-0.803	0.467
2005	-1.12	0.333			
2006	**-2.646*****	**0**			
2007	**-4.057*****	**0**			
2008	-0.643	0.8			
2009	-0.994	0.667			
2010	-1.108	0.733			
Austria					
			1997	-0.016	0.421
			1998	-0.004	1
			1999	0.006	0.947
2000	0.421	0.833			
2001	0.838	0.333			
2002	0.255	0.667			
2003	0.522	0.667			
2004	**2.645*****	**0**			
2005	**1.964***	**0.056**			
2006	**2.052***	**0.056**			
2007	2.073	0.167			
2008	1.657	0.278			
2009	2.158	0.111			
2010	1.378	0.444			
Belgium					
			1999	-0.777	0.500
			2000	-1.348	0.286
			2001	-0.759	0.643
2002	-0.484	0.571			
2003	-0.396	0.786			
2004	-0.755	0.214			
2005	**-1.254***	**0.071**			
2006	-0.733	0.429			
2007	-0.861	0.357			
2008	**-1.984***	**0.071**			
2009	-1.619	0.143			
2010	-1.317	0.286			
Denmark					
			1992	0.269	0.467
			1993	-0.663	0.333
			1994	0.655	0.667
1995	-0.478	0.333			
1996	-0.073	1			
1997	**-1.875***	**0.067**			
1998	-0.178	0.933			
1999	-0.232	0.933			
2000	0.712	0.533			
2001	1.501	0.267			

		France			
			1990	**-1.103***	**0.077**
			1991	-0.747	0.308
			1992	-0.969	0.308
1993	-0.303	1			
1994	-0.75	0.615			
1995	**-1.485***	**0.077**			
1996	**-2.131***	**0.077**			
1997	-2.397	0.154			
1998	-2.394	0.154			
1999	-2.021	0.077			
2000	-1.723	0.154			
2001	**-2.36***	**0.077**			
		Germany			
			1989	0.022	1
			1990	-0.348	0.538
			1991	-0.482	0.462
1992	-0.493	0.231			
1993	-0.636	0.231			
1994	-1.16	0.154			
1995	**-1.352*****	**0**			
1996	-0.723	0.154			
1997	-0.297	0.692			

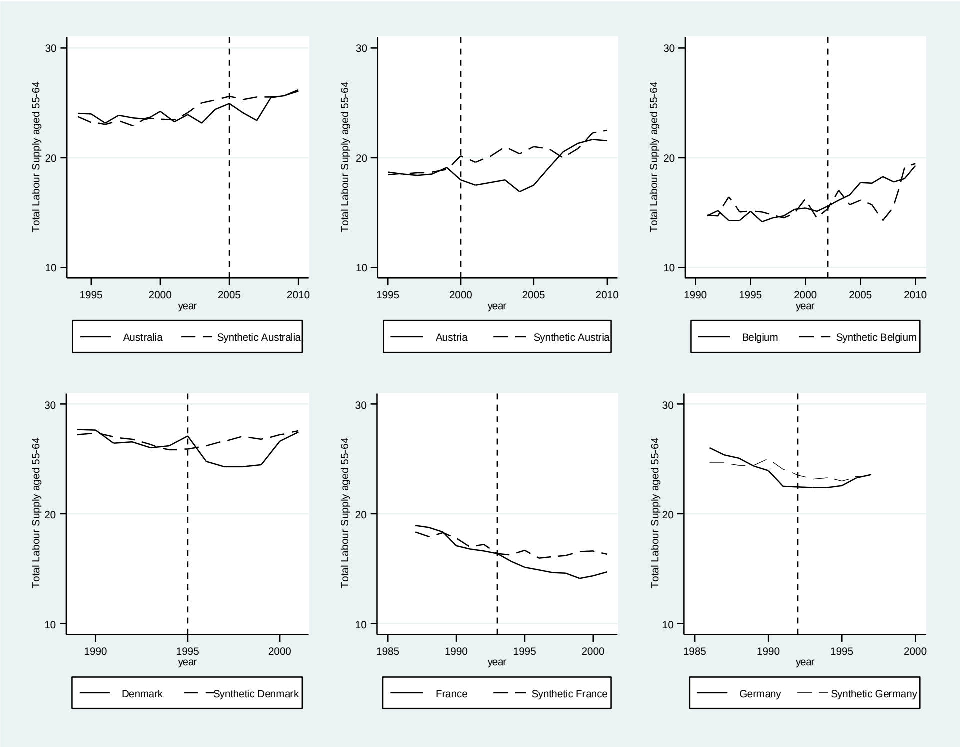
Note: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Source: Börsch-Supan et al. (2018) and own calculations.

4.3 Total labor supply

As a final step, I show the effects of flexibility reforms on total labor supply for males in the age group 55-64. As total labor supply is measured as the product of labor force participation and working hours for those who participated, the time periods covered are determined by the availability of data on these two variables. Time series data on labor force participation are available for slightly more years compared to working hours. That is why the time periods covered on total labor supply is mainly determined by the availability of the time series of working hours.

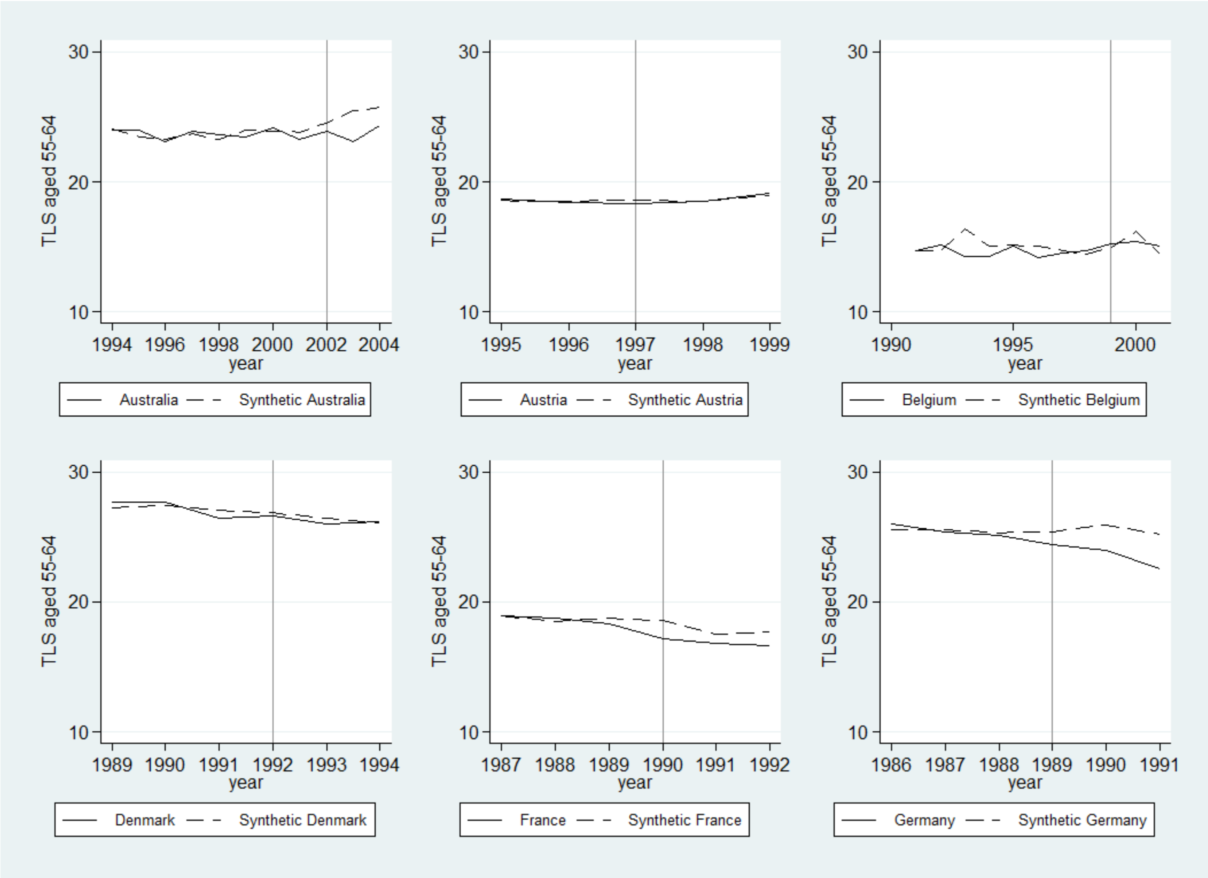
Figure 5: Trends in total labor supply: treated vs. synthetic control. Actual reform year



Source: Börsch-Supan et al. (2018).

