Late-career Risks in Changing Welfare States
Heisig, Jan Paul

Published by Amsterdam University Press

Heisig, Jan Paul.
Late-career Risks in Changing Welfare States: Comparing Germany and the United States since the 1980s.
Amsterdam University Press, 2015.
Project MUSE. muse.jhu.edu/book/66622.

For additional information about this book
https://muse.jhu.edu/book/66622
4 Data and methods

The remaining parts of this study will investigate how two late-career trigger events – exit from work and job loss – affect the financial situation of older workers and their families. My broader aim is to better understand how observed income trajectories are brought about by the interplay of welfare state provisions with ‘family buffering’ and labor market trajectories.

This chapter outlines the empirical approach for identifying the impact of the focal trigger events. I first provide a general description of my conceptual framework and analytical strategy (Section 4.1). Building on the counterfactual account of causality, I argue that simple before-after comparisons may provide a satisfactory picture of the impact of voluntary retirement, but that identifying the impact of late-career job loss requires comparison to an adequate control group. I identify difference-in-differences (DID) matching as a promising and increasingly popular empirical approach for constructing such a comparison group. Section 4.2 then discusses technical aspects of my implementation of DID matching. It also describes a method to account for compositional changes in the treatment group. In Section 4.3, I then summarize essential details of the data sets used in the subsequent analysis and describe the main outcome and event variables. More specific aspects of the analysis are covered in the individual chapters.

4.1 Conceptual framework and analytic strategy

To explore the economic effects of retirement and job loss, I will use longitudinal data that enable me to compare the situation of affected workers before and after the occurrence of an event. As just noted, I will describe technical aspects of the databases and key measures in Section 4.3 below. This section elaborates my general conceptual framework and analytical strategy.

Let $y_i$ represent an income-based measure of a worker’s economic situation, for example, an indicator for having low income below the poverty line or the level of income, perhaps expressed in percent of pre-event income (to focus on relative changes). Let $e_i = 1$ indicate that an individual $i$ experienced a given event $e$ such as late-career job loss between two time points $t = 0$ and $t = 1$ and let $e_i = 0$ indicate that an individual was not exposed to the event. Further, let $y_{i1}$ and $y_{i0}$ stand for $i$’s value of $y$ at $t = 1$ and $t = 0$, respectively. A natural starting point for investigating the impact
of e on y then is to focus on changes in y between \( t = 0 \) and \( t = 1 \) among those who experienced e. For individual i this difference is equal to \( y_{it} - y_{io} \) and its expected value can be written as:

\[
E(y_{it} - y_{io}|e_i^* = 1)
\]  

(4.1)

This quantity is readily estimated by computing the corresponding sample average. However, the change in y among those who experienced e may not be a satisfactory conceptualization of the effect of e. More specifically, the so-called counterfactual account of causality (Morgan and Winship 2007) – sometimes also referred to as the ‘Neyman-Rubin causal model’ or the ‘potential outcomes approach’ – suggests that the treatment effect of e on y for individual i, say \( \text{TE}_{i} \), should be conceptualized as the difference between changes in y under exposure and non-exposure (see, for example, Gangl 2010a: 23). To express this difference formally, it is necessary to define another variable \( e^*_i \) that is used to represent the two potential states of the world where i experienced e (\( e^*_i = 1 \)) and where i did not experience e (\( e^*_i = 0 \)). Then \( \text{TE}_i \) can be defined as follows:

\[
\text{TE}_i \equiv (y_{it} - y_{io}|e^*_i = 1) - (y_{it} - y_{io}|e^*_i = 0) \equiv \Delta y_{it} - \Delta y_{oi}
\]  

(4.2)

where \( \Delta y_{it} \equiv (y_{it} - y_{io}|e^*_i = 1) \) and \( \Delta y_{oi} \equiv (y_{it} - y_{io}|e^*_i = 0) \). Equation 4.2 conceptualizes the effect of e as the difference between two outcomes (each of which is itself a difference): The change in y conditional on exposure to e minus the change in y conditional on non-exposure to e. The ‘fundamental problem of casual inference’ (Holland 1986) is that only one of these outcomes is observed because individuals either experience e or not: For those exposed to e, we only observe \( \Delta y_{it} \), but not \( \Delta y_{oi} \), and vice versa for those not exposed to e. If we are interested in the effect of e on the ‘treated’, that is, on those who actually experienced it, inferences about their counterfactual trajectories under non-exposure will have to be based on the actual trajectories of individuals who were not exposed to e. The crucial task is to exploit information on the actual trajectories of non-exposed workers in such a way as to obtain good approximations of the counterfactual trajectories exposed workers would have experienced.

---

1 The label ‘treatment effect’ alludes to experimental studies where subjects are randomly allocated to a group receiving a treatment (such as a newly developed medicine) and a control group receiving no or only a placebo treatment.
under non-exposure. In analogy to experimental designs, this can be thought of as the task of finding an adequate control (or comparison) group for the treated.

For most events, the (unobservable) difference defined in Equation 4.2 will be a more convincing conceptualization of their effect than the simple (observable) within-person difference defined in Equation 4.1. The case seems to be different, however, for voluntary retirement. One way to see this is to consider the analogy to the experimental situation and to ask oneself whether it is possible to conceive a suitable experiment for identifying the effect of an event. Even though such an experiment would be unethical, the idea of designing an experiment where we randomly assign workers to a treatment group who lose their job and a control group who do not is perfectly sensible. However, the idea of assigning workers to a treatment group who retire voluntarily appears inherently contradictory because assignment by a third person contradicts the very idea of voluntary retirement. For this reason, I will focus on the simple within-person difference from Equation 4.1 in my analysis of voluntary retirement and on the difference defined in Equation 4.2 in my analysis of late-career job loss. The following discussion of how to conceptualize and estimate the effect defined in Equation 4.2 is therefore primarily relevant to my analysis of income changes around job loss.3

Figure 4.1 illustrates the counterfactual conception of causality using a hypothetical example. The solid line represents the actual income trajectory of a hypothetical person, say Mallory, who was exposed to an event, say late-career job loss, in year 0. Let us assume that Mallory’s reemployment prospects were not too bright and that she therefore retired immediately after job loss. The dashed line represents the counterfactual income trajectory that Mallory would have experienced if she had not lost her job. Under that scenario, she would have enjoyed further earnings/income growth for another three years. She would then have retired (voluntarily) and her income would have declined, though to a lesser extent than it actually did when retiring after late-career job loss (perhaps because she incurred actuarial deductions for early retirement in the latter case). What are

2 The same, of course, holds vice versa if one is interested in estimating the counterfactual trajectories non-exposed workers would have experienced under exposure.

3 This approach accords with extant research: Previous studies on the consequences of labor force exit have generally eschewed comparisons with non- or later-retiring workers (see, for example, Grad 1990; Bardasi and Jenkins 2002; Motel-Klingebiel and Engstler 2008), whereas studies on the consequences of job loss routinely make use of comparisons with non-displaced workers (e.g., Brand 2006; Schwerdt et al. 2010; Dieckhoff 2011; Ehlert 2013).
Mallory’s total income losses six years after job loss? According to the simple within-person comparison (see equation 4.1), total losses are equal to the sum of rectangles I, III, and IV. According to the counterfactual account (see equation 4.2), losses are instead given by the sum of rectangles II, III, IV. During the first three years after job loss, losses are greater than suggested by the within-person comparison (because of foregone earnings/income growth). During years four to six, they are smaller (because voluntary retirement would have led to a decline in income anyway). The counterfactual conception thus isolates the additional income loss attributable to job loss.

Figure 4.1 makes clear that country or period differences in the impact of events can stem from differences in trajectories after exposure and/or differences in trajectories after non-exposure. For example, a decline in income replacement rates should affect the costs of late-career job loss primarily by exacerbating the losses incurred by workers actually exposed to the event, that is, by changing the path of the solid line in Figure 4.1. A general trend toward later retirement, by contrast, can be expected to also and perhaps even primarily affect income trajectories under non-exposure (although it may also affect the trajectories of those exposed to late-career job loss, for example, because they become more likely to reenter
employment). This suggests that it may often be illuminating to investigate treatment and control group trajectories separately rather than only their difference. In my empirical analysis of income mobility around job loss in Chapter 8, I will therefore often examine the actual ‘non-differenced’ trajectories of displaced workers in addition to DID estimates (i.e., differences between displaced workers and a comparison group of non-displaced workers).

