Empirical Monte Carlo Evidence on Estimation of Timing-of-Events Models*

Stefano Lombardi† Gerard J. van den Berg‡ Johan Vikström§

November 8, 2018

Abstract

This paper uses an Empirical Monte Carlo simulation approach to study estimation of Timing-of-Events (ToE) models. We exploit rich Swedish data on unemployed individuals with information on participation in a training program to simulate placebo treatment durations. We then estimate ToE models by omitting some of the covariates previously used to simulate the placebo treatments. This generates unobserved heterogeneity correlated across the treatment ad outcome durations. When estimating ToE models, we use a discrete support point distribution for the unobserved heterogeneity, and we compare different specifications of the model. We find that the model performs well, in particular when time-varying covariates in the form of calendar-time variation are exploited for identification. For the discrete support distribution, we find that both over-correcting for unobserved heterogeneity with too many mass points and under-correcting with too few mass points leads to large bias. We also find that information criteria that penalize parameter abundance are a very useful way to select the number of support points. On the other hand, information criteria characterized by little penalty should be avoided because they lead to over-correction problems.

JEL-codes: C14; C15; C41; J64

*We thank Oskar Nordström Skans, seminar participants at IFAU, and conference participants at EEA and EALE for useful suggestions. Vikström and Lombardi acknowledge support from FORTE.

†Uppsala University, IFAU-Uppsala and Uppsala Center for Labor Studies.

‡University of Bristol, IFAU-Uppsala, IZA, ZEW, CEPR and J–PAL.

§IFAU-Uppsala, Uppsala Center for Labor Studies and Uppsala University.

1 Introduction

The Timing-of-Events (ToE) approach as of Abbring and van den Berg (2003) concerns the identification of the effect of a treatment given while in an initial state on the time spent in the same state. The authors specify a Mixed Proportional Hazard (MPH) model and establish conditions under which all parts of the model, including the treatment effect, are non-parametrically identified. One key feature of the model is that it allows both the exit rate from the initial state and the treatment rate to be affected by correlated unobserved determinants. This is one reason why the ToE approach has been used in many settings, in particular when quasi-experimental approaches are unfeasible. One early example is provided by Abbring et al. (2005), who study the effect of benefit sanctions on time in unemployment. In their setting, the time to a benefit sanction (treatment) and time in unemployment (outcome) are affected by related unobserved factors. Other studies include Lalive et al. (2008), Richardsson and van den Berg (2013), and van der Klaauw and van Ours (2013) on the effect of active labor market programs; Van Ours and Williams (2009, 2012), and McVicar et al. (2017) on cannabis use; van Ours et al. (2013), van den Berg and Gupta (2015), Palali and van Ours (2017) on health settings; Bijwaard et al. (2014) on migration; Jahn and Rosholm (2013) on temporary work; and Baert (2013) on over-education. Many other papers have used the ToE in the field of labor economics. These include Crépon et al. (2013), Caliendo et al. (2016), Busk (2016), Lindeboom et al. (2016), Holm et al. (2017), and Bergemann et al. (2017).

Several factors must be taken into account when estimating the ToE model. First, the model is often specified by approximating the unknown bivariate unobserved heterogeneity distribution by means of a discrete distribution (Lindsay, 1983; Heckman and Singer, 1984). In practice, however, this could be done in several ways. One is to pre-specify a (relatively low) number of support points and increase their number until computational problems arise. Alternatively, one could use information criteria to select the number of support points. Second, sample size may also be a relevant factor, since estimation of non-linear MPH models can be problematic with small samples. Third, different sources of exogenous variation can be used when estimating the model, such as variation from time-varying covariates.

In this paper, we use a new simulation design based on actual data to evaluate these and related specifications issues that arise when estimating the ToE model. To this end, we modify the novel Empirical Monte Carlo design (EMC) proposed by Huber et al. (2013). In their study, they compare different methods to estimate treatment effects under unconfoundedness.[1] The key idea is to use actual data on treated units to simulate placebo treatments for non-treated units, and then base the simulations on these placebo treatments. This ensures that the true treatment effect is zero, that the selection model is known, and that the unconfoundedness assumption holds by construction. The fact that real data is used instead of a data generating process chosen by the researcher makes the entire simulation exercise arguably more relevant for real applications.

In our simulation design, we use rich administrative data of Swedish unemployed individuals, with information on participation in a training program (the treatment). For each unemployed, we create detailed background information in the same vein as in Lechner and Wunsch (2013). We estimate a duration model for the time to treatment using data on both treated and non-treated individuals, and we use the estimated model to simulate placebo treatment durations for each non-treated unit. By construction, the effect of these placebo treatments is zero, and the treatment assignment process is known. With these simulated data we estimate alternative ToE models. The key aspect of our simulation design is that when estimating ToE models we exclude subsets of the covariates that were used to simulate the placebo treatments, and this generates a bivariate duration model with correlated unobserved determinants. This is because the excluded covariates were used to generate the placebo treatments, and the same covariates also affect the outcome duration. This new simulation design allows us to examine the ToE specification issues using simulations based on actual data.

How to best specify the distribution of unobserved heterogeneity has been an actively researched question for a very long time. Initial simulation evidence is provided by Heckman and Singer (1984), Ridder (1987), and Hu and Sickles (1994). More recently, Baker and Melino (2000) study a univariate duration model with unobserved heterogeneity and duration dependence. One of their main conclusions is that model specifications with too many support points over-correct for unobserved heterogeneity (through an overly-dispersed unobserved heterogeneity distribution), which leads to bias in all model components. Gaure et al. (2007) also use simulated data to examine a bivariate duration model similar as the one analyzed in this paper. One of their findings is that a discrete support points approach is generally very reliable if the sample

[1] Other studies using the EMC simulation design include Huber et al. (2016) on the performance of parametric and semiparametric estimators commonly used in mediation analysis; Frölich et al. (2017) study the performance of a broad set of semi and nonparametric estimators for evaluation under conditional independence; Lechner and Strittmayer (2017) compare different procedures to deal with common support problems; and Bodory et al. (2016) consider several inference methods for matching and weighting methods.

is large and there is some exogenous variation in the hazard rates. In particular, they highlight that calendar time is a particularly useful source of exogenous variation. On the other hand, unjustified restrictions, such as pre-specifying a very low the number of support points for the unobserved heterogeneity, may cause substantial bias.