Treatment effects on total labor supply of men aged 55-64 are shown graphically in Figure 5 and the yearly effect sizes and pseudo p-values are reported in Table 3. Figure 5 again recalls the results of Börsch-Supan et al. (2018) and graphically shows that in all countries the overall change in total hours worked per week after the reform is negative. The only exception is Belgium where the synthetic control reflects higher values in the majority of post-treatment years compared to the treated country values. Pseudo p-values given in Table 3 mostly prove what Figure 5 graphically displays. Total labor supply significantly decreased in Australia (2007) and Austria (2000-2006). After a significant increase in Denmark in the reform year (1995), the total number of working hours significantly decreases in later years (1997-1999). In Belgium, total labor supply significantly increases in 2007, and the total labor supply did not change significantly after the flexibility reforms in France and Germany.

Figure 6: Trends in total labor supply: treated vs. synthetic control. Placebo reform year



Source: Own calculations.

Eventually, Figure 6 graphically shows the treatment effects when reassigning the reform years to placebo reform years. As for labor force participation and working hours, placebo reform years are reassigned three years earlier compared to the actual reform years. Data availability is again the crucial factor to construct synthetic controls that reproduce the actual trend in males' total labor supply more precisely. However, in particular for Australia, Austria, and in the first years of the observation period for France and Germany, the synthetic controls come close to the actual trends.

Table 3 finally compares yearly treatment effect sizes and its significance from Börsch-Supan et al. (2018) with the results from the robustness checks with the placebo reforms years. A similar transition can be ascertained as for labor force participation and working hours: When estimating treatment effects with placebo reforms years, the synthetic control method does not produce significant results anymore. This holds true for Austria, Belgium, Denmark, and Germany and all years. Only Australia (2003) and France (1990) are exemptions. As total labor supply is the product of extensive margin (labor force participation) and intensive margin (working hours), the effects found for Australia and France most likely are translated from what was found for working hours (see Section 4.2): For France, the pre-treatment period is only three years which does not allow the establishment of a stable synthetic control, and for Australia the placebo reform year may most likely capture other reforms of the superannuation system that happened at that time (Warren 2008).

Table 3: Post-treatment results for total labor supply of males aged 55-64, effects and pseudo p-values

"In-space placebos" (actual reform year, Börsch-Supan et al. 2018)			**"In-time placebos"** (placebo reform year)		
(1)	(2)	(3)	(4)	(5)	(6)
year	estimates	pseudo p-values	year	estimates	pseudo p-values
Australia					
			2002	-0.633	0.133
			2003	**-2.372*** **	**0**
			2004	-1.359	0.133
2005	-0.628	0.733			
2006	-1.189	0.4			
2007	**-2.194*** **	**0.067**			
2008	-0.052	1			
2009	0.016	1			
2010	0.154	0.933			
Austria					
			1997	-0.227	0.389
			1998	-0.040	1
			1999	0.113	0.944
2000	**-2.193*** **	**0**			
2001	**-2.095 *** **	**0**			
2002	**-2.407*** **	**0**			
2003	**-3.056*** **	**0**			
2004	**-3.455*** **	**0**			
2005	**-3.459*** **	**0**			
2006	**-1.786* **	**0.056**			
2007	0.535	0.444			
2008	0.485	0.5			
2009	-0.603	0.333			
2010	-0.92	0.167			

		Belgium			
			1999	0.349	0.929
			2000	-0.826	0.786
			2001	0.641	0.786
2002	0.19	0.929			
2003	-0.873	0.714			
2004	0.869	0.643			
2005	1.605	0.5			
2006	1.932	0.5			
2007	**4.011***	**0**			
2008	2.231	0.143			
2009	-1.008	0.643			
2010	-0.226	0.929			
		Denmark			
			1992	-0.271	0.733
			1993	-0.374	0.467
			1994	0.131	0.867
1995	**1.193***	**0**			
1996	-1.457	0.133			
1997	**-2.302***	**0.067**			
1998	**-2.744***	**0.067**			
1999	**-2.339***	**0.067**			
2000	-0.58	0.667			
2001	-0.124	1			
		France			
			1990	**-1.409***	**0**
			1991	-0.715	0.462
			1992	-1.077	0.231
1993	0.024	0.923			
1994	-0.592	0.538			
1995	-1.559	0.308			
1996	-1.104	0.385			
1997	-1.446	0.462			
1998	-1.591	0.615			
1999	-2.438	0.308			
2000	-2.225	0.308			
2001	-1.557	0.462			
		Germany			
			1989	-0.984	0.231
			1990	-1.962	0.154
			1991	-2.713	0.231
1992	-1.107	0.692			
1993	-0.802	0.615			
1994	-0.892	0.769			
1995	-0.447	0.846			
1996	-0.155	0.923			
1997	0.127	1			

Note: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Source: Börsch-Supan et al. (2018) and own calculations.

5. SUMMARY AND CONCLUSION

Over the past decades, several OECD countries have enacted flexible retirement reforms. According to politicians, flexibilization is supposed to provide an elegant way to increase older workers' labor supply. The model in Börsch-Supan et al. (2018), however, has shown from a theoretical point of view that flexibility reforms may have ambiguous effects on labor supply. Therefore, the question whether flexibility reforms have actually helped to increase older workers' labor supply remains an empirical one.

In the empirical analysis of Börsch-Supan et al. (2018), the synthetic control method is used to investigate the effect of flexibility reforms enacted in different OECD countries between 1992 and 2006. As large sample inference techniques are problematic in comparative case studies, placebo studies can provide a remedy to facilitate inference. Following the nomenclature of Abadie et al. (2015), there are two possibilities of placebo studies: in-space placebos and in-time placebos. To investigate the significance of the estimates, in-space placebo studies were applied in Börsch-Supan et al. (2018). This meant an artificially reassigning of the treatments (i.e. flexibility reform) to members of the synthetic control group which were not directly exposed to a reform at the same time. The results show that labor force participation of males aged 55-64 has very little if at all increased in some countries and years due to the flexibility reforms introduced since the 1990s. At the same time, we find that older men of the same age group have decreased their weekly working hours. In sum, the reforms have produced zero to negative effects on total labor supply.

The aim of this paper is to scrutinize these results by applying in-time placebo studies. The strategy of in-time placebo studies is to reassign the treatment to dates when then actual reform did not take place (i.e. a reassignment with respect to time). The idea is to find out whether the synthetic control method produces significant effects when applied to dates other than the actual reform date. If this were the case, the confidence about the validity of the results presented in Börsch-Supan et al. (2018) would dissipate.

The application of in-time placebo studies, however, reveals that the results of Börsch-Supan et al. (2018) are stable to this robustness check. In contrast to the results from the in-space placebo study, reassigning the reform years to placebo reform years display no perceivable effects. Significant effects diminish in almost all cases when reassigning the reform dates three years earlier. Remaining effects more likely stem from poor synthetic controls resulting from data restrictions or from the concurrence of other reforms.

Finding a good synthetic control country is crucial for the estimation quality. Yet, a key challenge of the synthetic control method is that it requires a substantial amount of data. Particularly time series on working hours for older workers are hard to obtain. The application of placebo reform years to earlier years reduces the length of pre-treatment periods even more. A higher number of pre-treatment periods would very likely substantially improve the pre-treatment fit. However, more data is not available. If one nevertheless does not want to forgo the potential in-time placebo robustness check, data restrictions remain Achilles' heel.