It is common to differentiate individuals with respect to their actual exposure to the treatment and to define conditional average treatment effects on that basis (Imbens 2004; Gangl 2010a,b). More specifically, the Average Treatment on the Treated (ATT) and the Average Treatment Effect on the Non-treated (ATN) can be defined as follows:

\[
\text{ATT} \equiv E(\Delta y_{1i} - \Delta y_{0i} | e_i = 1) \tag{4.3}
\]

\[
\text{ATN} \equiv E(\Delta y_{1i} - \Delta y_{0i} | e_i = 0) \tag{4.4}
\]

Finally, the (unconditional) Average Treatment Effect (ATE) can be expressed as a weighted average of ATT and ATN. With \(p(e = 1)\) denoting the (population) share of treated units, the ATE is given by:

\[
\text{ATE} \equiv p(e = 1) \cdot E(\Delta y_{1i} - \Delta y_{0i} | e_i = 1) + (1 - p(e = 1)) \cdot E(\Delta y_{1i} - \Delta y_{0i} | e_i = 0) \tag{4.5}
\]

It is common to differentiate between the population and sample versions of these quantities, that is, between \(\text{PATT}, \text{PATN}, \text{and PATE}\) (with ‘P’ standing for ‘Population’) on the one hand and \(\text{SATT}, \text{SATN}, \text{and SATE}\) (with ‘S’ for ‘sample’) on the other hand (Imbens 2004). Equations 4.3, 4.4, 4.5 define the population versions (because they define average effects in terms of expectations rather than sample averages). The distinction between population and sample variants emphasizes the classical problem of drawing inferences about population parameters from a limited sample. However, there is an important difference here to the problem of estimating, say, population means on the basis of sample means: While no uncertainty is involved in calculating the sample mean of a given variable (at least if we ignore issues of measurement error), identification of sample (average) treatment effects does involve an important element of estimation as it requires identification of the unobserved potential outcomes. This second inferential problem – the
identification of counterfactual outcomes – is in fact the more vexing challenge in drawing inferences about the effect of e.4

Like most applied research, my focus in this study will be on estimating the effect of job loss on those who were actually exposed to it, that is, the ATT or, more precisely, a feasible variant of this effect that is characterized in more detail below. What is required, therefore, is an estimation of the changes in y that displaced workers would have experienced if they had not lost their job.5 As noted above, there is usually no other empirical basis for estimating these trajectories than the actual trajectories of individuals who were not exposed to job loss. However, there are two general reasons why in non-experimental settings the typical trajectory of a non-exposed individual (i.e., E(Δy0|e = 0)), may be very different from the typical counterfactual trajectory of exposed workers (i.e., from E(Δy0|e = 1)) (cf. Gangl 2010a: 25): Outcomes of exposed and non-exposed workers might have differed even in the absence of exposure (Gangl refers to this as ‘heterogeneity’) and the probability of exposure might be related to the size of TE (Gangl refers to this as ‘endogeneity’).

There are different approaches to estimating treatment effects using non-experimental data (for recent overviews, see Morgan and Winship [2007] or Gangl [2010a]). A fundamental question is whether one can hope to fully account for heterogeneity and endogeneity by conditioning on observed characteristics (Morgan and Winship 2007). To ‘fully account for heterogeneity and endogeneity’ here means that the distribution of potential outcomes is independent of treatment status conditional on observed covariates. The assumption that this is the case is often referred to as the conditional independence, ignorability, or unconfoundedness assumption.6 While more or less convincing indirect tests may sometimes be available (cf. Imbens 2004: 21-22), direct tests of the unconfoundedness assumption are impossible because they would require knowledge of the counterfactual outcomes.

4 Because this is the key challenge, it is also generally the case that ‘a good estimator for one ATE [i.e., PATE or SATE, J.P.H.] is automatically a good estimator for the other’ (Imbens 2004: 6).
5 Estimation of ATN or ATE would also require estimation of the trajectories non-exposed individuals would have experienced under exposure.
6 The classical and strongest form of this assumption is that (y0, y1) ⊥ e|X (Imbens 2004: 7), with ‘⊥’ denoting statistical independence. A weaker form that is sufficient for estimating the effect of treatment on the treated is y0 ⊥ e|X, that is, conditional on X, the distribution of outcomes under non-exposure should be the same for treated and control cases (Imbens 2004: 8). Finally, if the focus is only on estimating average treatment effects the y’s and y’s in these equations can be replaced by the corresponding expected values, i.e., E(y0) and E(y1) (Heckman et al. 1997; Imbens 2004).
If unconfoundedness holds, causal effects can in principle be recovered through classical regression analysis (Imbens 2004; Imbens and Wooldridge 2009). However, researchers embracing the counterfactual conception of causality often prefer non- or semi-parametric matching approaches. The most common matching procedure, propensity score matching (PSM; Rosenbaum and Rubin 1983), proceeds by estimating a probability model of treatment assignment (or exposure to an event) and then matches treatment and control cases on the basis of their propensity scores, that is, the predicted probabilities from the assignment model. Matching approaches have some key advantages over classical regression analysis. First, they require researchers to more explicitly address the process of selection into treatment (Gangl and DiPrete 2004). Second, by ‘pairing observationally close, if not identical, observations’ (Gangl 2010a: 31) matching procedures highlight common support problems, that is, situations where reasonably similar comparison cases are missing for some units. Finally, and relatedly, classical parametric regression analysis makes strong assumptions about the functional form of relationships. This may render results highly ‘model dependent’ in the sense of being sensitive to specification choices, especially when extrapolation is involved, that is, when treatment and control cases are very dissimilar and the region of common support is limited (King and Zeng 2006). Ho et al. (2007: 201) therefore describe matching as a form of ‘preprocessing’ that ‘makes estimates based on the subsequent parametric analyses far less dependent on modeling choices and specifications’.

If unconfoundedness does not hold, neither regression analysis nor matching can go all the way towards providing unbiased effect estimates, although they may substantially reduce bias compared to naive (unadjusted) comparisons of treated and non-treated cases. Common methods for identifying causal effects in the absence of unconfoundedness which I will not discuss here are instrumental variables estimation or regression-discontinuity designs (see the overviews mentioned above for further details). In the context of the present study it is more important to highlight an important strength of panel data: Panel data, which provide repeated measurements on individuals (or other units of analysis such as firms or countries) enable researchers to estimate the impact of a variable solely on the basis of within-person (or, more generally, within-unit) variability. As is well known (see, for example, Halaby 2004), this eliminates any bias arising from time-invariant unobserved characteristics that would lead to violations of the unconfoundedness assumption in cross-sectional settings. To see this, consider the following outcome equation for individual i, which is a slight modification of equation 20 in Gangl (2010a: 34):

\[ y_{it} = \alpha + \beta x_{it} + \epsilon_{it} \]

where:
- \( y_{it} \) is the outcome for individual i at time t,
- \( x_{it} \) is the variable of interest for individual i at time t,
- \( \alpha \) is the intercept,
- \( \beta \) is the coefficient for the variable of interest, and
- \( \epsilon_{it} \) is the error term.
Here, $y_{it}$ represents the level of the outcome variable for individual $i$ at time $t$ and $x_{it}$ is a (row) vector of time-varying control variables with associated coefficient (column) vector $\beta$. $w_i$ is a vector of observed time-invariant covariates with associated coefficient vector $\gamma$. $e_i$ is the variable of interest (e.g., an indicator for having recently experienced late-career job loss) with associated (treatment) effect $\tau$. $\alpha_i$ is a time-invariant individual-specific error term that captures the effects of all unobserved time-invariant variables. $\lambda_t$ captures unobserved period effects that can be modeled by including appropriate dummy variables. Finally, $\epsilon_{it}$ is a random individual- and period-specific error term. Importantly, if $\alpha_i$ is correlated with $e_{it}$, cross-sectional estimates of $\tau$ obtained by regressing $y_{it}$ on $x_{it}$, $w_i$, and $e_{it}$ will be biased. For example, cross-sectional estimates of the consequences of late-career job loss based on comparisons between employed and unemployed individuals could be biased because the unemployed are negatively selected with respect to unobserved attributes such as health status, field of study, or cognitive skills. The so-called first-difference (FD) estimator addresses this potential source of bias by analyzing within-person differences:

$$y_{it} - y_{it-1} = (x_{it} - x_{it-1})\beta + (e_{it} - e_{it-1})\tau + (\lambda_t - \lambda_{t-1}) + (\epsilon_{it} - \epsilon_{it-1})$$  

(4.7)

In this equation $\alpha_i$, the person-specific error term capturing the combined impact of all unobserved time-invariant variables, drops out of the equation (and so does the combined impact of observed time-invariant variables whose effects can therefore not be estimated in the FD framework). This is the crucial advantage of within estimation over cross-sectional designs.