Our study adds to this evidence by using a simulation design based on actual data. This leads to several conclusions. First, by excluding covariates from the model, the estimated treatment effect is severely biased. However, with already two support points, large part of the bias is eliminated. Second, we find a substantial risk of over-correcting for unobserved heterogeneity. With a large number of support points the average bias is more than twice as large as with a few support points, and the variance of the estimates increases with the number of support points. Third, we find that information criteria are a very useful way to select the number of support points. In particular, the Akaike information criterion (AIC), Bayesian information criterion (BIC), and Hannan-Quinn information criterion (HQIC) all perform well. They protect against over-correction by penalizing parameter abundance. They also guard against under-correction by rejecting models specified with little or no unobserved heterogeneity mass points. Fourth, information criteria imposing very little penalty for parameter abundance, such as those solely based on the maximum likelihood (ML criterion), should be avoided altogether. This is because they favor models with too many support points, and this leads to over-correction problems. Moreover, in some specifications the less restrictive AIC criteria outperforms the more restrictive BIC and HQIC criteria, but the opposite occurs in other settings. This means that no criterion is superior in all settings.

In our analyses we mainly focus on the above mentioned specification choices. However, the simulation results taken as a whole also indicate that the ToE model indeed is able to adjust for a large proportion of the bias. This holds in our baseline model, where the only source of exogenous variation derives from time-fixed covariates. With time-varying covariates in the form calendar-time variation, the bias is further reduced and almost equal to zero. This holds even when the setting is characterized by substantial heterogeneity induced by omitting large set of covariates, including a wide range of short- and long-term labor market history variables. This echoes the results from Gaure et al. (2007).

As a background to the main analyses, and in a similar way as in Lechner and Wunsch (2013), we evaluate the relevance of different set of covariates in terms of bias reduction. We find that short-term labor market history variables are particularly important. Moreover, adjusting for employment history is relatively more important than doing so with unemployment, earnings and welfare history. We also find that controlling for information about long-term labor market history (last 10 years) on top of short-term history (last two years) does not meaningfully reduce the bias. This latter result is in line with what found in Lechner and Wunsch (2013).

This paper is also related to several other strands of the literature. In the spirit of Lalonde (1986), two recent papers by Kastoryano and van der Klaauw (2011), and Muller and van der Klaauw (2017) evaluate the ToE model and other dynamic evaluation approaches by comparing their estimates with an experimental benchmark.

The paper proceeds as follows. In the next section we present the Timing-of-Events model proposed by Abbring and van den Berg (2003). Section 3 describes the simulation design and the data used in the simulations. Section 4 describes the selection model that is used to simulate the placebo treatments and presents bias results when different sets of covariates are included in the model. In Section 5 we present the EMC simulation results. Section 6 concludes.

2 The Timing-of-Events model

This section presents the Timing-of-Events (ToE) approach as of Abbring and van den Berg (2003). They specify a bivariate duration model for the duration in an initial state and the duration until the treatment of interest: T_e and T_p, with t_e and t_p being their realisations. In the model we have observed individual characteristics, X, and unobserved individual characteristics V_e and V_p, with realizations (x, v_e, v_p). Abbring and van den Berg (2003) assume that the exit rate from the initial state, $\theta_e(t|D(t), x, V_e)$, and the treatment rate, $\theta_p(t|x, V_p)$, follow the Mixed Proportional Hazard (MPH) form, where t is the elapsed duration:[2]

$$\begin{aligned} \ln \theta_e(t|x, D, V_e, t_p) &= \ln \lambda_e(t) + x'\beta_e + \delta D(t) + V_e, \\ \ln \theta_p(t|x, V_p) &= \ln \lambda_p(t) + x'\beta_p + V_p, \end{aligned} \tag{1}$$

where $D(t)$ is an indicator function taking the value one if the treatment has been imposed before t, δ represents the treatment effect, and ($\lambda_e(t)$, $\lambda_p(t)$) capture duration dependence. Also, let G denote the joint distribution of $V_e, V_p|x$ in the inflow into unemployment.

Abbring and Van den Berg (2003) show that all components of this model, including the treatment effect, δ, and the unobserved heterogeneity distribution, G, are identified

[2]This is the most basic ToE model, but note that Abbring and Van den Berg (2003) also allow for time-varying treatment effects as well as other extensions of this basic model.

under the following assumptions. The first assumption is no-anticipation, which means that future treatments are not allowed affect current outcomes. This holds if the units do not know the exact time of the treatment or if they do not react on such information.[3] A second assumption is that X and V should be independently distributed, implying that the observed characteristics are uncorrelated with the unobserved characteristics. A third assumptions is the proportional hazard structure (MPH model). These assumptions are discussed in more detail when we describe our simulation design. Abbring and van den Berg (2003) also impose several regularity conditions.

Identification is semi-parametric, in the sense that given the MPH structure, the ToE model does not rely on any other parametric assumptions. Moreover, unlike many other approaches, the Timing-of-Events method does not require any exclusion restrictions. Instead, identification of the treatment effect follows from the variation in the moment of the treatment and the moment of the exit from the initial state. If treatment is closely followed by an exit from the initial state, regardless of the time before the treatment, then this is evidence of a causal effect, while any selection effects due to dependence of V_p and V_e do not give rise to the same type of quick succession of events. However, this requires some exogenous variation in the hazard rates. The most basic exogenous variation is generated through the time-invariant covariates, x, in the model, which create variation in the hazard rates across units. Strictly speaking, this is the only variation that is needed for identification.

Previous studies suggest that calendar-time variation, for instance due to business cycle variation and seasonal variation, is a useful and more robust source of additional exogenous variation (Eberwein et al., 1997; Gaure et al., 2007). The intuition is that calendar-time variation and other time-varying covariates shift the hazard rates, and this helps to identify the influences of the unobserved heterogeneity. More specifically, current calendar-time factors have an immediate impact on the exit rate, whereas past calendar-time factors affect the current transition probabilities only through the selection process (see van den Berg and van Ours, 1994; 1996 for a more detailed discussion). In this paper, we first simulate and estimate models with only variation deriving from time-fixed observable covariates. Later, we exploit calender-time variation in the form of time-varying local unemployment rate.

[3] The no-anticipation assumption also implies that any anticipation of the actual time of the exit from the initial state do not affect the current treatment rate.

3 Simulation approach

3.1 The basic idea

The idea behind EMC is to simulate using real data instead of using a data generating process that is entirely specified by the researcher, such as in a typical Monte Carlo study. The argument is that real data is more closely linked to real applications with real outcomes and real covariates, and provides arguably more convincing simulation evidence. As a background to our simulation design, consider the EMC design adopted by Huber et al. (2013). They use real data on job-seekers in Germany to compare the performance of alternative estimators of treatment effects under conditional independence. They proceed in the following way. Initially, the real data of both treated and non-treated units is used to capture the treatment selection process. The estimated selection model is then used to simulate placebo treatments for all non-treated units in the sample, effectively partitioning the sample of non-treated into placebo treated and placebo controls. This ensures that the selection process used for the simulations is known and that the conditional independence assumption holds by construction, even if real data is used in the simulations. Moreover, by construction, the true effect of the placebo treatments is zero. Huber et al. (2013) use the resulting simulated data to analyze the performance of various propensity score estimators.