REFERENCES

Acemoglu, D., S. Johnson, A. Kermani, J. Kwak, and T. Mitton (2016). 'The value of connections in turbulent times: Evidence from the United States', *Journal of Financial Economics*, 121(2), 368-391.

Abadie, A. and J. Gardeazabal (2003). 'The economic costs of conflict: A case study of the Basque Country', *American Economic Review*, 93(1), 113-132.

Abadie, A., J. Diamond and J. Hainmueller (2010). 'Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco Control Program', *Journal of the American Statistical Association*, 105(490), 493-505.

Abadie, A., A. Diamond and J. Hainmueller (2015). 'Comparative politics and the synthetic control method', *American Journal of Political Science*, 59(2), 495-510.

Athey, S. and G. W. Imbens (2017). 'The state of applied econometrics: Causality and policy evaluation', *Journal of Economic Perspectives*, 31(2), 3–32.

Barro, R. and J.-W. Lee (2013). 'A new data set of educational attainment in the World, 1950-2010', *Journal of Development Economics*, 104, 184-198.

Börsch-Supan, A., T. Bucher-Koenen, N. Goll, and V. Kutlu Koc (2018). 'Dangerous flexibility – Retirement reforms reconsidered', *Economic Policy*, 33(4), 315–355.

Börsch-Supan, A. and C. Coile (2019). 'Introduction to 'Social Security Programs and Retirement around the World: Reforms and Retirement Incentives'', preliminary draft, in: Börsch-Supan, A. and Coile, C. (eds, forthcoming): *Social Security Programs and Retirement around the World: Reforms and Retirement Incentives*, forthcoming from University of Chicago Press, available [https://www.nber.org/chapters/c14190.pdf], [Accessed 21.10.2019].

Bilgel, F. and B. Galle (2015). 'Financial incentives for kidney donation: A comparative case study using synthetic controls', *Journal of Health Economics*, 43, 103–117.

Billmeier, A. and T. Nannicini (2013). 'Assessing economic liberalization episodes: A synthetic control approach', *The Review of Economics and Statistics*, 95(3), 983–1001.

Bohn, S., M. Lofstrom, and S. Raphael (2014). 'Did the 2007 Legal Arizona Workers Act reduce the state's unauthorized immigrant population?', *The Review of Economics and Statistics*, 96(2), 258–269.

Card, D (1990). 'The impact of the Mariel Boatlift on the Miami labor market', *Industrial and Labor Relations Review*, 43(2), 245–57.

Cavallo, E., S. Galiani, I. Noy, and J. Pantano (2013). 'Catastrophic natural disasters and economic growth', *The Review of Economics and Statistics*, 95(5), 1549-1561.

Duval, R. (2003). 'The retirement effects of old-age pension and early retirement schemes in OECD countries', *OECD Economics Department Working Papers*, 370, OECD: Paris.

Ferman, B., C. Pinto, and V. Possebom (2020). 'Cherry picking with synthetic controls', *Journal of Policy Analysis and Management*, accepted and forthcoming.

Freire, D. (2018). ‘Evaluating the effect of homicide prevention strategies in São Paulo, Brazil: A synthetic control approach’, *Latin American Research Review*, 53(2), 231–249.

Galiani, S. and B. Quistorff (2016). ‘The synth_runner package: Utilities to automate synthetic control estimation using synth’ [econweb.umd.edu/~galiani/files/synth_runner.pdf], [accessed 20.12.2019].

Gobillon, L. and T. Magnac (2016). ‘Regional policy evaluation: Interactive fixed effects and synthetic controls’, *The Review of Economics and Statistics*, 98(3), 535-551.

Hahn, J. and R. Shi (2017). ‘Synthetic control and inference’, *Econometrics*, 5(4), 52.

Hinrichs, P. (2012). ‘The Effects of affirmative action bans on college enrollment, educational attainment, and the demographic composition of universities’, *The Review of Economics and Statistics*, 94(3), 712–722.

Kaul, A., S. Klößner, G. Pfeifer and M. Schieler (2015). ‘Synthetic control methods: Never use all pre-intervention outcomes as economic predictors’, [http://www.oekonometrie.uni-saarland.de/papers/SCM_Predictors.pdf], [Accessed 11.11.2016].

Kaul, A., S. Klößner, G. Pfeifer and M. Schieler (2018). ‘Synthetic control methods: Never use all pre-intervention outcomes together with covariates’, [http://www.oekonometrie.uni-saarland.de/papers/SCM_Predictors.pdf], [Accessed 06.11.2019].

Kleven, H. J., C. Landais and E. Saez (2013). ‘Taxation and international migration of superstars: Evidence from the European football market’, *American Economic Review* 103(5), 1892-1924.

Kreif, N., R. Grieve, D. Hangartner, A.J. Turner, S. Nikolova, and M. Sutton (2016). ‘Examination of the synthetic control method for evaluating health policies with multiple treated units’, *Health Economics*, 25, 1514–1528.

Lindecke, C., D. Voss-Dahm and S. Lehndorff (2007). ‘Altersteilzeit, Erfahrungen und Diskussionen in Deutschland und anderen EU-Ländern’, *Arbeitspapier* 142, Hans-Böckler-Stiftung: Düsseldorf.

Liu, S. (2015). ‘Spillovers from universities: Evidence from the land-grant program’, *Journal of Urban Economics*, 87, 25–41.

Nannicini, T., and A. Billmeier (2011). ‘Economies in transition: How important is trade openness for growth?’, *Oxford Bulletin of Economics and Statistics*, 73(3), 287–314.

OECD (2011). *Pensions at a Glance 2011*, OECD Publishing: Paris.

OECD (2013). *Pensions at a Glance 2013*, OECD: Paris.

OECD (2016a). ‘Level of GDP per capita and productivity’, OECD.Stat, Data, Economy, Productivity, GDP per capita and productivity levels.

OECD (2016b). ‘Life expectancy at birth’, OECD.Stat, Data, Health Status, Life expectancy.

O’Neill, S., N. Kreif, R. Grieve, M. Sutton, and J. S. Sekhon (2016). ‘Estimating causal effects: considering three alternatives to difference-indifferences estimation’, *Health Services and Outcomes Research Methodology*, 16, 1–21.

Saia, A. (2017). ‘Choosing the open sea: The cost to the UK of staying out of the euro’, *Journal of International Economics* 108, 82-98.

Stearns, J. (2015). ‘The effects of paid maternity leave: Evidence from temporary disability insurance’, *Journal of Health Economics*, 43, 85–102.

Social Security Administration (1985-2014). *Social Security Programs throughout the World*, SSA: Washington, D. C., Geneva.

Warren, D. (2008). ‘Australia’s retirement income system: Historical development and effects of recent reforms’, *Melbourne Institute Working Paper*, 8, Melbourne.

APPENDIX

A.1 Pre- and post-treatment periods by country

Table A.1.1: Treated countries, placebo reform years and time periods for labor force participation (LFP) and working hours (WH)

Treated country	Actual reform year	Placebo reform year	Pre-treatment time period (in-time placebo study)	Post-treatment time period (in-time placebo study)
Australia	2005	2002	LFP: 1994-2001 WH: 1994-2001	LFP: 2002-2004 WH: 1994-2004
Austria	2000	1997	LFP: 1994-1996 WH: 1995-1996	LFP: 1997-1999 WH: 1997-1999
Belgium	2002	1999	LFP: 1992-1998 WH: 1991-1998	LFP: 1999-2001 WH: 1999-2001
Denmark	1995	1992	LFP: 1985-1991 WH: 1989-1991	LFP: 1992-1994 WH: 1992-1994
France	1993	1990	LFP: 1983-1989 WH: 1987-1989	LFP: 1990-1992 WH: 1990-1992
Germany	1992	1989	LFP: 1983-1988 WH: 1986-1988	LFP: 1989-1991 WH: 1989-1991
Sweden	2000	1997	LFP: 1985-1996 WH: 1990-1996	LFP: 1997-1999 WH: 1997-1999

Note: Post-treatment time periods are restricted to the year before actual reform year. Otherwise, the actual reform effect would affect the results.