As noted above, the analysis of late-career job loss in Chapter 8 will use a special case of FD estimation that is often referred to as difference-in-differences (DID) estimation. The label ‘difference in differences’ comes from the fact that in applications with a binary treatment variable such as exposure to late-career job loss, the quantity of interest is a difference between two differences in the outcome variable $y$, namely the difference between the (average) change in $y$ among the treated and the average change in $y$ among the non-treated. The analogy to the counterfactual effect of an event $e$ defined in equation 4.2 is obvious.

---

7 Fixed-effects (FE) estimation, a common alternative to FD estimation, achieves the same thing through so-called demeaning, that is, by subtracting the unit-specific means of all variables from their unit- and period-specific values (cf. Halaby 2004; Gangl 2010a).
DID estimation is a powerful tool for identifying causal effects, but it still rests on non-trivial assumptions. The crucial assumption required for DID estimation to yield unbiased estimates of the ATT is often referred to as the ‘common trends assumption’ (e.g., Lechner 2011). This assumption requires that, conditional on covariates \( X \), the expected change in the outcome variable for the control group is equal to the expected (counterfactual) change that the treatment group would have experienced in the absence of treatment. Using the potential outcomes framework, this assumption can be expressed as follows (cf. Lechner 2011: 12):

\[
E(y_{i1}|e_i = 1, e_i^* = 0, X = x) - E(y_{i0}|e_i = 1, e_i^* = 0, X = x) = E(y_{i1}|e_i = 0, e_i^* = 0, X = x) - E(y_{i0}|e_i = 0, e_i^* = 0, X = x)
\]  

(4.8)

This assumption is untestable because \((y_{i1}|e_i = 1, e_i^* = 0, X = x)\) and \((y_{i0}|e_i = 1, e_i^* = 0, X = x)\) are unobservable. If \( e \) does not affect the level of \( y \) before its occurrence, \((y_{i0}|e_i = 1, e_i^* = 0, X = x)\) will be equal to \((y_{i0}|e_i = 1, e_i^* = 1, X = x)\) (i.e., the pre-treatment level of \( y \) for individuals who were exposed to \( e \)), which is observable. In general, the existence of pre-treatment effects of \( e \) on \( y \) constitutes a problem for DID estimation (Lechner 2011). If longer panels are available, a simple strategy for minimizing the impact of potential pre-treatment effects is to ensure that pre-treatment measures are not taken too shortly before the occurrence of the treatment. Thus, I will generally use income measures from one or two years before the occurrence of job loss as reference measures in my analysis of late-career job loss.

In my application, a simple comparison of displaced and non-displaced workers would probably not meet the common trends assumption: For example, displacement tends to be concentrated among less-educated workers whom previous research has shown to retire earlier than higher-educated workers even in the absence of displacement. This in turn suggests that displaced and non-displaced workers differ in their overall retirement propensity and hence in their earnings trends under non-exposure. An

---

8 Relatedly, Gangl (2010a: 34) notes that DID estimation requires exogeneity of the differenced error terms, that is, the differenced error term \( \epsilon_{it} - \epsilon_{it-1} \) needs to be uncorrelated with the differenced causal variable \( e_{it} - e_{it-1} \).

9 Such pre-treatment effects may occur because individuals anticipate the occurrence of and \( e \) or because assignment to (or selection into) the treatment depends on the level of \( y \). A prominent example is the assignment to training programs on the basis of earnings, which will result in program participants having exceptionally low earnings prior to the treatment: This phenomenon is known as ‘Ashenfelter’s dip’ (Ashenfelter 1978), and will lead to upward bias in estimated program effects (because of ‘regression to the mean’).
obvious strategy for rendering the common trends assumption more plausible is to control for differences between treated and comparison cases that are thought to be related to trends under non-exposure, that is, to condition on a rich set of variables X. Hence, DID designs are often combined with parametric regression methods or other approaches that adjust for (pre-treatment) differences between treated and comparison units. In the present study, I combine DID estimation with a semiparametric matching technique, that is, I compare (within-person) changes among affected workers to changes among a ‘matched’ control group of observationally similar workers (Gangl 2010b). The goal of matching is to render trends among comparison workers a better approximation to the counterfactual trends that workers exposed to e would have experienced under non-exposure. DID matching is no panacea: Estimated treatment effects will still be biased if there is systematic variation in income trends after accounting for the variables included in the matching procedure. In general, however, matching can be expected to reduce bias compared to naive unadjusted DID estimates, while being sensitive to common support problems and avoiding restrictive functional form assumptions.

DID matching is an increasingly popular approach for addressing problems of causal inference in non-experimental settings. Extensive discussions and comparisons with cross-sectional matching estimators are provided by Heckman et al. (1997), Heckman et al. (1998), and Smith and Todd (2005). Recent studies that use DID matching to investigate the impact of job loss or unemployment on a variety of outcomes include Gangl (2006), Strauß and Hillmert (2011), and Dieckhoff (2011).

4.2 Implementation of DID matching and compositional adjustments

Matching procedure

The crucial task in the application of DID matching is the construction of an adequate control (or comparison) group. Most previous applications of DID matching (and cross-sectional matching) use some variant of propensity score matching (Rosenbaum and Rubin 1983). As discussed above, PSM begins by estimating a probability model (usually a logit or probit model) of being treated. A unit’s propensity score is the predicted probability from this model. In the second step, a control group is constructed by matching treated units to non-treated units with similar propensity scores. Different
algorithms such as \((k)-\)nearest-neighbor matching or kernel matching are used in practice. Caliendo and Kopeinig (2008) provide a thorough discussion of the most common algorithms and other crucial steps in the application of PSM.

As with other matching methods, the ultimate goal of PSM is to eliminate or at least greatly reduce differences between treatment and control cases with respect to the matching variables, that is, the variables used in the probability model (Ho et al. 2007). Rosenbaum and Rubin (1983) therefore refer to the propensity core as one member of a more general class of ‘balancing scores’ that ‘balance’ treated and control units with respect to a given set of covariates (ideally, these will be the variables necessary to achieve unconfoundedness). In practice, it has to be checked post hoc whether a particular PSM solution effectively balances the data. If comparisons of treated and control cases reveal substantial differences with respect to one or more of the covariates, researchers will usually begin anew, running a different specification of the probability model and/or changing the matching algorithm (Caliendo and Kopeinig 2008: 47-49).

A classical alternative to PSM is exact matching (EM) which matches treated units with control units that are exactly identical with respect to the matching variables, that is, treated and control units are matched only if they fall into the same cell of the multidimensional table spanned by the matching variables \(X\). Exact matching has some attractive properties: Treatment and control groups will be perfectly balanced by construction. Exact matching also guarantees similarity of the higher-order moments of \(X\) across treatment and control groups whereas balance checks performed in the context of PSM are often restricted to differences in means (Iacus et al. 2011, 2012). In practice, however, EM is rarely feasible because most data sets will lack exact matches for a large number of treated units. Difficulties to find exact matches rise rapidly (exponentially) with the number of matching variables\(^{10}\), a problem that is sometimes referred to as the ‘curse of dimensionality’.

In this study, I use Coarsened Exact Matching (CEM), a new matching method introduced by Iacus et al. (2011, 2012).\(^{11}\) Like classical exact matching, CEM exactly matches on observable variables in the sense that treated and control units are matched if and only if they belong to the

\(^{10}\) For example, if matching is done on 4 dichotomous variables, there are a total of \(16 = 2^4\) possible combinations. With 8 dichotomous variables, there are \(256 = 2^8\) combinations.

\(^{11}\) The Stata implementation used to obtain the results in this study is described in Blackwell et al. (2009).
same cell – or ‘stratum’ – of the multidimensional table spanned by X. This multidimensional table, however, is not based on the original metric of the variables in X, but rather uses ‘coarsened’ versions of at least some of these variables. Thus, the analyst will typically ‘coarsen’ continuous measures like employer tenure into categorical variables, ideally in ways that are consistent with substantive knowledge about the variable in question. Categorical variables can also be further collapsed. For example, an eleven-point happiness scale could be reduced to a four-category measure (very unhappy, unhappy, happy, very happy). While this procedure leads to somewhat greater dissimilarities between treated and control units than EM, it greatly reduces curse-of-dimensionality-type problems. Within the imbalances permitted by the chosen coarsenings, CEM by construction also balances treatment and control group with respect to ‘all multivariate nonlinearities, interactions, moments, quantiles, comoments, and other distributional differences’ (Iacus et al. 2012: 8).