We tweak this simulation design in some key dimensions with the aim of using the EMC approach to study the ToE model. Our simulations are also based on real data. We use rich Swedish register and survey data of unemployed individuals, with information on participation in a labor market training program. The outcome duration, T_e, is the time in unemployment, while the treatment duration, T_p, is time to the training program. The data (described below) is also used to create detailed background information for each unit in the same vein as Lechner and Wunsch (2013). We use this data to generate placebo treatments, but we do this in a slightly different way than Huber et al. (2013). In particular, instead of simulating binary treatment indicators as they do, we use an hazard model for the treatment duration, and use this to simulate placebo treatment durations. As for the standard EMC approach, the effect of these placebo treatments is zero by construction. Unobserved heterogeneity is then generated by leaving out blocks of the covariates used in the true selection model. That is, we leave out some covariates that were used when generating the placebo treatment durations. This leads to a bivariate duration model with correlated unobserved determinants, since the excluded covariates affect time in unemployment (the outcome) and, by construction, the treatment duration.

With the resulting simulated sequences based on real data, we perform various simulation exercises. Our aim is to evaluate several specification issues arising when estimating ToE models. Some of these issues were raised by previous Monte Carlo simulations studies (Gaure et al., 2007; Baker and Melino, 2000), such as the specification of the unobserved heterogeneity distribution and of the baseline hazard. In our analyses we exclude different blocks of covariates, with the aim of studying how the ToE approach performs with different types of unobserved heterogeneity. The simulations are also performed under different simulation designs, for instance, with and without exogenous variation in the form of calender-time variation.

One important reason to use the Swedish unemployment spell data is that there are many examples of evaluations that estimate ToE models using unemployment spell data.[4] The use of unemployment spell data also affects how we design our simulation study. Unemployment durations and labor market program entries are typically measured at the daily level. We therefore use a discrete-time hazard model to generate the placebo treatment durations measured at the daily level. However, we use a continuous-time ToE model, implicitly assuming that the daily spell data is approximately continuous. There are several reasons for doing this. First, continuous-time-models is often used in the literature, even with daily data.[5] Second, Abbring and van den Berg (2003) establish identification results for continuous-time data and it is unclear to what extent the identification results carry over to discrete and interval-censored data. Third, Gaure et al. (2007) estimate both continuous- and discrete-time models and conclude that a discrete-time model outperforms the continuous-time model if the data is truly discrete. Despite this, the continuous-time model used in this paper delivers some optimistic results for the performance of the ToE model.

Since we do not simulate the job transitions, the true effect of the treatment is zero. By construction, the no-anticipation assumption holds, because the units cannot anticipate and react to placebo treatments. However, there are other ToE assumptions that may not hold in this simulation design. First, the assumption requiring independence between X and V (random effects assumption). This may not hold in our simulations, since the excluded covariates representing unobserved heterogeneity may be correlated with the covariates that were actually used in the ToE estimation. Second, since the

[4] Examples include Abbring et al. (2005), Lalive et al. (2005), Røed et al. (2006), Lalive et al. (2008), Kyyrä (2010), Richardsson and van den Berg (2013), Kyyrä et al. (2013), Arni et al. (2013), and Van den Berg and Vikström (2014).

[5] Several works estimate continuous-time models using discrete data (daily, weekly, monthly or yearly): Palali and van Ours (2017), Tatsiramos (2010), Jahn and Rohsolm (2013), Kyyrä et al (2013), McVicar et al (2018), Muller et al. (2017), van Ours and Williams (2009), van Ours and Williams (2012), and van Ours et al (2013).

outcome duration is not modeled, the outcome hazard (re-employment rate) may not follow the MHP structure. Third, although we use a very rich set of covariates to estimate the selection process, if there are some omitted characteristics then the model will be miss-specified.

All these three potential violations of the ToE assumption arise because we use a simulation design based on real data, which most likely does not follow a MPH structure. However, one may argue that this is the benefit of our approach, because we explore estimation of the ToE model using arguably more realistic data.

3.2 Data and the training program

Training program. One often studied treatment for job-seekers is labor market training. This motivates our use of data on a Swedish vocational training program called AMU (Arbetsmarknadsutbildning). The program and the administrative data that we use resemble those of other countries. The main purpose of the program, which typically lasts for around 6 months, is to improve the skills of unemployed individuals so to enhance their chances of finding a job. Important courses include manufacturing, machine operators, office/warehouse work, health care, and computer skills. Previous evaluations of the training program include Harkman and Johansson (1999), de Luna et al. (2008), Richardson and van den Berg (2013), and Vikström and van den Berg (2017). These papers describe the program in more detail.[6]

Data sources and sampling. We combine data from several administrative registers and surveys. The Swedish Public Employment Service provides daily unemployment and labor market program records of all unemployed individuals in Sweden. We use this information to construct spell data on the treatment duration (time to training program) and the outcome duration (time to employment), both measured in days. We sample all unemployment spells starting during the period of 2002–2011.[7] The analyses are restricted to the prime-age population (age 25–55) since younger workers are subject to different labor market programs and to avoid early retirement decisions. We also exclude disabled workers. In total, the sampled spells are 2.6 million, of which 3% involve participation in training. The mean unemployment duration in the sample is 370 days. In case a job-seeker enters into training multiple times, only the first one is considered.

[6]The basic eligibility criterion is to be 25-year old. During the training, participants receive a grant. Those who are entitled to unemployment insurance (UI) receive a grant equal to their UI benefits level, while for those not entitled to UI the grant is lower-sized. In all cases, training is free of charge.

[7]Any ongoing spells are right-censored in December 31, 2013.

Covariates and outcome measure. For each spell, we construct detailed information on individual-level characteristics (summarized in Table 1). We follow Lechner and Wunsch (2013) and construct similar variables as in their study. The population register called LOUISE provides basic socio-economic information, such as country of origin, civil status, regional indicators and level of education. Matched employer-employee data (RAMS) and wage statistics from Statistics Sweden are used to construct information on the characteristics of the last job (wages, type of occupation, skill-level), and to retrieve information on the characteristics of the last firm (firm size, industry and average worker characteristics). Data from the Public Employment Service is used to construct unemployment history variables. It is also used to construct information on the regional unemployment rate. Earnings records and information on welfare participation are used to construct employment, out-of-labor force and earnings histories. For the history variables, we construct both short-run history (last 2 years) and more long-run history (last 10 years). Altogether, this captures all aspects of the workers employment and earnings history in the last 2/10 years. From Unemployment Insurance (UI) records we obtain information on UI eligibility.[8]

The outcome considered in this paper is the re-employment rate (job exit rate). We consider as an exit to employment a transition to a part-time or full-time job that is maintained for at least 30 days.

Since our aim is to estimate duration models, we use some additional variables that are not used by Lechner and Wunsch (2013). In particular, we use information on the time in the last unemployment spell and an indicator for having at least one previous unemployment spell. The idea is that previous unemployment durations may capture important aspects of unobserved heterogeneity. This also allows us to compare the relative importance of controlling for employment history, unemployment history, and duration related information on the last unemployment spell.