Source: Own calculations.

A.2 Synthetic control weights

Table A.2.1: Synthetic control weights, outcome variable: Labor force participation

Untreated countries	Treated countries						
	Australia	Austria	Belgium	Denmark	France	Germany	Sweden
Belgium	-	-	-	0	0.331	-	-
Canada	0.649	0	0	0	0	0	0.398
Czech republic	0	0.021	-	-	-	-	-
Estonia	0	-	-	-	-	-	-
Finland	-	-	-	0	0	-	-
Greece	0	0	0	0	0	0	0
Hungary	0	0	0.357	-	-	-	-
Iceland	0	0	0	-	-	-	-
Ireland	0	0	0	0	0	0.487	0
Israel	0.188	0	-	0	-	-	0.101
Italy	0	0	0.102	0	0.257	0.161	0
Japan	0.027	0.005	0	0	0	0	0.418
Korea	0	0	0	0.145	0	0	0.083
Luxembourg	0.053	0.591	0.541	0.198	0.413	0.327	0
The Netherlands	-	-	-	-	-	-	0
New Zealand	0	0	0	0	0	0	0
Norway	0	0	0	0	0	0	0
Poland	0	0.051	0	-	-	-	-
Portugal	0	0	0	0	0	0.025	0
Slovak Republic	0	-	-	-	-	-	-
Spain	0	0.235	0	0.05	0	0	0
Sweden	-	-	-	0.607	-	-	-
Switzerland	0	0	0	-	-	-	-
UK	0	0	0	0	0	0	0
US	0.083	0	0	0	0	0	0
Time periods covered	1994-2004	1994-1999	1992-2001	1985-1994	1983-1992	1983-1991	1985-1999

Note: “-” means that the corresponding country is not included in the estimation. The upper bound of the time periods covered is the pre-reform year of the actual flexibility reform.

Source: Own calculations.

Table A.2.2: Synthetic control weights, outcome variable: weekly working hours

Untreated countries	**Treated countries**						
	Australia	Austria	Belgium	Denmark	France	Germany	Sweden
Australia	-	-	-	-	-	0	-
Belgium	-	-	-	0.038	0.206	0	-
Canada	0	0	0	0	0	0	0
Czech republic	-	-	-	-	-	-	-
Estonia	-	-	-	-	-	-	-
Finland	-	-	-	0.441	-	-	-
Greece	0	0	0	0	0	0	0
Hungary	-	0.108	-	-	-	-	-
Iceland	-	0	-	-	-	-	-
Ireland	0	0.095	0	0	0.168	0	0
Israel	-	0	-	-	-	-	-
Italy	0	0.228	0.056	0	0	0	0
Japan	0	0	0	0.005	0	0	0
Korea	0	0	0	-	-	-	-
Luxembourg	0	0	0	0.136	0.195	0.18	0
The Netherlands	-	-	-	0.38	-	0.331	0.86
New Zealand	0	0	0	0	0	0	0
Norway	0.449	0.320	0.074	0	-	-	0.14
Poland	-	-	-	-	-	-	-
Portugal	0	0	0	0	0	0.402	0
Slovak Republic	0.073	0	-	-	-	-	-
Spain	0	0.249	0.507	0	0.431	-	0
Sweden	-	-	-	-	-	-	-
Switzerland	0.478	0	0	-	-	-	-
UK	0	0	0	0	0	0.007	0
US	0	0	0.363	0	0	0.108	0
Time periods covered	1994-2004	1995-1999	1991-2001	1989-1994	1987-1992	1986-1991	1990-1999

Note: "-" means that the corresponding country is not included in the estimation. The upper bound of the time periods covered is the pre-reform year of the actual flexibility reform.

Source: Own calculations.

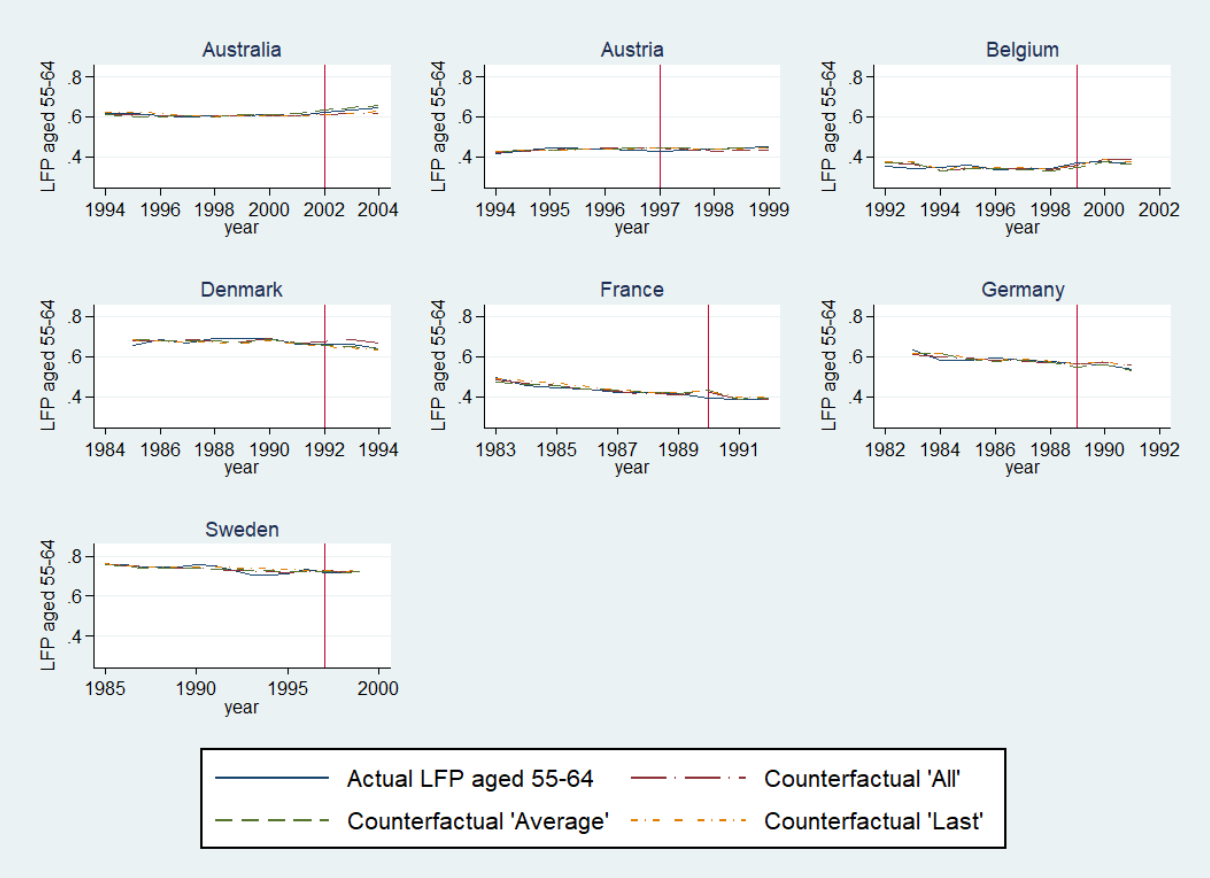
A.3 Robustness of the treatment effects

As in Börsch-Supan et al. (2018), I check the robustness of the treatment effects for the case of the placebo reform years by using a method developed by Kaul et al. (2015, updated in Kaul et al. 2018). Kaul et al. (2015) recommend applying the synthetic control method at least twice: one estimation should incorporate only the average of the outcome variable's pre-treatment values in addition to the set of covariates. Another version should only include the last pre-treatment value of the outcome variable (i.e. in the last period before the treatment) in addition to the other covariates. If the two estimation versions yield similar results in terms of similar weights of the corresponding synthetic units and, therefore, in terms of similar patterns of the predicted counterfactuals that come close to each other, the treatment effects are unbiased.