CEM inherits from EM the property that some treated cases may not be successfully matched because the data include no control units sharing their combination of coarsened variables. There is a clear trade-off here in that wider coarsenings will reduce the number of treated units without matches, yet also permit greater covariate imbalance between treatment and control groups. However, it is worth stressing that problems to find matches for treated units may be more prominent with CEM, but that analysts using PSM also often discard cases without close matches (in terms of the propensity score) by enforcing so-called caliper or common support restrictions (Bryson et al. 2002; Crump et al. 2009).

To the extent that treated units cannot be matched, one will be estimating a restricted or feasible version of the population (or sample) average treatment effect on the treated, that is, one will be estimating F\text{PATT} (or F\text{SATT}) (Iacus and King 2012). However, this restriction to a feasible subset of treated cases will often be rewarded with substantial bias reduction.

To limit the number of unmatched treated units one may often have to use somewhat wider coarsenings than would be desirable under ideal circumstances. Fortunately, it is possible to mitigate this problem by combining matching with conventional regression analysis to adjust for remaining differences between treated and control cases (Blackwell et al. 2009; Iacus et al. 2012). As noted above, the ultimate goal of matching methods is to achieve covariate balance between treated and control units, and in practice this is achieved by applying a set of matching weights (see below). CEM, PSM and other matching methods are thus simply different approaches for obtaining a set of weights that improves covariate balance between treated
and control units. Additional adjustments for remaining differences can be carried out by applying conventional regression methods to the reweighted data, an approach that is sometimes also taken in applications of PSM (e.g., Dehejia and Wahba 1999). Given that the matched and reweighted sample will be much better balanced than the unmatched sample, estimated treatment effects should be relatively robust to the inclusion of additional controls and to different specification choices — that is, they should exhibit substantially less ‘model dependence’ (Ho et al. 2007; Iacus et al. 2012). In my own analysis of the impact of job loss in Chapter 8, I will adopt this ‘matching-plus-regression’ or ‘augmented matching’ approach.

As just noted, CEM and other matching techniques can be viewed as reweighting approaches which seek to ensure that treatment and control groups are ‘balanced’ in the sense of being (reasonably) similar with respect to the matching variables. More specifically, standard CEM assigns the following weight \( w_i \) to matched unit \( i \) in stratum \( s \in S \) — where \( S \) denotes the (sub)set of matched strata that include both treated and control units, and \( T^* \) and \( C^* \) denote the sets of matched treated and control units, respectively (cf. Iacus et al. 2012: equation 6 on p. 8):

\[
w_i = \begin{cases} 
1, & i \in T^* \\
\frac{m_T^s m_C^s}{m_T^s m_C^s}, & i \in C^s 
\end{cases}
\]  

Here, \( m_T^s \) and \( m_C^s \) denote, respectively, the number of treated and control cases in stratum \( s \). \( m_T \) and \( m_C \) denote the total number of matched treated and control units, that is, \( m_T = \sum_{s \in S} m_T^s \) and \( m_C = \sum_{s \in S} m_C^s \). In this formula, the crucial expression for achieving comparability of treated and control units with respect to the coarsened matching variables is \( \frac{m_T^s}{m_C^s} \). This factor ensures that the weight received by a given control unit equals the number of treated units it ‘represents’. \( \frac{m_C}{m_T} \) is a scaling factor that ensures that the CEM-weighted ratio of matched treated and control units equals their unweighted ratio. Unmatched treated and control units are pruned from the analysis.

---

12 A straightforward way of assessing the relative effectiveness of different matching methods therefore is to see how successful they actually are in reducing differences between treated and control units (King et al. 2011).
Incorporation of survey weights

Like many observational data sets, the ones used in this study provide survey weights to correct for design effects (variable selection probabilities) and selective non-response. For example (see Section 4.3 below), low-income households were overrepresented in the original PSID sample, as were recent immigrants in the original SOEP sample. In longitudinal applications it may also be necessary to account for differential attrition rates. This raises the question whether to account for sampling weights when applying CEM. Generally speaking, if treatment effects are heterogeneous in such a way that size of treatment effect and size of sampling weight are correlated (e.g., because persons experiencing severe income losses have lower selection probabilities by design or are more likely to leave the panel), then unbiased estimation of (feasible) population average treatment effects seems to require that analysts take sampling weights into account.

For example, assume that we are interested in estimating the impact of job loss on the risk of entering poverty. If displaced workers have lower incomes than non-displaced workers they would presumably have a higher risk of entering poverty than the average non-displaced worker even in the absence of displacement (e.g., because of other events such as spousal job loss). Matching on predisplacement income addresses this problem by ensuring that displaced workers and comparison cases have similar incomes. However, low-income workers presumably face greater risks of becoming poor because of displacement than higher-income workers (e.g., because their earnings-related unemployment benefits are less likely to lift them above the poverty line). If, like the PSID, the data set used to study the impact of displacement oversamples low-income households, they will be overrepresented among displaced workers in the sample, compared to their share in the ‘population’ of displaced workers. Ignoring survey weights (and thus failing to account for this oversampling), will result in exaggerated estimates of the poverty-triggering effect of displacement.

Despite the increasing popularity of matching approaches, and of PSM in particular, for identifying treatment effects using survey data, there has been little systematic discussion of how to account for sampling weights. In the case of PSM, there are two questions: whether to use weights when estimating the assignment model and whether to account for them in the subsequent estimation of treatment effects (Bryson et al. 2002). In CEM, primarily the latter question arises. Here, researchers using PSM with sampling weights seem to have mostly applied weights to treated cases only (cf. Bryson et al. 2002: 29-30; Dolton et al. 2006: 46). This is also what is
tentatively suggested by Leuven and Sianesi (2012) in the help file for their widely used Stata implementation of psm. The rationale for weighting only the treatment group presumably is as follows: As noted above, matching techniques themselves can be interpreted as reweighting approaches that aim to balance the composition of treated and non-treated units with respect to the matching variables. More specifically, if the quantity of interest is a (feasible) \textit{patt\textsuperscript{13}}, the goal of matching is to reweight non-treated units in such a way that (certain moments of) the reweighted multivariate distribution of the matching variables in the control group approximate the distribution in the treatment group. In this sense, the weights applied to control cases in the final estimation of treatment effects are derivative of the covariate distribution in the treatment group. Sampling weights have to be applied to the treated to ensure that their composition corresponds to the composition of the ‘population’ of the treated. The balance-achieving weights for control units are then determined by applying the matching procedure and need not take sampling weights into account.

However, as Zanutto (2006: 73) notes, it may be the case ‘the weights [...] contain information that is not available in the covariates’. For example, in longitudinal settings, longitudinal weights may capture systematic differences in attrition rates beyond those accounted for by the matching variables. In principle, it may be more compelling to use a richer matching specification that explains these differences (for an analogous argument in the context of traditional regression analysis, see Winship and Radbill 1994). In practice, however, such respecification may be difficult due to missing information or curse-of-dimensionality problems. In this study, I will therefore take the longitudinal survey weights of both treated and control units into account. More specifically, I construct the \textit{cem} weights using a modification of Equation 4.09 that replaces the terms from that equation with their survey-weighted equivalents. Let \( l_i \) denote the longitudinal survey weight for unit \( i \) and let \( T^m \) and \( C^m \) denote the sets of all matched treated and control units, respectively. I then construct \textit{cem} weights as follows:

\[
\begin{align*}
    w_i^* &= \begin{cases} 
    l_i, & i \in T^s \\
    \frac{\sum_{i \in C^s} l_i}{\sum_{i \in T^s} l_i}, & i \in C^s
    \end{cases} \\
    & \quad \text{ (4.10)}
\end{align*}
\]

\textsuperscript{13} If the goal is to estimate the \textit{pate}, sampling weights must also be applied to control cases to ensure that they are representative of the population of interest.
The weight of matched treated units is simply their longitudinal survey weight. As before, unmatched treated and control units are removed from the analysis. The weight assigned to matched control unit $i$ in stratum $s$ is the product of that unit's longitudinal survey weight with two additional factors. The first of these factors is the weighted equivalent of the scaling factor in Equation 4.09, and the second is the ratio of the weighted number of treated units in $s$ to the weighted number of controls in $s$.