Sample statistics. Table 1 presents sample statistics for the variables used in the simulations. The table shows that immigrants from outside Europe, males, married and less educated are overrepresented among the training participants. Training participants also more likely to be employed in firms with lower wages, and there fewer previous managers and more mechanical workers among the treated workers. All labor market history measures point in the same direction: training participants have worse unemployment and welfare characteristics in the last 2 and 10 years.

[8]Due to differences between the Swedish data and the German data, there are some differences as compared to Lechner and Wunsch (2013) data. The classification of occupations differs, we lack some firm-level characteristics and we have less information on UI claims. We also use information on welfare benefits to construct measures of out-of-labor-force.

3.3 Simulation details

Selection model. The first step of the EMC design is to estimate the treatment selection model. In our duration framework with daily information on unemployment and program entries, we use a discrete-time hazard model and estimate a complementary log-log model for the treatment hazard, $\theta_p(t|x)$, at time, t, conditional on a set of time-invariant covariates, x:

$$\theta_p(t|x) = 1 - \exp[-\exp(\lambda_p(t) + x\beta_p)]. \tag{2}$$

The baseline hazard, $\lambda_p(t)$, is taken as piecewise constant, with $\ln \lambda_p(t) = \alpha_m$ for $t \in [t_{m-1}, t_m)$, where m is an indicator of the mth time interval. We use eight time intervals, with splits after 31, 61, 122, 183, 244, 365 and 548 days. The observed variables, X, include all covariates described in Table 1. The model estimates reported in Table 1 show that the daily treatment rate peaks after roughly 300 days. They also confirm the same patterns found for the sample statistics: immigrants, younger workers, males, high-school graduates, and UI recipients are more likely to be treated. Short- and long-term unemployment and employment history variables are also important determinants of treatment assignment.

After estimating the selection model using the full population of actual treated and controls, the treated are discarded and play no further role in the simulations. The next step is to use (2) to simulate the placebo times to treatment for each non-treated. To this aim, we perform a transition lottery for each unit, where the realized time to treatment is generated by comparing each units estimated treatment probability with random drawings from uniform distributions.[9] Simulated treatments that occur after the actual exit from unemployment are ignored. Thus, the placebo treated are those with a placebo treatment before the exit to job. During this procedure, $\hat{\theta}_p(t|x_i)$ is multiplied by a constant γ, which selected such that the share of placebo treated is around 20%. This assures that there is a fairly large number of treated in each sample. A similar approach is adopted by Huber et. al (2013).

The above steps describe the baseline simulation design, where the selection model ignores calendar-time variation (besides inflow dummies). In later simulations, we also exploit calender-time variation in the form of time-varying local unemployment rate. To this end, another set of placebo treatments is created where calendar time is taken into account by adding spell-varying local unemployment rate to (1).

[9] For each time-interval, θ_p is compared to a random uniform realization $u \sim U[0,1]$. An exit to treatment occurs if $\theta_{pit} < u$.

Simulations. The placebo treatments are simulated for all non-treated. Next, we draw random samples of size N from this full sample (independent draws with replacement). We set N = 10,000, 20,000 and 40,000 because ToE models are rarely estimated with very small sample sizes. If the estimator is N-convergent, increasing the sample size by a factor of 4 (by going from 10,000 to 40,000) should reduce the standard error by 50%. Later, we consider larger samples with N = 80,000 and 160,000. For each specification we perform 500 replications.

3.4 Implementation of the bivariate duration model

We estimate a continuous-time ToE model for the treatment and outcome hazards as defined in Equation (1). The unknown distribution of the unobserved heterogeneity is approximated by a discrete mass points distribution (Lindsay, 1983; Heckman and Singer, 1984; Gaure et al., 2007).

Likelihood function. For each unit $i = 1, \ldots, N$ we formulate the conditional likelihood contribution, $L_i(v_i)$, conditional on the vector of unobserved variables $v_i = (v_{ei}, v_{pi})$. Then, the individual likelihood contribution, L_i, is obtained by integrating $L_i(v_i)$ over the distribution of the unobserved heterogeneity, G. For the duration dependence ($\lambda_e(t)$, $\lambda_p(t)$), we use a piecewise constant specification with $\lambda_s(t) = \exp(\alpha_{sm})$ where the spell-duration indicators are $\alpha_{sm} = \mathbb{1}\left[t \in [t_{m-1}, t_m)\right]$, for $m = 1, \ldots, M$ cut-offs. In the baseline setting we fix the cut-offs to $31, 61, 122, 183, 244, 365, 548, 2160$. The actual covariates X used in the model are explained below.

To set up $L_i(v_i)$, we split the spells into parts where all right-hand side variables in (1) are constant. Splits occur at each new spell-duration indicator and when the treatment status changes. Later, calendar-time variation lead to additional (monthly) splits. Spell-part j for unit i is denoted by c_{ij}, and has length l_{ij}. Let C_i be the set of spell-parts for unit i. Each part, c_{ij}, is fully described in terms of l_{ij}, α_{sm}, x_i and the outcome indicator, y_{sij}, which equals one if the spell part ends with a transition to state s and zero otherwise. There are two two possible states (job exit and treatment start). Then, with approximately continuous durations, $L_i(v_i)$ is:

$$L_i(v_i) = \prod_{c_{ij} \in C_i} \left[\exp\left(-l_{ij} \sum_{s \in S_{it}} \theta_s(t, x_i, D_{it}, v_{si}|\cdot) \right) \times \prod_{s \in S_{it}} \theta_s(t|\cdot)^{y_{sij}} \right], \qquad (3)$$

with

$$\theta_s(t|\cdot) = \begin{cases} \lambda_e(t)\ \exp(x_i'\beta_e)\ \exp(\delta D_{it})\ v_{ei} \\ \lambda_p(t)\ \exp(x_i'\beta_p)\ v_{pi}. \end{cases}$$

L_i is obtained by integrating $L_i(v_i)$ over $G(V)$. Let p_w be the probability associated with support point, w, with $w = 1, \ldots, W$, such that $\sum_{w=1^W} p_w = 1$. Then, the log-likelihood function is:

$$\mathcal{L} = \sum_{i=1}^{N} \left(\sum_{w=1}^{W} p_w \ln L_i(v_w) \right) \equiv \sum_{i=1}^{N} L_i. \tag{4}$$

Search algorithm. In order to estimate the discrete support points, we use the iterative search algorithm in Gaure et al. (2007).[10] For each replication we estimate models with up to $\overline{W}$ support points. We can then select the appropriate model using alternative information criteria (see below). Let $\hat{\vartheta}_W$ be the maximum likelihood (ML) estimate with W support points. The search algorithm is:

Step 1: Set $W = 1$ and compute the ML estimate $\hat{\vartheta}_W$.