Another specification could include all pre-treatment values in addition to the covariates. Kaul et al. (2018) state that it actually becomes increasingly popular in applications of synthetic control methods to include the entire pre-treatment path of the outcome variable as economic predictors and following Cavallo et al. (2013) including the entire pre-treatment path seems the obvious choice. Including all pre-treatment values of the outcome variables is exactly was has been done, e.g., in Bilgel and Galle (2015), Billmeier and Nannicini (2013), Bohn et al. (2014), Hinrichs (2012), Kreif et al. (2016), Liu (2015), Nannicini and Billmeier (2011), O'Neill et al. (2016), and Stearns (2015). Kaul et al. (2018) demonstrate, however, both theoretically and empirically that incorporating all outcome lags causes all other covariates to be irrelevant in the estimation. This finding holds irrespective of how important these covariates are for accurately predicting post-treatment values of the outcome and therefore threatens the estimator's unbiasedness.

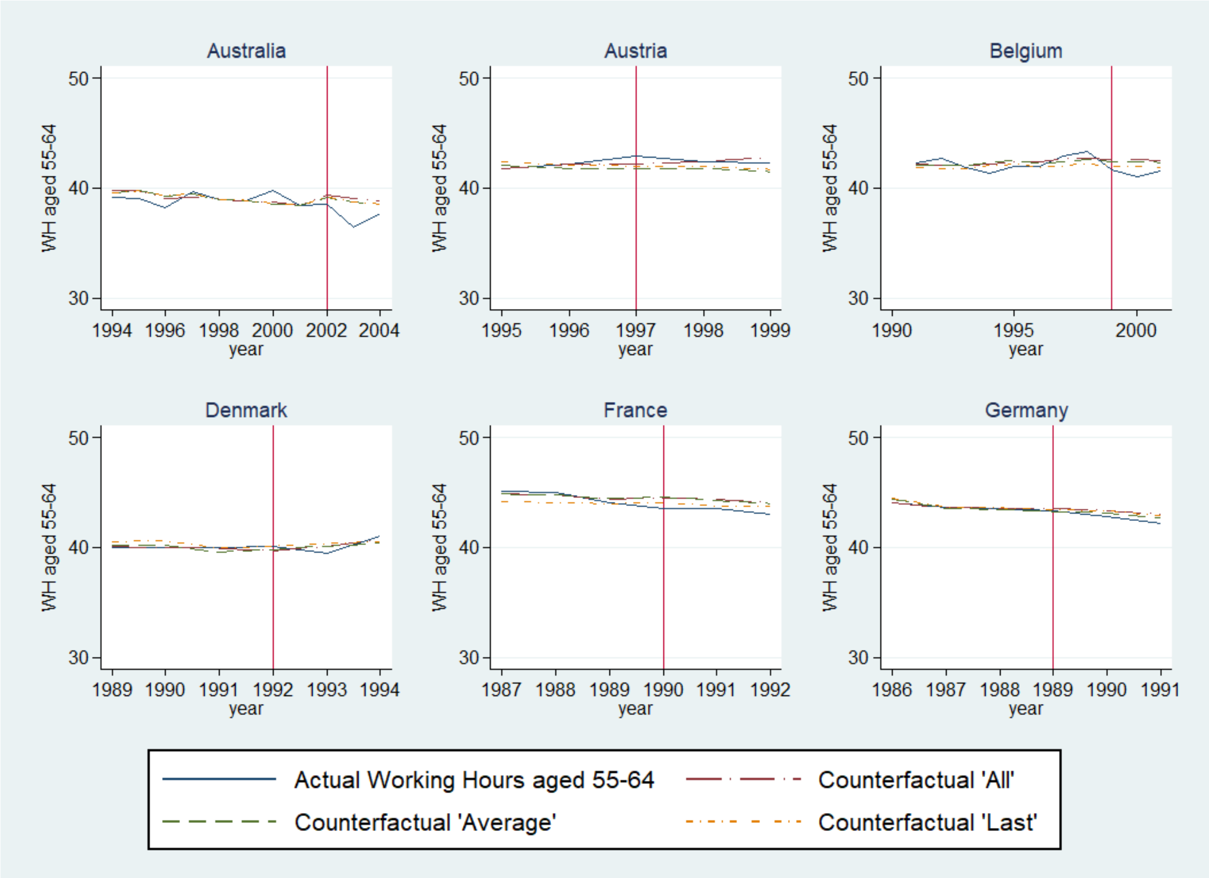
Following the recommendation of Kaul et al. (2015) to apply the synthetic control method at least twice, I report treatment effects on labor force participation (Figure A.3.1) and working hours (Figure A.3.2) for males aged 55-64, under different specifications and for all treated countries. In Figure A.3.1 and Figure A.3.2, the red vertical line indicates the placebo reform year. The blue solid line depicts the actual outcome trajectory of the treated country and the red long dashed-dotted line shows the synthetic control country when controlling for the entire pre-treatment path of the outcome variable ("All") in addition to the set of other covariates. While the green dashed line represents the synthetic control when including the average pre-treatment value of the outcome variable plus covariates, the orange short dashed-dotted line depicts the specification when controlling for the last pre-treatment outcome value only plus covariates. The comparison of the green dashed line and the orange short dashed-dotted line reveals that controlling for the pre-treatment average or the last pre-treatment value yields similar results. This holds true for both outcome variables labor force participation (Figure A.3.1) and working hours (Figure A.3.2). Since the use of the average and the last pre-treatment values give similar results, I use the average value of the pre-treatment outcome values in addition to the set of covariates in the analysis.

Figure A.3.1: Trends in males' LFP aged 55-64, robustness of the treatment effects. Placebo reform year



Source: Own calculations.

Figure A.3.2: Trends in males' working hours aged 55-64, robustness of the treatment effects. Placebo reform year



Source: Own calculations.

A.4 Quality of pre-treatment characteristics

Table A.4.1 and Table A.4.2 compare the pre-treatment characteristics of the countries experienced the reform and its synthetic control country. Table A.4.1 shows the characteristics for the outcome variable labor force participation, and Table A.4.2 depicts the values for the outcome variable working hours. Overall, the tables allow the same conclusion as in Börsch-Supan et al. (2018): The synthetic control countries provide a good approximation for the treated country in the years before the flexibility reform. Only when it comes to GDP per capita, in few countries there is a discrepancy between the actual country values and the values of the synthetic country. This stems from the same fact as in Börsch-Supan et al. (2018) namely that per-capita GDP has the lowest predictive power especially for labor force participation among all predictor variables. I also use the statutory eligibility age instead of possible years of early retirement for some countries as the quality of the pre-treatment matches increased in these cases.

Table A.4.1: Labor force participation predictor means before the flexibility reform

	Australia	Synthetic Australia	Austria	Synthetic Austria
LFP aged 55-64	0.607	0.607	0.433	0.434
LFP aged 25-54	0.909	0.908	0.934	0.935
SEA	65	65	65	64.96
GDP per capita	34,125	34,134	32,687	45,223
Years of schooling	11.2	11.1	9.6	9.6
Life expectancy	75.8	75.5	73.4	73. 4
	Belgium	Synthetic Belgium	Denmark	Synthetic Denmark
LFP aged 55-64	0.344	0.345	0.678	0.678
LFP aged 25-54	0.920	0.912	0.942	0.950
Years of early retirement	5	4.2	7	6.2
GDP per capita	31,302	41,084	31,796	29,461
Years of schooling	10.2	9.9	9.6	10.04
Life expectancy	73.6	73.6	72.0	72.5
	France	Synthetic France	Germany	Synthetic Germany
LFP aged 55-64	0.441	0.442	0.589	0.589
LFP aged 25-54	0.958	0.946	0.926	0.939
Years of early retirement	5	5.005	2	2.927
GDP per capita	25,946	32,755	26,886	25,752
Years of schooling	7.2	8.8	8.19	9.14
Life expectancy	71.6	71.4	71.53	71.24
	Sweden	Synthetic Sweden		
LFP aged 55-64	0.737	0.737		
LFP aged 25-54	0.934	0.942		
Years of early retirement	6.5	4.50		
GDP per capita	29,218	26,814		
Years of schooling	10.54	10.55		
Life expectancy	75.04	74.43		

Note: Years of early retirement is defined as the difference between the statutory eligibility age and the earliest eligibility age.