Adjusting for compositional changes in the treatment group

A major goal of my analysis is to better understand the implications of recent welfare state change for citizens’ well-being by comparing the financial consequences of retirement and late-career job loss across historical time. But can the period differences emerging from such an analysis be attributed to country-period differences in welfare state arrangements? In Chapter 2, I noted one important alternative possibility, namely that such differences might also reflect the changing composition of those who experience an event: As a simple example, assume that rates of poverty entry after late-career displacement can be shown to have risen over time. Such a result could, for example, be due to displacement being increasingly concentrated among low-income workers near the poverty line – rather than to changes in welfare state provisions that raise the risk of entering poverty conditional on being a low- or high-income worker.

One straightforward way to address the possibility that period differences reflect such compositional changes is to adjust the composition of treated units from different periods so as to match a common reference distribution. Such an adjustment provides an answer to the question: What would period differences in the economic consequences retirement or displacement have looked like, if the composition of displaced workers had not changed? The difference between the unadjusted effects and the composition-adjusted effects is a measure of how important compositional trends have been in shaping observed trends in the effects of events. If a given trend in effects is reinforced after adjusting for compositional changes, one would conclude that compositional trends have attenuated the trend. Conversely, if a trend is attenuated by applying compositional adjustments, one would conclude that compositional trends (rather than welfare state change) have been an important source of observed changes in effects. Other examples of studies that use reweighting methods to adjust for compositional changes are DiNardo’s (1996) pioneering analysis of changes in the American wage...

To implement the compositional adjustments, I employ the entropy balancing (EB) method introduced by Hainmueller (2012). While using a different approach for obtaining the weights, EB serves the same aim as the matching methods discussed above, namely to construct a set of weights that balances two or more groups with respect to a set of covariates. EB is a powerful tool, but can run into convergence problems, particularly with small sample sizes, which is a practical reason why I prefer CEM for matching treated to control units within periods. In addition and as noted above, CEM automatically balances all possible interactions and higher-order moments within the constraints set by coarsening choices. EB does not have this desirable property (although it is of course possible to explicitly include appropriate terms in the EB procedure).

I will generally carry out compositional adjustments for all variables included in the CEM procedure. As for the ‘reference distribution’, that is, the distribution to be approximated by the reweighted distribution of the period-specific treatment groups, I will use the average country-specific composition of treated units over the whole observation period. This differs from the approach taken by DiNardo et al. (1996) and Giesecke et al. (2015), who adjust the composition of workers from later subperiods to match the composition of workers during the earliest subperiod of their observation period. My reason for choosing the average composition instead is that the number of treated cases per period is sometimes quite small. Occasionally, sampling variability might therefore result in rather untypical period-specific samples and I want to avoid using such samples as a benchmark for the compositional adjustments.

Let \( E \) be the set of weights, obtained via entropy balancing, that balances the period-specific matched treatment groups with respect to a reference distribution. Technical details of the algorithm are described in Hainmueller (2012). Let \( e_i \in E \) denote the weight for treated unit \( i \). By straightforward modification of equation 4.10, the CEM weights for the composition-adjusted analysis are then defined as:

\[
W_i^* = \begin{cases} 
  e_i, & i \in T^s \\
  \frac{\sum_{i \in C^s} e_i}{\sum_{i \in T^s} e_i}, & i \in C^s
\end{cases}
\]

As the goal of the EB procedure is to balance the composition of the treatment group, weights for control units are still based on their longitudinal
survey weight \( l \), as are the sums of weights calculated over control units.\(^{14}\) As discussed above, the (augmented) matching approach ensures compositional similarity of the composition-adjusted treatment group and the control group. I obtained weights for the treatment group using the Stata implementation of \( \text{eb} \) developed by Hainmueller and Xu (2012), employing the \text{basewt} option to preserve information contained in the original longitudinal survey weights of the treated units (see Hainmueller [2012] and Hainmueller and Xu [2012] for further details).

4.3 Data and key measures

4.3.1 Data and general sample restrictions

My empirical analysis is based on data from the Panel Study of Income Dynamics (PSID, Hill 1992; Brown et al. 1996) and the German Socio-Economic Panel (SOEP, Wagner et al. 2007), two of the longest-running household panel studies in the world. The original PSID sample was drawn in 1968 and the original SOEP sample in 1984. Several key variables, including the income measures used in this study, come from the Cross-National Equivalent File (CNEF, Frick et al. 2007), which provides consistent and internationally comparable variables based on the original surveys. I use PSID data from survey years 1980 to 2005 and SOEP data from survey years 1984 to 2010. As I further describe below, all income measures refer to the previous calendar year, so the most recent years for which I have income data are 2004 in the American and 2009 in the German case. PSID/CNEF data from the 2007 wave (providing income data for 2006) were available at the time of writing. Unfortunately, exploratory analysis revealed severe problems with a crucial income component – household private pension income – in the 2007 PSID/CNEF data set. Because I was not able to construct a consistent time series from the original PSID variables, I had to exclude the latest wave of PSID/CNEF data from the analysis.

A major difference between the two studies is that the SOEP conducts personal interviews with all adult household members, whereas the PSID

\(^{14}\) The sums of weights calculated over treated units are based on the \( \text{eb} \) weights. However, the sum over all treated units is normalized to equal the sum of the original survey weights (i.e., \( \sum_{i \in T} e_i = \sum_{i \in T} l_i \)). This does not, of course, hold within individual strata: In general, for most or all \( T \neq l \), which is necessary for \( \text{eb} \) to fulfil its function of holding the composition of (matched) treated units constant.
obtains all information from a single interview with the so-called ‘head’ of the household. In opposite-sex couple households, it is usually the man who is assigned the status of head – except in a very small number of situations such as when the man is cognitively impaired. Information on adult household members other than the household head or his partner is very limited. To ensure comparability, I therefore excluded persons other than the ‘head of household’ (Haushaltsvorstand) or his/her partner from the German sample. However, income received by other household members is included in the household income measures from both surveys.

Both PSID and SOEP are based on probability samples of the residential population with oversampling of certain groups (e.g., immigrants in the SOEP, low-income households in the PSID), making the use of sampling weights imperative. New respondents can enter the studies in two principal ways: by moving into households with existing sample members or by being selected for inclusion in refreshment or enhancement samples. As for the first possibility, both SOEP and PSID collect information on persons entering the panel by moving into an existing SOEP / PSID household. However, the two surveys differ in their following rules, that is, in the extent to which persons who are not original sample members are followed when they no longer coreside with an original sample member, where an original sample member is a person that entered the sample through selection into the initial sample or an enhancement/refreshment sample (for details, see Schonlau et al. 2010).

As for the addition of new samples, the SOEP has been refreshed and enhanced repeatedly since the beginning of the study (Wagner et al. 2007). Despite panel attrition, the number of SOEP households roughly doubled from approximately 6,000 to approximately 12,000 households between 1984 and the early 2000s. The PSID did not draw new samples before 1990. As was widely recognized by then, this lack of enhancement samples compromised the representativeness of the PSID, as post-1968 immigration was not reflected in the panel. To alleviate this problem, a so-called ‘Latino sample’ of 2,000 Mexican, Cuban, and Puerto Rican households was drawn in 1990. However, the Latino sample was discontinued completely after 1995, rendering available panels too short for the present study. A smaller ‘immigrant sample’ that included other immigrant groups in addition to Latinos was added to the study in 1997.15 Again, this sample is largely excluded because of short panel lengths. 1997 was also the year when funding constraints forced

---

15 The 1997 sample included 441 households, another 70 households were added in 1999.
the PSID to switch to two-year interview intervals. From 1997 onwards, data were thus only collected in odd-numbered years. As explained below, this causes a number of complications.

Like any non-mandatory household panel study, SOEP and PSID are subject to attrition of households and individual respondents from the panel. Attrition not only diminishes sample sizes. To the extent that it is selective it may also result in biased estimates of population parameters. However, two studies of the impact of attrition in the PSID until the late 1980s found no evidence that it had caused major distortions (Fitzgerald et al. 1998; Lillard and Panis 1998). In addition, longitudinal weights can be an effective means of correcting for selective attrition. In the present study, I use a generic set of ‘comparability optimized’ longitudinal weights whose construction is described in Kohler (2011). These weights are optimized for comparability in two respects. First, individual attrition probabilities used in the construction of longitudinal weights are estimated on the basis of comparable specifications. Second, in the case of the PSID, positive weights are assigned to individuals entering the panel by moving into core sample households, and these weights are constructed on the basis of the corresponding SOEP procedures. The PSID does not provide sample weights for this group.