Step 2: Increment W by 1. Fix all ϑ_W elements but (v_W, p_W) to $\hat{\vartheta}_{W-1}$. Use the simulated annealing method (Goffe et al., 1994) to search for an additional support point, and return the $(\tilde{v}_W, \tilde{p}_W)$ values for the new support point.

Step 3: Perform ML maximization with respect to the full parameters vector $\vartheta_W = (\beta, v, p)$ by using $\hat{\vartheta}_W$ and $(\tilde{v}_W, \tilde{p}_W)$ as initial values. Return $\hat{\vartheta}_W$.

Step 4: Store $\{\hat{\vartheta}_W,\ \mathcal{L}(\hat{\vartheta}_W)\}$. If $W < \overline{W}$ return to Step 2, else stop.

Step 1 corresponds to a model without unobserved heterogeneity, since $\hat{v}$ cannot be distinguished from the intercept in x. In *Step 2* the algorithm searches for a new support point in the $[-3, 3]$ interval.[11] In this step, all other parameters of the model are fixed. This explains why in *Step 3* we perform a ML maximization over all parameters, including the new support point. At the end of the procedure we obtain $\overline{W}$ ML estimates: $\{\hat{\vartheta}_W,\ \mathcal{L}(\hat{\vartheta}_W)\}_{W=1}^{\overline{W}}$.

Information criteria. We use different approaches to choose between the $\overline{W}$ estimates. First, we report results where we pre-specify the number of support points (1–6 points). An alternative approach is to increase the number of support points until there is no further improvement of the likelihood (maximum likelihood (ML) criterion). We also use information criteria that penalize parameter abundance. Specifically, the

[10]Estimations were performed by using Matlab R2017a Parallel Computing Toolbox on resources provided by the Swedish National Infrastructure for Computing (SNIC) at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

[11]As starting values we set $v_W = 0.5$ and $p_W = \exp(-4)$. The simulated annealing is stopped once it finds a support point with a likelihood improvement of at least 0.01. In most cases, the algorithm finds a likelihood improvement within the first 200 iterations.

Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the Hannan-Quinn information criterion (HQIC). The latter two are more restrictive since they impose a larger penalty on parameter abundance. Formally, $AIC = \mathcal{L}(\hat{\vartheta}_W) - k$, $BIC = \mathcal{L}(\hat{\vartheta}_W) - 0.5k \cdot \ln N$ and $HQIC = \mathcal{L}(\hat{\vartheta}_W) - k \cdot \ln(\ln N)$, where $k \equiv k(W)$ is the number of estimated model parameters and N is the total number of spell parts used for the estimation.[12] The ML criterion is defined as $ML = \mathcal{L}(\hat{\vartheta}_W)$, where only likelihood increases greater than 0.01 are considered. The criteria are calculated for each replication, so that the selected number of support points may vary both across replications and criteria. This allows us to compute the average bias and the mean square error for each information criteria.

4 Selection model and different sets of covariates

Before proceeding to the simulation results, we evaluate the relevance of different sets of covariates in a similar way as Lechner and Wunsch (2013). The aim is to assess the relative importance of different covariates in the evaluation of training program for unemployed individuals. Specifically, we include various blocks of covariates and compare the size of the bias of the estimated treatment effect across specifications, knowing that the true effect of the placebo treatments is zero (see Table 1 for a list of the covariates in each block). All covariates are a subset of those used to generate the placebo treatments. Depending on the covariates included, each time we create a different type of unobserved heterogeneity.[13,14] Note that we use similar covariates as in Lechner and Wunsch (2013), so that we can examine to what extent their results extend to other countries. One obvious difference is that we consider a duration outcome framework. For each block of covariates, the full sample of placebo treated and placebo non-treated are used to estimate parametric proportional hazard (PH) models, and then we compare the resulting bias across specifications.

[12] We follow Gaure et al. (2007) and use N_c. One alternative is to use N instead, but our simulations indicate that this is of minor importance in practice.

[13] Lechner and Wunsch (2013) provide good arguments as to why conditional independence (CIA) should be valid in their German setting when they use their full set of covariates. This allows them to study which covariates need to be adjusted for. We acknowledge that CIA may not hold in our setting, for instance because of treatment selection based on unobserved motivation and skills. In a strict sense, this means that the relative importance of different types of covariates are evaluated without arguing that this captures all covariates relevant in the selection process.

[14] Before Lechner and Wunsch (2013), several other papers have examined the importance of different covariates. Some studies have compared experimental estimates to non-experimental ones that control for different sets of covariates. Other studies use novel survey data to assess the importance of usually unobserved variables such as personality traits.

Results. Table 2 reports the estimated biases of the effect of training when the model is extended sequentially by adding blocks of covariates.[15] Panel A begins with a model with a set of baseline socio-economic characteristics, which returns a positive and sizable bias of around 8.4% (hazard rate estimates from a PH model). Then, Panel A stepwise includes controls for calendar time and regional conditions (regional dummies and local unemployment rate), but these covariates turn out to be relatively less important (bias is reduced from 8.4% to 7.5%). Here, the excluded covariates include short- and long-term labor market history, so that the positive bias means that training participants tend to have more favorable labor market histories.

In Panel B, we adjust for short-term labor market history. First, we adjust for the duration of the last unemployment spell.[16] The idea is that previous unemployment durations may capture important aspects of unobserved heterogeneity. However, it turns out that adjusting for previous unemployment duration only reduces the bias from 7.5% to 7.0%. All other blocks of short-term labor market history variables also reduce the bias. However, adjusting for short-term employment history is relatively more important than adjusting for unemployment, earnings and welfare history (out-of-labor-force). If we adjust for unemployment history and earnings history, the bias drops to 6.7% and 5.0%, respectively, whereas if the model includes employment history the bias is close to zero. In fact, the sign of the bias is even reversed (slightly negative, -1.9%) when adjusting for short-term employment history. We conclude that participants in labor market training to a large extent are selected based on their previous employment records. One explanation may be that caseworkers aim to select job-seekers with an occupational history appropriate to the vocational training program.

Next, Panel C of Table 2 shows that adding information about long-term labor market history (last 10 years) on top of short-term history (last 2 years) has minor impact on the bias of the estimated treatment effect. The same holds when in Panel D we adjust for various characteristics of the last job (e.g., previous wage and occupation) as well as for detailed information about the last firm (e.g., industry and composition of worker).

Lechner and Wunsch (2013) also find that, after controlling for calendar time, regional conditions and short-term labor market history, adding additional covariates such as long-term labor market history is relatively unimportant. However, one difference compared to their study is that in this setting adjusting for short-term employment

[15] The order adding covariates is similar to that used by Lechner and Wunsch (2013), who argue that the order resemble the ease, likelihood and cost of obtaining the respective information for researchers.

[16] We also include an indicator for any previous unemployment spell within the last five years.

history is enough to obtain small bias, while Lechner and Wunsch (2013) find that it is important to also adjust for all aspects of the short-term history (employment, unemployment, out-of-labor-force status, earnings, and non-firm characteristics of the last job) in order to obtain a low bias.