Source: Own calculations.

Table A.4.2: Working hours predictor means before the flexibility reform

	Australia	Synthetic Australia
WH aged 55-64	38.99	38.10
WH aged 25-54	41.41	41.61
SEA	65	65.533
GDP per capita	34,125	45,052
Years of schooling	11.2	10.8
Life expectancy	75.8	75.4
	Austria	Synthetic Austria
WH aged 55-64	41.9	41.9
WH aged 25-54	41.4	41.6
SEA	65	64.4
GDP per capita	33,077	33,134
Years of schooling	9.7	9.7
Life expectancy	73.5	73.8
	Belgium	Synthetic Belgium
WH aged 55-64	42.28	42.28
WH aged 25-54	40.69	42.26
SEA	65	64.94
GDP per capita	31,107	31,933
Years of schooling	10.1	9.9
Life expectancy	73.5	73.9
	Denmark	Synthetic Denmark
WH aged 55-64	39.98	39.99
WH aged 25-54	40.77	40.51
Years of early retirement	7	5
GDP per capita	32,336	32,380
Years of schooling	9.8	9.4
Life expectancy	72.17	72.34
	France	Synthetic France
WH aged 55-64	44.7	44.7
WH aged 25-54	42.28	42.87
Years of early retirement	5	4.3
GDP per capita	27,161	27,079
Years of schooling	7.87	8.55
Life expectancy	72.27	72.40

	Germany	Synthetic Germany
WH aged 55-64	43.76	43.80
WH aged 25-54	42.49	42.90
Years of early retirement	2	2.42
GDP per capita	27,814	26,685
Years of schooling	8.47	8.53
Life expectancy	71.67	71.57

Note: Years of early retirement is defined as the difference between the statutory eligibility age and the earliest eligibility age.

Source: Own calculations.

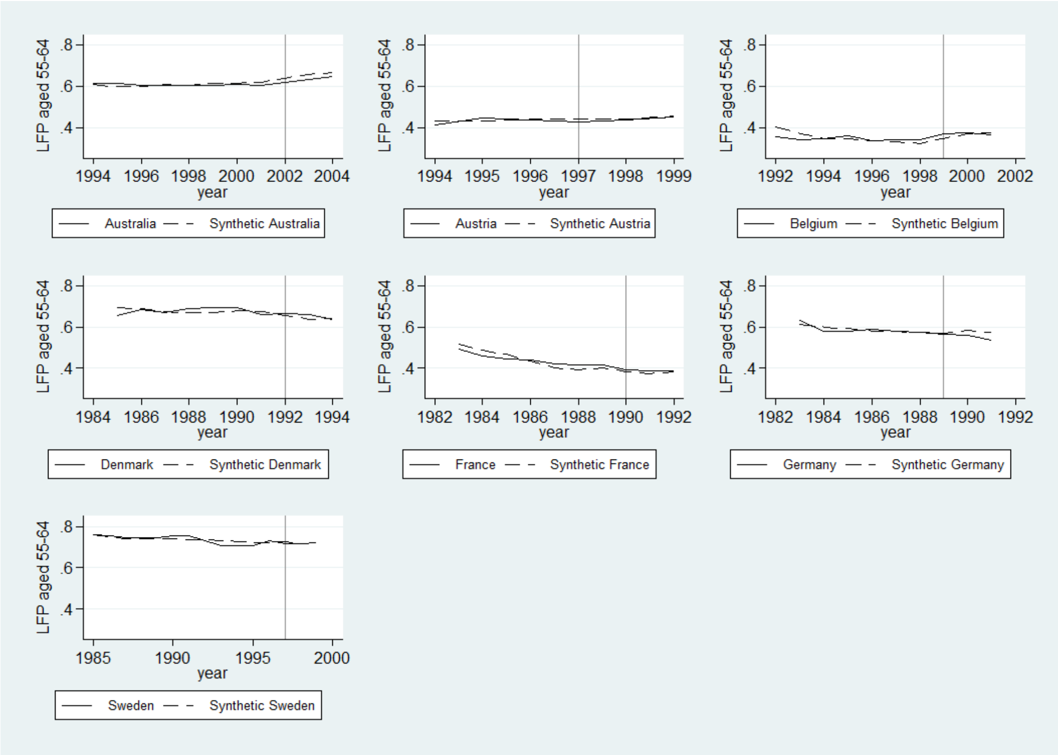
A.5 Sensitivity check: dropping Luxembourg from the synthetic control groups

Abadie et al. (2015) state that the selection of comparison units is crucial to avoid erroneous conclusions. Differences in outcome variables between treated units and control units may merely reflect disparities in the units' characteristics, if comparison units are not sufficiently similar to the units of interest. A potential problem in this paper might be that the synthetic counterfactual in many cases heavily relies on Luxembourg. In other words, for constructing the synthetic controls, in particular for the case of labor force participation, Luxembourg has a rather high weight for almost all countries (see Table A.2.1). Luxembourg, however, is a relatively small country, with very close labor market ties to France, Germany and Belgium. Thus, labor market developments in Luxembourg might not be completely independent of the developments in the surrounding countries. I therefore drop Luxembourg from the synthetic control group as a robustness check.

Figure A.5.1, Figure A.5.2, and Figure A.5.3 show the trends in outcome variables for the treated country and the synthetic controls when excluding Luxembourg from the synthetic control group for all treated countries. Figure A.5.4 and Figure A.5.5 display the robustness of the treatment effects for different specifications.

The post-treatment results regarding labor force participation (Table A.5.1) do not at all show significant effects anymore when excluding Luxembourg. Therefore, the significant effect found for France in 1990 actually seems to have been driven by Luxembourg's outlier in labor force participation (see Section 4.1). For the outcome variable working hours, the exclusion of Luxembourg from the synthetic controls maintains the significant negative effects found for Australia (2003) and France (1990) (see Section 4.2). For France, this most likely stems from a poor synthetic control. Due to data availability, the pre-treatment period is only three years and does constitute a solid basis to develop a stable synthetic control. However, the effect for Australia may be due to a set of reforms of the superannuation system that happened in 2002 and 2003 (see Section 4.2 and Section 4.3). Both effects persist when excluding Luxembourg from the synthetic controls. Excluding Luxembourg from the synthetic controls yields significant negative effects on working hours for Germany (1989, 1990, 1991). The effects for Germany, however, very likely stem from a reform which first came into effect in 1989 and initialized part-time employment before retirement (*Altersteilzeit*). This scheme comprised a reduction of working hours in a specific period before full retirement (Lindecke et al. 2017). In practice, the scheme achieved its breakthrough only after a revision in the later 1990s, but may explain the negative effects found on working hours. While the negative effect for Australia (2003) translates also to total labor supply (Table A.5.3), no other effects remain significant. Overall, the robustness check excluding Luxembourg suggest that the results are not driven by the inclusion of Luxembourg in the synthetic controls.

Figure A.5.1: Trends in males' LFP aged 55-64: treated vs. synthetic control. Placebo reform years. Without Luxembourg in the synthetic control groups



Source: Own calculations.