I generally restrict my analysis to individuals who were 50 or older at the time of the event in question. I also generally require that individuals have worked a non-negligible number of hours (10 or more per week) prior to the occurrence of the event. I focus on changes from one or two years before an event (the preretirement or predisplacement reference year) until four years later (‘four-year changes’) and six years later (‘six-year changes’). The reasons for taking the pre-event measurement one year before the event in some and two years before the event in other cases are related to the PSID’s switch to biennial interviewing (see Section 4.3.2 below). My unit of analysis can be thought of as an episode. An individual i contributes a treatment episode if she experienced e during the observation period and if sufficient income measures from adjacent years are available, so that the income change of interest can be calculated. I define retirement as a singular, non-reversible transition, so a given individual can contribute at most one treatment episode. By contrast, my definition of job loss allows for multiple job losses, so one and the same person can contribute multiple treatment episodes. Further details on the event variables are provided in Section 4.3.2 below.

I include a given (treatment) episode in the sample if it contains sufficient information for computing at least one of the two changes of interest, that
is, a four-year change or a six-year change. The survey weight assigned to an episode (i.e., the term \( l_i \) in equations 4.10 and 4.11 above) equals the longitudinal weight for the final year of that episode. Individuals with missing values on at least one of the variables used in the CEM or EB procedures are excluded from the analysis. These variables are listed and further described in the empirical chapters.

In the analysis of job loss, treatment episodes are matched with control episodes to obtain DID matching estimates of the impact of job loss. Potential control episodes must meet the following requirement: The contributing worker must have been ‘at risk’ of experiencing job loss at the relevant age (see below), but must not have actually experienced job loss in any of the years included in the potential control episode.

A concrete example may be helpful (for simplicity I focus on the case of four-year changes): I generally perform exact matching on age, because it plays a crucial role both for a displaced worker’s reemployment prospects and her (early) retirement options. Now, assume that individual \( i \) experienced job loss at age 58, that sufficient income measures are available to compute four-year income changes (from age 56 to 60), and that no matching variables are missing. A second individual \( j \) provides a potential comparison episode for \( i \)’s four-year episode if the following conditions are met: \( j \) was observed from 56 to 60, was ‘at risk’ of experiencing job loss at age 58, (i.e., worked more than ten hours per week at age 56), but did not actually lose a job between ages 56 and 60, and no relevant data are missing. \( j \) is not required to have remained ‘at risk’ of job loss after age 58 in order to provide a potential comparison episode. For example, \( j \) may have retired (voluntarily) at age 60, but she must have been ‘at risk’ at age 58. In reality, my dataset of course contains many \( i \)’s, that is, many episodes of workers who were displaced at age 58 and were observed long enough for four-year changes to be computed, and many \( j \)’s, that is, many episodes of workers who were at risk of experiencing job loss at age 58, but did not actually experience it during the relevant observation window. These episodes are then matched according to the CEM algorithm outlined above. Details on the matching variables and coarsenings are provided in Chapter 8.

Finally, \( j \) may also contribute potential control episodes for workers displaced at ages other than 58. Assume that \( j \) was observed from age 52

---

16 To ensure that I am comparing well-matched samples of treated and control units for both types of changes, I created separate sets of CEM weights for four-year and six-year changes. Weights generated by simply matching units in the year of the event would become inaccurate as treated and control units drop out of the sample.
until age 62. In that case, j may also contribute a potential control episode for workers displaced at, say, age 54. Again, all that is required is that no relevant data are missing, that j was at risk of experiencing job loss at age 54, and that she did not actually experience it between ages 52 and 56. Under such circumstances, j’s level income at age 56 will be used multiple times. For example, it will be used in computing j’s four-year change from 52 to 56 (a potential comparison change for estimating the impact of job loss at age 54) and it will also be used in computing the four-year change from age 56 to 60 (a potential comparison change for estimating the impact of job loss at age 54). To account for the non-independence of observations arising from this setup, I generally compute cluster-robust standard errors with clustering at the person level. I also obtain such standard errors in the analysis of income changes around retirement where I pool four-year and six-year changes, as they tend to be very similar.

4.3.2 Key measures

4.3.2.1 Income measures
The primary outcome variables examined in this study are different components and aggregates of individual and household income. Both PSID and SOEP collect detailed retrospective information on various types of income for the calendar year before the interview. On this basis, the CNEF provides several cross-nationally comparable income measures. Table 4.1 lists the individual income components and provides definitions of key income aggregates (preceded by an equality sign) such as household pre-government income. Obviously, the income components are not elementary in any fundamental sense, as they could in principle be further disaggregated (e.g., into different types of asset income or public transfers). Total household taxes are not elicited directly from respondents, but estimated using tax simulation programs that are based on relevant regulations and empirical information (e.g., average tax rates for different income brackets). Further details on the simulation of household taxes are provided in the CNEF codebooks (Goebel et al. 2012; Lillard et al. 2012).17

17 For recent survey years, some of the income variables can take negative values in the US. To ensure consistency, I recoded negative values on these income components and aggregates to zero. The variable total household taxes can take negative values in both countries and at all time points and does so for a considerable and increasing number of households in the US, which attests to the (growing) importance of the EITC. I did not recode negative household taxes to zero in order to capture the income-smoothing effects of negative income taxation (cf. Chapter 3).
Table 4.1  Income components and income aggregates

<table>
<thead>
<tr>
<th>Individual labor earnings</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Earnings by other household members</td>
</tr>
<tr>
<td>+ Household asset income</td>
</tr>
<tr>
<td>+ Household private transfers</td>
</tr>
<tr>
<td>+ Household private retirement income</td>
</tr>
<tr>
<td>= Pre-government household income</td>
</tr>
</tbody>
</table>

| = Pre-tax post-transfer household income |
| + Household public transfers |
| + Household public retirement income |

| = Post-government household income |
| + Household (direct) taxes |

To account for differences in household needs (see Chapter 2), I adjust pre- and post-government household income according to the modified OECD scale (Hagenaars et al. 1994). More specifically, I divide these income variables by an equivalence weight that is constructed as follows:

\[ 1 + 0.5 \cdot (HHM_{15+} - 1) + 0.3 \cdot (HHM_{0-14}) \]

where \(HHM_{15+}\) refers to the number of household members ages 15 and above and \(HHM_{0-14}\) refers to the number of household members ages 0 to 14. Adults living on their own thus receive a weight of one, that is, incomes are expressed in terms of the needs of a single-person household.

As discussed in Chapter 2, needs-adjusted post-government (or disposable) income occupies a central place in my analysis because of its presumably close relationship with economic well-being. I also noted that I will focus on two aspects of changes in needs-adjusted disposable income: relative changes with respect to pre-event levels and changes with respect to the poverty line.

My approach for estimating relative changes in the various income measures is straightforward. Consider the case of a worker who experienced event \(e\) in year \(\tau\). One could now consider many different kinds of income changes. For example, one might be interested in the change from two years before until two years after the event, that is, from \(\tau - 2\) to \(\tau + 2\), or in the change from one year before until five years after event, that is, from \(\tau - 1\)
to $\tau + 5$. More generally, given annual income data, the set of potentially interesting changes can be characterized in terms of the distance between the two income measurements $\alpha$ and the timing of the two measurements relative to $\tau$. More formally, let $y_t$ represent income in year $t$. The set of possible $\alpha$-year changes given exposure to $e$ in $\tau$ is then given by:

$$\Delta_{[−a,b]} = y_{\tau+b} - y_{\tau-a} \text{ with } a,b \geq 0 \land a+b = \alpha > 0$$

(4.12)

To analyze relative changes between $\tau - a$ and $\tau + b$, I simply express $y_{\tau+b}$ as a proportion of $y_{\tau-a}$ and then calculate:

$$\Delta_{rel}^{[−a,b]} = \left(\frac{y_{\tau+b}}{y_{\tau-a}} - 1\right) \cdot 100 = \left(\frac{y_{\tau+b}}{y_{\tau-a}} - \frac{y_{\tau-a}}{y_{\tau-a}}\right) \cdot 100$$

(4.13)

The second equality shows that this is a ‘proper’ difference in that it is taken with respect to one and the same measure, namely income expressed as a multiple of income at $\tau - a$. This difference ranges between $-100$ and $+\infty$. I cap this measure at $+100$ to limit the impact of positive outliers.