Finally, we take a closer look at the block of short-term employment history covariates, and add these separately together with the baseline covariates. The aim is to understand which aspects of employment history are important. The results from this exercise are reported in Panel A of Table 3 and show that several employment covariates single-handedly capture a large part of the bias. For instance, by only adjusting for months employed in the last six months before the unemployment spell reduces the bias from 7.5% to 2.5%. Other employment history covariates have similar impact on the estimated bias. However, in all cases the bias is positive, so that the reversal of the sign of the bias from a positive to a negative bias seems to occur only once all short-term employment history variables are included together. This suggests that these covariates capture different aspects of the treatment selection. As a comparison, Panel B of Table 3 reports estimates from a similar exercise where we control for the short-term unemployment history variables one at a time. In this case we find that these covariates have a modest impact on the estimated bias.

5 Main simulation results

This section presents the main simulation results. The main focus is on the estimation of the (placebo) treatment effect. We study to what extent the ToE model is able to adjust for the bias observed in the previous section and which specification of the model that leads to the best results. Results are presented in the form of average bias, variance of the estimates, and mean squared error (MSE). Here, the only source of exogenous variation is due to variation in the observed covariates. In Section 5.1, we examine exogenous variation in the form of calendar-time variation.

Baseline results. Table 4 reports results from the baseline simulations where we compare different specifications of the discrete unobserved heterogeneity distribution. In these simulations we adjust for baseline socio-economic characteristics, calendar time and regional indicators (covariates in Panels A–B, Table 1). First, consider the results for a sample size of 10,000 in Columns 1–3. In Panel A, we fix the number of support points to a pre-specified number in all replications. The first row shows that the baseline model without any unobserved heterogeneity (1 support point) leads to large bias (7.4%).[17] This confirms that under-correcting for unobserved heterogeneity may lead to substantial bias.

However, already with two support points the bias is reduced from 7.4% to 1.7%.[18,19] For three or more support points, the average bias is increasing and always goes in the same direction when adding additional support points. In fact, with six support points the average bias (3.8%) is more than twice as large as the average bias with two support points (1.8%). Moreover, the variance as well as the MSE are increasing in the number of support points (Columns 2–3). The increased bias due to too many support points is in line the results from Baker and Melino (2000), who argue that specifications with too many (spurious) mass points can over-correct for unobserved heterogeneity. This happens because too many support points lead to an overly dispersed distribution of unobserved heterogeneity. Thus, in order to fit the data, this is compensated by changes (bias) in the treatment effect, and presumably also to the duration dependence. This pattern contradicts the general intuition that one should adjust for unobserved heterogeneity in the most flexible way in order to avoid bias due to unaccounted unobserved heterogeneity.

To better understand the result with over-correction with too many spurious support points, Figure 1 shows the distribution of the treatment effect estimates for 1, 2 and 6 support points. With one support point, the estimates are centered around a bias of around 7% and the variance of the estimates is rather low. With two support points the entire distribution shifts towards zero (the average bias is non-zero), but the variance is larger than for one support point. With six support points, there is a further increase in the variance. Perhaps more importantly, the entire distribution of the estimates shifts to the right (larger positive bias). This means that the increased bias is not explained by a few extreme estimates. Instead, the overly-dispersed distribution of the unobserved heterogeneity has a more general effect for almost all replications.

Interestingly, the problem with over-correcting for unobserved heterogeneity does not occur to the same extent with the simulated data used by Gaure et al. (2007). They

[17] This is roughly the same bias as in the corresponding model estimated with the full sample in Panel A of Table 2. The minor difference is due to sampling variation since here we report the average bias from random drawings, whereas estimates in Table 2 are for the full the sample of placebo treated and placebo non-treated.

[18] Here, we focus on the bias of the treatment effect, but previous simulation studies using simulated data confirm that failing to account for unobserved heterogeneity also leads to biased spell-duration and covariate effects (see e.g. Gaure et al., 2007).

[19] Gaure et al., 2007 also find a sharp reduction of the bias with two support points in their setting with simulated data. However, they also point out that a low number of support points does not recover the true duration dependence (spell-duration effects).

highlight that the main problem is under-correction with too few support points.[20] Our simulation results based on real data, instead suggests that both under- and over-correction are important problems when estimating ToE models. Thus, finding a way to identify the appropriate number of support points appears to be very important.

Information criteria. Next, Panel B of Table 4 provides simulation results when the distribution of the unobserved heterogeneity (number of support points) is specified by using alternative information criteria. Panel C also reports the average number of support points that were selected for each criterion. The ML criterion, according to which the number of support points is increased as long as the likelihood is improved, leads to 4.38 support points on average. The bias and variance are large compared to simply pre-specifying a low number of support points. Hence, the ML criterion tends to select too many support points, leading to an over-correction problem (too many spurious support points are included). On the other side, the results for AIC, BIC and HQIC are much more encouraging. All three criteria produce models with rather few unobserved heterogeneity support points (often two support points). In this setting, this corresponds to the specifications with the lowest bias, compared to any pre-specified number of support points. We conclude that these more restrictive information criteria protect against over-correction problems due to too many support points. They do this by penalizing the number of parameters in the discrete heterogeneity distribution. They also guard against under-correction problems (too few support points) by favoring models with unobserved heterogeneity over models without unobserved heterogeneity (1 support point).

A comparison between the AIC, BIC and HQIC criteria reveals rather small differences. As expected, the two more restrictive information criteria (BIC and HQIC) lead to models with fewer support points, but the average bias is virtually the same as for the less restrictive AIC criterion. The variance is lower for BIC and HQIC than for AIC. This is because these more restrictive criteria tend to select fewer support points and the variance of the estimated treatment effects is increasing in the number of support points. However, when the covariates that are excluded from the model (different unobserved heterogeneity) are varied, there are cases where BIC and HQIC tend to under-correct for unobserved heterogeneity, and this leads to larger bias than for the less restrictive AIC criteria.

The main interest here is in providing background information on the alternative

[20] In their main simulations, Gaure et al. (2007) find no evidence that too many support points over-corrects for unobserved heterogeneity. However, when they reduce the sample size they also find evidence of over-correction. Here, the sample size is 10,000 observations, but it will also be shown that over-correction also is a problem with larger sample sizes.

specification choices. However, Table 4 also provides some insights on the overall idea of using ToE models to adjust for unobserved heterogeneity. Overall, the table shows that the ToE approach corrects for a large share of the bias due to unobserved heterogeneity. The bias is reduced from 7.4% for the model without unobserved heterogeneity to around 1.6% when the information criteria are used to select the number of support points (see Column 1 of Table 4). This holds even though the only source of exogenous variation in these baseline simulations are due to variation in the observed covariates, X. In subsequent analyses, we explore if additional sources of exogenous variation in the form calendar-time variation are able to eliminate the bias entirely.