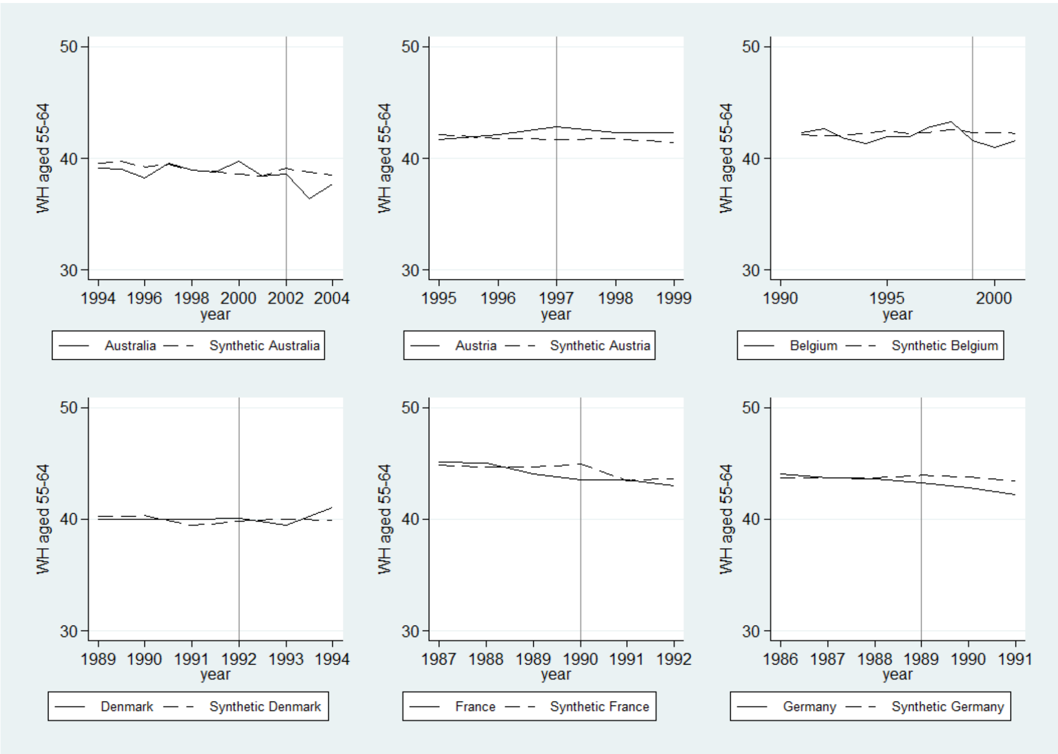
Table A.5.1: Post-treatment results regarding LFP of males aged 55-64, effects and pseudo p-values. Without Luxembourg in the synthetic control group

Australia			Austria		
year	estimates	pseudo p-values	year	estimates	pseudo p-values
2002	-0.019	0.150	1997	-0.014	0.667
2003	-0.023	0.100	1998	-0.003	1
2004	-0.017	0.350	1999	-0.004	1
Belgium			**Denmark**		
year	estimates	pseudo p-values	year	estimates	pseudo p-values
1999	0.021	0.750	1992	0.007	0.938
2000	0.007	0.875	1993	0.021	0.250
2001	-0.013	0.688	1994	-0.001	1
France			**Germany**		
year	estimates	pseudo p-values	year	estimates	pseudo p-values
1990	0.010	0.929	1989	-0.006	0.833
1991	0.011	1	1990	-0.023	0.833
1992	0.005	1	1991	-0.040	0.500
Sweden					
year	estimates	pseudo p-values			
1997	-0.011	0.571			
1998	-0.003	0.929			
1999	-0.002	1			

Note: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Source: Own calculations.

Figure A.5.2: Trends in males' working hours aged 55-64: treated vs. synthetic control. Placebo reform years. Without Luxembourg in the synthetic control group



Source: Own calculations.

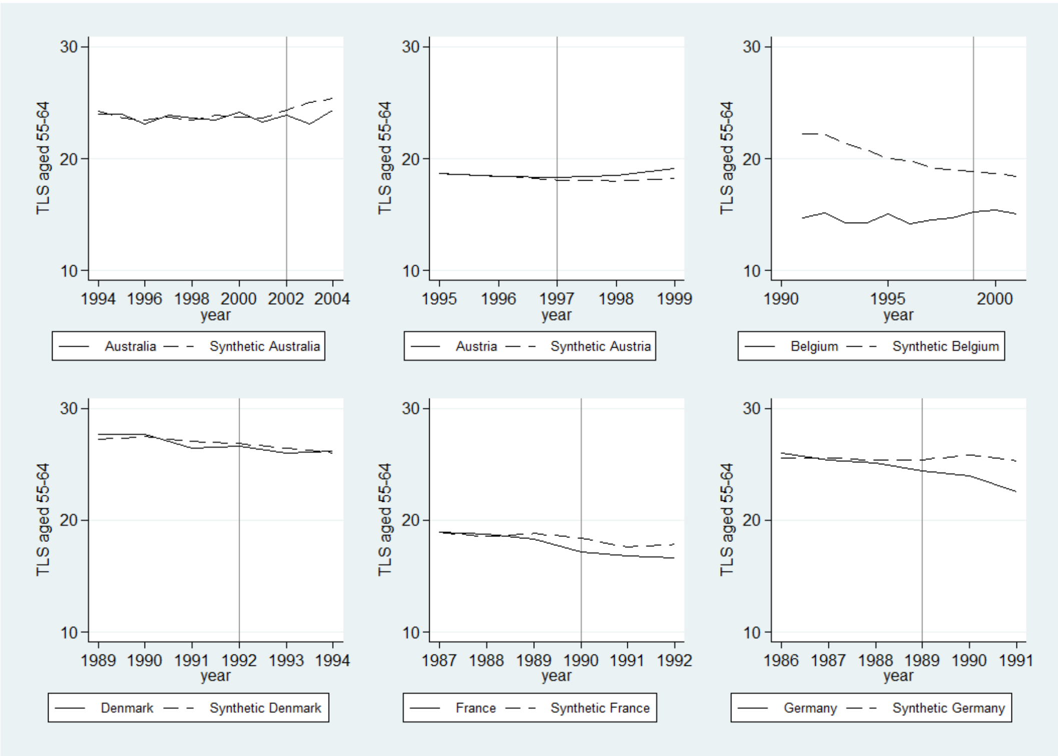
Table A.5.2: Post-treatment results regarding working hours of males aged 55-64, effects and pseudo p-values. Without Luxembourg in the synthetic control group

Australia			**Austria**		
year	estimates	pseudo p-values	year	estimates	pseudo p-values
2002	-0.577	0.400	1997	-0.014	0.667
2003	**-2.354***	**0.071**	1998	-0.003	1
2004	-0.805	0.500	1999	-0.004	1
Belgium			**Denmark**		
year	estimates	pseudo p-values	year	estimates	pseudo p-values
1999	-0.733	0.692	1992	0.266	0.500
2000	-1.304	0.231	1993	-0.554	0.643
2001	-0.704	0.846	1994	1.135	0.357
France			**Germany**		
year	estimates	pseudo p-values	year	estimates	pseudo p-values
1990	**-1.435***	**0.091**	**1989**	**-0.636*****	**0**
1991	0.040	1	**1990**	**-0.966*****	**0**
1992	-0.609	0.545	**1991**	**-1.242***	**0.083**

Note: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Source: Own calculations.

Figure A.5.3: Trends in males' total labor supply aged 55-64: treated vs. synthetic control. Placebo reform years. Without Luxembourg in the synthetic control group



Source: Own calculations.