I examine several statistics that are based on $\Delta_{rel}$. In addition to averages, I will examine its standard deviation to gauge the variability of income changes after an event. I will also estimate the proportion of workers whose income declines by more than a third (‘large loss’) or more than half (‘very large loss’) of pre-event income (see Gosselin and Zimmerman [2008] for a similar approach).

I define poverty as relative income poverty using a threshold of 60% of median needs-adjusted post-government income. As is common in analyses of group-specific poverty risks, this threshold is calculated on the basis of the whole adult population (rather than only on the basis of older workers) and calculated separately by country and year. The 60% threshold is commonly used in European research (e.g., Vandecasteele 2010) and is part of the so-called ‘Laeken indicators’ for social monitoring in the European Union (Atkinson et al. 2002; Krause and Ritz 2006). In the American literature, it is more common to use the federal poverty line (e.g., Johnson et al. 2010) which is an absolute poverty line based on a ‘basket of goods’ approach (Fisher 1992). Even if one believes that the term ‘poverty’ should be reserved for the latter type of measure, the measure used here can still be considered a useful indicator of having low income relative to the population at large. My analyses of poverty risks after the focal events are restricted to persons with pre-event incomes above the poverty line, that is, I examine the effect of the focal events on the likelihood of entering poverty.
In addition to changes in needs-adjusted post-government income, that is, relative changes and poverty entries after taxes and transfers, I will often provide similar results for needs-adjusted pre-government income (i.e., income before public pensions, other public transfers, and direct taxes; see Table 4.1). The comparison of income trajectories before and after taxes and transfers can provide useful information concerning the extent of ‘welfare state buffering’ (e.g., Goodin et al. 1999; DiPrete and McManus 2000; Ehlert 2012). This comparison will often be illuminating, but it is important to acknowledge that it is one piece of a complex puzzle rather than a clear and definitive answer to the question how effectively the welfare state cushions the effects of adverse life events. This is because pre-government trajectories are not independent of the system of taxes and transfers. For example, high public pension replacement rates can be expected to crowd out private retirement savings. Besides, a smaller difference between pre- and post-government income need not signal a ‘weaker’ welfare state, but may also reflect exogenous variation in problem pressures: There will be less need for welfare state buffering when displaced older workers are facing a strong labor market and can easily find well-paying reemployment.

Throughout the analysis I will focus on four- and six-year changes. That is, I will choose a and b such that \(a + b = \alpha \in \{4, 6\}\). While longer-term changes would be interesting, panel attrition and the fact that income b years after an event is only observed if that event occurred at least b years before the end of the observation period severely restrict the possibilities for such an analysis. The focus on even-numbered intervals (i.e., four- and six-year rather than three- or five-year changes) is not motivated by substantive considerations, nor by a whimsical preference of mine, but rather dictated by the PSID’s biennial interview intervals after 1997.

Unfortunately, this is not the only complication arising from the change in the survey interval. This change also affects how accurately the timing of two of my focal events, exit from work and declines in health, can be ascertained (while I do not examine the impact of declines in health in this study, I do use an indicator of negative health shocks to distinguish between voluntary and involuntary retirement, see below). This is because information on the occurrence of these events is not elicited directly from respondents, but instead inferred from changes in ‘state’ variables (annual work hours in the year before the interview and health status at the time.

---

18 I apply the poverty threshold based on median post-government income when examining poverty entries for pre-government income, as this simplifies the interpretation of differences in entry rates before and after transfers.
of interview). For example, I treat a respondent as having retired in $t$ if she reported no or only marginal work hours for a given year $t$ and reported a substantial number of work hours for $t - 2$ (I also require work hours to have been low in $t + 2$, but this is not relevant to the present discussion). My measure of job displacement is not subject to these inaccuracies because it is based on direct questions about the occurrence and timing of certain work-related events (however, as I elaborate below, the change to biennial interviewing leads to another set of problems in constructing the displacement indicator).

Let $s_e$ be the state variable whose change signals the occurrence of event $e$, for example, an indicator of low work hours in the case of retirement\(^{19}\) or a measure of health problems in the case of health shocks. If interviews are conducted annually, I can ascertain whether $s_e$ changed between the interviews in $t - 1$ and $t$, an interval that is usually about one year long. When interviews are conducted only every other year, I can only be sure that a change occurred between the interviews in $t - 2$ and $t$, an interval that is usually about two years long. For simplicity, let me adopt the convention of saying that $e$ occurred in $t$, or that $\tau = t$, if $s_e$ changed between $t - 1$ and $t$, noting that the change in $s_e$ may well have occurred after the interview in $t - 1$ but before January 1 in $t$.\(^{20}\) As just noted, in the case of biennial interviewing, we only know that the change in $s_e$ occurred between the interviews in $t - 2$ and $t$. Assuming that $e$ occurs with constant probability, we know that $\tau = t - 1$ for approximately half of respondents and $\tau = t$ for the other half, but we cannot tell which of the two groups a given respondent belongs to.

Recall that Equation 4.12 defined a specific type of $\alpha$-year change in terms of $a$ and $b$. Importantly, different types of $\alpha$-year changes will be very different if income trajectories follow systematic pre- and/or post-event trends, for example, because some workers reduce their hours before retirement or because ‘health-shocked’ workers recover and increase theirs. As long as we have annual data on both $s_e$ and income, nothing prevents us from calculating all possible $\alpha$-year changes for every individual. With biennial data, however, not only becomes the analysis confined to even-numbered intervals: The timing of $e$ now also inevitably affects the types of changes

---

\(^{19}\) The retirement indicator is based on retrospective information about work hours in the previous calendar year (rather than work hours at the time of interview), but this does not substantively affect the problem described here.

\(^{20}\) Both PSID and SOEP conduct the majority of interviews (approximately 90%) between January and July.
that can be calculated. To see this, recall that \( \tau = t - 1 \) for some (perhaps about half) of the workers with a change in \( s_e \) from \( t - 2 \) to \( t \) and that \( \tau = t \) for the remaining workers. If we now calculate (four-year) income changes from \( t - 2 \) to \( t + 2 \) for these respondents, we will be calculating \( \Delta_{[-1,3]} \) for the first and \( \Delta_{[-2,2]} \) for the second group. Again, without further information, there is no way of deciding which group an individual belongs to. When averaging these four year changes across all exposed (or treated) workers, we will be estimating the following (weighted) average:

\[
\frac{p(\tau = t - 1) \cdot \Delta_{[-1,+3]} + p(\tau = t) \cdot \Delta_{[-2,+2]}}{p(\tau = t - 1) + p(\tau = t)}
\tag{4.14}
\]

where \( p(\tau = t - 1) \) and \( p(\tau = t) \) denotes the proportions of workers for whom \( \tau = t - 1 \) and \( \tau = t \), respectively. In the special case where the probability of \( e \) is constant \( p(\tau = t - 1) = p(\tau = t) \).

In terms of practical implications, this discussion suggests that it is crucial to ensure that the same quantities are estimated when comparing income changes across countries and/or periods with different survey intervals. There are different strategies for achieving this goal. One of the simplest and the one that I will be using in this study is to estimate averages such as the one given in Equation 4.14 also during those periods where annually spaced interviews are available. I provide further details on my approach when describing the construction of the event variables below and in the individual empirical chapters.

### 4.3.2.2 Event variables

Part II of this study explores the economic consequences of retirement in the sense of exit from work and Part III the consequences of involuntary job loss. As noted in Chapter 2, these events are interrelated: Job loss has been shown to be an important trigger of involuntary early retirement. Moreover, there are reasons to suspect that country differences and recent changes in welfare state arrangements have different implications for workers who enjoy smooth late careers and for workers whose careers are interrupted by involuntary job loss or other unexpected events that limit their control over retirement. In my analysis of income trajectories around retirement, I will therefore make a basic distinction between involuntary and voluntary retirees and focus on the latter group of retirees. Consistent with previous research showing job loss and health problems to be the primary triggers of involuntary early retirement, I classify retirement as involuntary if it was preceded by job loss or the onset of health problems (see Barrett and Brzozowski [2010] for a similar approach). Even though I will not examine
the consequences of late-career health shocks explicitly, this event thus plays a crucial role in distinguishing voluntary from involuntary retirees. I now describe the construction of the three event indicators.