Sample size. In Columns 4–9 of Table 4, the sample size is increased to 20,000 and 40,000 observations, respectively. For both these sample sizes we see that two support points are associated with the lowest bias, but here the increase in the bias after three support points is smaller than for 10,000 observations. For instance, with 10,000 observations, going from 2 to 6 support points increases the bias from 1.7% to 3.8% and with 40,000 observations, it increases from 1.9% to 2.5%. This pattern is confirmed by the simulation results in Table 5, where the sample size is increased to 80,000 and 160,000. With these sample sizes, there is virtually no increase in the average bias with three or more support points (see e.g., Column 4 for the bias with 160,000 observations). This suggests that over-correction due too many support points is mainly a problem with small samples. Note that what constitutes a small sample size most likely differs across applications. For instance, the number of parameters in the model, the fraction of treated, the number of exit states, and the variation in the observed covariates may be important. Another result is that for larger sample sizes there are smaller differences between the ML criterion and the three other information criteria. For instance, with a sample size of 160,000, there are virtually no differences in the average bias between the four information criteria.

Excluded covariates. We next vary the unobserved heterogeneity by excluding different sets of covariates when estimating the ToE models. In the baseline simulations, the ToE model includes baseline socioeconomic characteristics, inflow time dummies and regional indicators. Here, we generate more unobserved heterogeneity by excluding additional covariates (all the socioeconomic characteristics reported in Panel A of Table 1) and less heterogeneity by excluding fewer covariates (previous earnings). Table 1 shows that these models generate a bias of 10% and 5.0%, respectively, in the full placebo treatment sample of (Panel and B). These values can be compared to the bias in the baseline setting of 7.5%. Before presenting the simulation results, we characterize the new distribution of the unobserved heterogeneity, which can be computed in the

EMC framework. A comparison with the baseline model shows that leaving out more covariates from the model leads to substantially more dispersion of the unobserved heterogeneity for the treatment rate (Figure 2) and for the job exit rate (Figure 2).

Panel A of Table 6 shows that for the model with more extensive unobserved heterogeneity, the average bias with one mass point (the specification that does not adjust for unobserved heterogeneity) is 10.1%, but it drops to 1.6% with two support points. Again, this shows that the ToE model adjusts for a large share of the bias due to unobserved heterogeneity. As before, we also see that over-correction may occur if too many spurious support points are included in the model. This leads to increased bias beyond three support points. The results in Panel B also confirm that it is important to use an appropriate information criterion to select the number of support points. The ML criterion leads to a model with too many support points, while the other three criteria perform much better. In this case with more extensive unobserved heterogeneity, all three criteria select a larger number of support points than in the baseline setting ((Panel C). All three criteria also give a lower bias than for two support points (the number with the lowest bias). For instance, the average bias with two support points is 1.6% whereas for the AIC criterion, the bias is 1.0%. This is because the AIC criterion on average favors 2.46 support points, so that it selects models with 2 or 3 support points, depending on the data in each replication.

These simulations also result in somewhat larger differences between the AIC, BIC and HQIC criteria. For the AIC criterion, the bias is 1.2% and for the BIC and HQIC it is around 1.6%. This means that the more restrictive information criteria (BIC and HQIC) may under-correct for unobserved heterogeneity by favoring models with too few support points, and this leads to larger bias. However, for the specification where we exclude fewer covariates (Columns 4–7, Table 6) the pattern is different. Here, the average bias is lower for the more restrictive BIC and HQIC criteria than for AIC. This is because for this specification, there seems to be a larger risk of over-correcting for unobserved heterogeneity, favoring information criteria with a larger penalty for parameter abundance. From this, we conclude that neither one of the information criteria is superior in all settings.

Finally, note that the variance of the estimates is lower for BIC and HQIC than for AIC (Column 2). This is because BIC and HQIC tend to select fewer support points and the variance is increasing in the number of support points. In some cases, this may create a trade-off between the bias and the variance in the choice of the information criterion.

Summary of findings. One finding is that criteria with very little penalty for parameter abundance, such as the ML criterion, should be avoided altogether., This is because it tends to favor models with too many support points, which leads to problems with over-correction for the unobserved heterogeneity. Another finding is that no information criterion among AIC, BIC and HQIC is superior in all settings. All three penalize parameter abundance, and this protects against problems of over-correction due to spurious support points. In some cases, the risk of under-correcting is relatively more important, and this favors the less restrictive AIC criterion. In other cases, instead, the risk of over-correcting is more important, and this favors the more restrictive BIC and HQIC criteria. Thus, using all three criteria and reporting several estimates as robustness check appear to be a reasonable approach.

5.1 Exogenous variation

Identification of the ToE model requires variation in the observed exogenous covariates, because this produces "exogenous" variation in the hazard rates. This was the only source of exogenous variation exploited in the baseline simulations above. It resulted in several insights on how to specify the unobserved heterogeneity when estimating ToE models. Another result was that the ToE model was able to adjust for a large part of the selection due to unobserved heterogeneity, but it did not eliminate all of the bias. This is one reason why we now consider additional sources of identification such as calendar-time variation (time-varying local unemployment rate).[21] The calendar-time variation should be useful for identification since it generates a shift in the hazard rates that helps to recover the distribution of the unobserved heterogeneity. To this end, we first re-estimate our treatment selection model by also including time-varying local unemployment rate. Then, as before, we simulate placebo-treated and non-treated durations, and finally we estimate ToE models, now also including the time-varying local unemployment rate in the model.

The results from this exercise are presented in Columns 1–3 of Table 7. The first row of Panel A shows that the bias without adjusting for unobserved heterogeneity (1 support point) is 6.9%. This is slightly smaller than the bias in Table 4 for the baseline setting. This is because the time-varying local unemployment rate is included in the selection model, and this leads to a slightly different selection process.[22] As before, additional support points are then stepwise included (Panel A). The results

[21]The time-varying unemployment rate is measured at monthly level and varies across counties (län). We refer to it as local unemployment rate.

[22]The estimates from the selection model are reported in Table A.1 in the Appendix. The estimate for the time-varying local unemployment rate is 0.0194.

confirm what was found in the baseline simulations. First, if we under-correct for unobserved heterogeneity (no unobserved heterogeneity) this leads to sizeable bias; if we over-correct for unobserved heterogeneity the bias is also large. Second, the ML criterion tends to select models with an overly-dispersed unobserved heterogeneity distribution, which is associated with large bias. Third, the three criteria that penalize parameter abundance (AIC, BIC and HQIC) all perform well, since they lead to models characterized by low bias.

However, one important difference compared to the baseline simulations is that the average bias for the three information criteria now is virtually zero. This confirms that exploiting time-varying covariates greatly helps the identification. Note that this result holds even though we have generated substantial and complex heterogeneity by omitting a large number of covariates, including a wide range of short- and long-term labor market history variables, as well as firm characteristics and characteristics of the last job. This produced substantial bias in the model without unobserved heterogeneity. The importance if calendar-time variation echoes the results from Gaure et al. (2007), who reach a similar conclusion. The only difference being that they use calendar-time dummies whereas we exploit time-varying local unemployment rate.