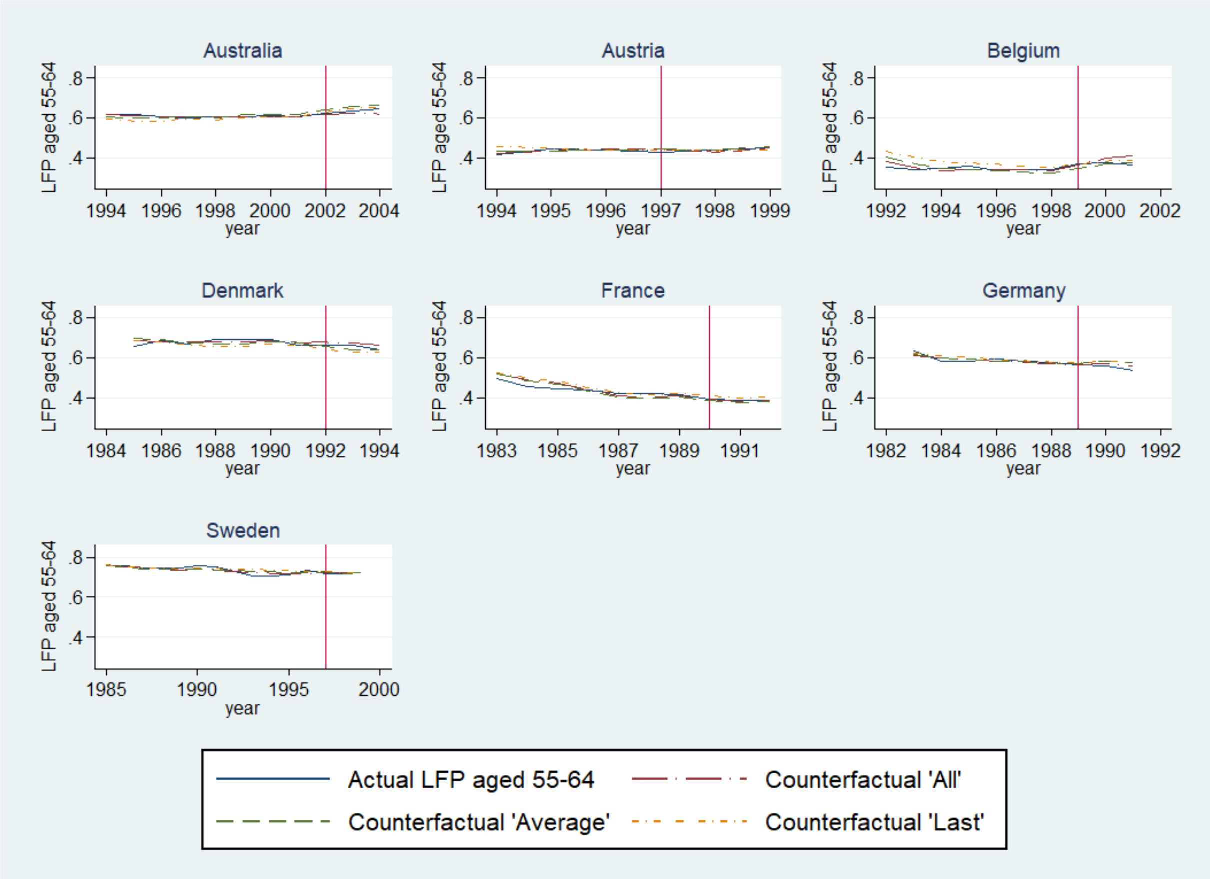
Table A.5.3: Post-treatment results regarding total labor supply of males aged 55-64, effects and pseudo p-values. Without Luxembourg in the synthetic control group

Australia			Austria		
year	estimates	pseudo p-values	year	estimates	pseudo p-values
2002	-0.409	0.133	1997	0.248	0.176
2003	**-1.970*** **	**0**	1998	0.456	0.176
2004	-0.996	0.143	1999	0.862	0.118
Belgium			**Denmark**		
year	estimates	pseudo p-values	year	estimates	pseudo p-values
1999	-3.582	0.923	1992	-0.273	0.786
2000	-3.315	0.769	1993	-0.369	0.571
2001	-3.307	0.846	1994	0.158	0.857
France			**Germany**		
year	estimates	pseudo p-values	year	estimates	pseudo p-values
1990	-1.295	0.182	1989	-1.028	0.250
1991	-0.796	0.545	1990	-1.885	0.167
1992	-1.225	0.273	1991	-2.790	0.250

Note: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

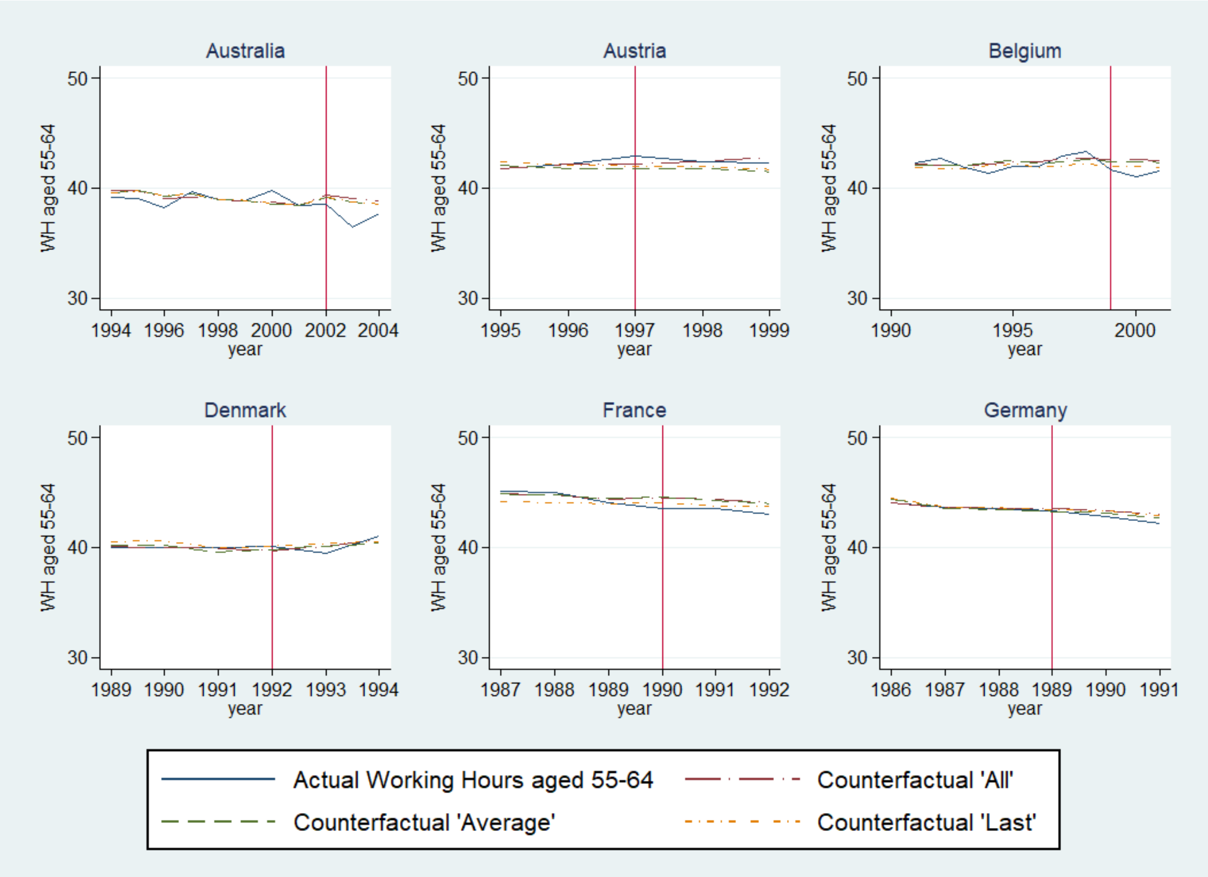
Source: Own calculations.

Figure A.5.4: Trends in males' LFP aged 55-64, robustness of the treatment effects. Placebo reform year. Without Luxembourg in the synthetic control group



Source: Own calculations.

Figure A.5.5: Trends in males' working hours aged 55-64, robustness of the treatment effects. Placebo reform year. Without Luxembourg in the synthetic control group



Source: Own calculations.