Retirement/Exit from work

Previous research has operationalized retirement in various ways, the three most prominent being self-reports (Drobnic 2002), pension take-up (Fasang 2008), and exit from work or the labor force (Radl 2010). In this study, I define retirement as exit from work because I am interested in the economic risks associated with retirement and because the decline in earnings associated with leaving employment is the primary relevant process in that regard. I focus on exit from work (in the sense of actual employment or work hours) rather than exit from the labor force, which includes the unemployed, that is, older people who are not working, but looking for work. This is because unemployment benefits often serve as de facto early retirement benefits (see Chapter 3) and because distinguishing between ‘genuine’ unemployment and unemployment that really is early retirement is exceedingly difficult.

I define retirement as the first prolonged spell without substantial employment after age 50. The analysis is restricted to persons who worked a substantial number of hours around and/or after age 50; a person who has never worked a substantial number of hours for a longer period of time in her later years is hence not ‘at risk’ of retirement. More specifically, to be ‘at risk’ of retirement, a person must have worked at least 15 hours per week on average between ages 48 and 50 or during her first three years in the PSID/SOEP if she entered the study at a higher age. The measure of weekly work hours is obtained by dividing annual work hours by 52. Annual work hours are based on retrospective questions about the previous calendar year and provided in the CNEF (for further details, see Goebel et al. 2012; Lillard et al. 2012).

For persons meeting this criterion, retirement is then defined as occurring in the first year of the first prolonged spell without substantial employment after age 50. More specifically, a person is treated as having retired in a given year \( \tau \) if she worked less than ten hours per week in \( \tau \) and worked less than ten hours per week two years later (i.e., in \( \tau + 2 \)).

Retirement is classified as involuntary if the worker experienced involuntary job

21 This definition does not rule out that a retiree ‘unretires’ (Maestas 2010) in the sense of working a substantial number of hours in years \( t + 4 \) and later, but such unretirements are rare in my data. In both countries, barely more than 5% of workers work ten hours or more in \( t + 4 \).
loss or a negative health shock at any time between two years before and one year after retirement.\textsuperscript{22}

As noted above, certain complications arise because the PSID switched to biennial interviewing in 1997. Thus, work hours are available for the years 1996, 1998, 2000, and so on, but not for the odd-numbered years in between. The same holds for the income measures which are also collected retrospectively for the previous calendar year. Now consider all persons who belong to the sample at risk and reported ten or more work hours for 1996, but fewer than 10 for 1998 and 2000. As discussed above, these persons may have worked fewer than ten hours already in 1997. Let $\tau$ denote the (unknown) actual year of retirement (i.e., 1997 or 1998). If the probability of retiring did not change from 1997 to 1998,\textsuperscript{23} then the four-year income change from 1996 to 2000 will be the change from $\tau - 1$ to $\tau + 3$ for one half and the change from $\tau - 2$ to $\tau + 2$ for the other half of retirees. In order to ensure comparability across countries and periods, I therefore mimic this situation for the country/periods where annual information is in fact available. That is, I compute four-year changes from $\tau - 1$ to $\tau + 3$ for workers retiring in odd-numbered years and from $\tau - 2$ to $\tau + 2$ for those retiring in even-numbered years. A completely analogous argument applies to six-year changes.

### Job loss

The analysis of late-career job loss is restricted to individuals who worked at least ten hours per week before losing their job and who were between ages 51 and 65 at the time of job loss. Involuntary job loss is defined as an involuntary separation for one of the following reasons: closure of a business or establishment, being fired, or the end of a fixed-term contract if the latter was accompanied by at least one month of unemployment, either in the year when the contract ended or in the following year. Information on the occurrence of job loss events is based on a set of broadly comparable questions concerning respondents’ recent job history that has been administered by both PSID and SOEP in all of the years included in this study. Self-employed workers are excluded from the analysis because the event is not well-defined for them.

\textsuperscript{22} The possibility that one of these events occurred after the year of retirement is primarily relevant in the context of health shocks, because retirees are mostly no longer at risk of losing a job.

\textsuperscript{23} For a given pair of adjacent years this assumption may sometimes be violated, but in the long-run retirement probabilities should not differ systematically between odd and even-numbered years.
The PSID’s switch to two-year interview intervals in 1997 again creates difficulties because some job events were no longer recorded after the change. The complications arise because the reference period of PSID’s job history questions was not changed after 1997: As before, respondents interviewed in $t$ were asked to report job events that had occurred in the year before the interview (i.e., in $t - 1$) or before the interview in $t$ (i.e., between January 1 and the day of the interview). Before the switch to biennial interviewing, information on job events occurring after the day of the interview in $t$, but before January 1 of the following year ($t + 1$) were thus recorded at the $t + 1$ interview. After the switch, however, no interviews were conducted in $t + 1$. Information on job events that occurred after the day of the interview in a given interview year is therefore missing. Hence I partially imputed the job loss indicator for interview years from 1997 onwards (i.e., $t_{\text{int97+}} \in \{1997, 1999, 2001, 2003, 2005\}$).

Because I do not distinguish between workers with only one and workers with multiple displacements in a given year, imputation was not necessary for those who reported having lost their job between January 1 and the day of the interview in $t_{\text{int97+}}$. To impute the displacement indicator for those who did not report having been displaced in the current year, I obtained a single imputation using the *mi impute logit* routine in *Stata 12*. The relevant outcome can be thought of as the likelihood of experiencing displacement in $t_{\text{int97+}}$ conditional on not having experienced it before the interview in that year. I therefore estimated the imputation model over all observations (i.e., including those from years before 1997) who did not report having been displaced in $t$ at the interview in $t$. The outcome variable in this model is whether respondents reported having been displaced in $t$ at the following interview in $t + 1$. This information is of course completely missing for observations from interview years 1997 and later. In addition to a large set of income- and employment-related variables, the imputation model includes dummies for the month of interview to account for the obvious relationship between interview timing and the likelihood of being displaced during the remainder of the year. To capture cyclical fluctuations as well as secular trends in displacement risks, I also included measures of overall displacement rates for adjacent years, that is, for $t - 1$ and $t + 1$. Job events during these years were still completely recorded after the switch to biennial interviewing.

Unlike with the retirement indicator, it thus remains possible to identify the exact year when a job loss occurred even after the PSID’s switch to biennial interviewing (subject, of course, to the uncertainty of the imputation), but there is no real gain from this greater precision in event timing.
because the income measures are available only for even-numbered years. Letting $\tau$ denote the year of job loss, I can therefore only compute four-year changes from $\tau - 1$ to $\tau + 3$ for workers displaced in odd-numbered years and from $\tau - 2$ to $\tau + 2$ for workers displaced in even-numbered years. The difference to the retirement case is that it is possible to tell, for each individual worker, which one of the two changes is calculated. For ease of presentation, however, I will proceed just as in the case of retirement and generally present averages of these two types of four-year changes. Again, as in the case of retirement, I will also do so for those country/periods where annual data are available in order to ensure comparability.

**Health shocks**

The indicator of negative health shocks captures relatively abrupt declines in health that show some persistence. I do not examine income or employment trajectories around health shocks in this study. However, as noted above, I do use information on health shocks to distinguish voluntary from involuntary retirees. Generally speaking, I define a health shock as having occurred in $\tau$ if a respondent reported ‘bad health’ in $\tau$ and $\tau + 2$, but did not report bad health in $\tau - 2$ and $\tau - 4$. The requirement that changes in health be persistent can be thought of as a proxy for severity (Schimmel and Stapleton 2012).

Unfortunately, it is not possible to construct the underlying measure of ‘bad health’ in a fully comparable fashion. In the American case, I use the disability variable provided by the cnef which classifies individuals as disabled if they report a physical or nervous condition that limits the amount of work they can do. The cnef treats German respondents as disabled if they report a legally recognized disability of at least 30%. However, in comparison to the US, this variable alone yields implausibly low shares of disabled persons. I therefore also classify German respondents as having bad health if they meet at least one of the following conditions (see Burkhauser and Daly [1998] for a similar approach and for empirical evidence suggesting that the resulting measure is broadly comparable to the American): They report very low health satisfaction (0-3) on an eleven-point scale or they report that their overall health strongly limits their ability to perform everyday tasks. At the same time, I raise the threshold for the attested disability measure from 30 to 50%, which is the official German threshold for being considered as ‘severely disabled’.

24 Although not specifically referring to (paid) work, this question is the one that is most similar to the psid’s. Unfortunately, it is missing for several waves.