6 Conclusions

In this paper, we have modified a recently proposed simulation technique, the Empirical Monte Carlo approach, to evaluate the Timing-of-Events model. This method allowed us to exploit rich administrative data information to generate realistic placebo treatment durations, overcoming the common critique that standard simulation studies are sensitive to the DGP chosen by the researcher.

For ToE models, one issue is the specification of the discrete support points distribution for the unobserved heterogeneity. From our simulations, we conclude that information criteria are a very reliable way to specify the support points distribution in the form of the number of support points to include in the model. This holds as long as the criteria includes a substantial penalty for parameter abundance. Information criteria with very little penalty for parameter abundance, such as the ML criteria, should be avoided altogether. Three criteria, which all perform well, are the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the Hannan-Quinn information criterion (HQIC). All three protect both against over-correction for unobserved heterogeneity (due to the inclusion of spurious support points) and against under-correction due to insufficient correction for unobserved heterogeneity. No single criterion are superior in all settings. All three protect both against the inclusion of spurious support points (unobserved heterogeneity over-correction) and against the inclusion of too few mass points (unobserved heterogeneity under-correction). No single criterion are superior in all settings.

Another key conclusion is that ”exogenous” variation in the form calendar-time variation is a very useful source of identification. This result holds even though ToE models that only rely on variation in the observed covariates also tend to produce good results, as long as an appropriate information criterion is used.

In the paper we have also evaluated the relevance of different sets of covariates. In this case the main conclusion is that it is important to adjust for short-term labor market histories when evaluating labor market program for unemployed workers, whereas adding long-term labor market histories are unimportant. This consistent with the results in Lechner and Wunsch (2013).

References

Abbring, Jaap H. and Gerard J. van den Berg (2003a). "The identifiability of the mixed proportional hazards competing risks model". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.3, pp. 701–710.

— (2003b). "The nonparametric identification of treatment effects in duration models". In: *Econometrica* 71.5, pp. 1491–1517.

Abbring, Jaap H., Jan C. van Ours, and Gerard J. van den Berg (2005). "The effect of unemployment insurance sanctions on the transition rate from unemployment to employment". In: *The Economic Journal* 115.505, pp. 602–630.

Arni, Patrick, Rafael Lalive, and Jan C. Van Ours (2013). "How effective are unemployment benefit sanctions? Looking beyond unemployment exit". en. In: *Journal of Applied Econometrics* 28.7, pp. 1153–1178.

Baert, Stijn, Bart Cockx, and Dieter Verhaest (2013). "Overeducation at the start of the career: Stepping stone or trap?" en. In: *Labour Economics* 25, pp. 123–140.

Baert, Stijn, Bas van der Klaauw, and Gijsbert van Lomwel (Feb. 2018). "The effectiveness of medical and vocational interventions for reducing sick leave of self-employed workers: Interventions for Reducing Sick Leave of Self-Employed". en. In: *Health Economics* 27.2, e139–e152.

Baker, Mikael and Angelo Melino (2000). "Duration dependence and nonparametric heterogeneity: A Monte Carlo study". In: *Journal of Econometrics* 96, pp. 357–393.

Berg, Gerard J. van den and Sumedha Gupta (Mar. 2015). "The role of marriage in the causal pathway from economic conditions early in life to mortality". en. In: *Journal of Health Economics* 40, pp. 141–158.

Berg, Gerard J. van den and Johan Vikström (2014). "Monitoring job offer decisions, punishments, exit to work, and job quality". In: *The Scandinavian journal of economics* 116.2, pp. 284–334.

Bergemann, Annette, Laura Pohlan, and Arne Uhlendorff (Aug. 2017). "The impact of participation in job creation schemes in turbulent times". en. In: *Labour Economics* 47, pp. 182–201.

Bijwaard, Govert E., Hans van Kippersluis, and Justus Veenman (2015). "Education and health: The role of cognitive ability". In: *Journal of Health Economics* 42, pp. 29–43.

Bijwaard, Govert E., Christian Schluter, and Jackline Wahba (July 2014). "The Impact of Labor Market Dynamics on the Return Migration of Immigrants". en. In: *Review of Economics and Statistics* 96.3, pp. 483–494.

Bijwaard, Govert E. and Qi Wang (Feb. 2016). "Return Migration of Foreign Students". en. In: *European Journal of Population* 32.1, pp. 31–54.

Bodory, Hugo et al. (2016). "The finite sample performance of inference methods for propensity score matching and weighting estimators". In: *IZA Discussion Papers, No. 9706*.

Busk, Henna (2016). "Sanctions and the exit from unemployment in two different benefit schemes". In: *Labour Economics* 42, pp. 159–176.

Caliendo, Marco, Steffen Künn, and Arne Uhlendorff (Oct. 2016). "Earnings exemptions for unemployed workers: The relationship between marginal employment, unemployment duration and job quality". en. In: *Labour Economics* 42, pp. 177–193.

Cockx, Bart and Matteo Picchio (2013). "Scarring effects of remaining unemployed for long-term unemployed school-leavers". en. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176.4, pp. 951–980.

Crépon, Bruno et al. (June 2018). "Information shocks and the empirical evaluation of training programs during unemployment spells". en. In: *Journal of Applied Econometrics* 33.4, pp. 594–616.

Frölich, Markus, Martin Huber, and Manuel Wiesenfarth (2017). "The finite sample performance of semi- and non-parametric estimators for treatment effects and policy evaluation". en. In: *Computational Statistics & Data Analysis* 115, pp. 91–102.

Gaure, Simen, Knut Røed, and Lars Westlie (2012). "Job search incentives and job match quality". en. In: *Labour Economics* 19.3, pp. 438–450.

Gaure, Simen, Knut Røed, and Tao Zhang (2007). "Time and causality: A Monte Carlo assessment of the timing-of-events approach". en. In: *Journal of Econometrics* 141.2, pp. 1159–1195.

Goffe, William L., Gary D. Ferrier, and John Rogers (1994). "Global optimization of statistical functions with simulated annealing". In: *Journal of Econometrics* 60.1-2, pp. 65–99.

Graaf-Zijl, Marloes de, Gerard J. van den Berg, and Arjan Heyma (2011). "Stepping stones for the unemployed: the effect of temporary jobs on the duration until (regular) work". en. In: *Journal of Population Economics* 24.1, pp. 107–139.

Ham, John C. and Samuel A. Rea Jr (1987). "Unemployment insurance and male unemployment duration in Canada". In: *Journal of Labor Economics* 5.3, pp. 325–353.

Heckman, J. and B. Singer (1984). "A Method for minimizing the impact of distributional assumptions in econometric models for duration data". In: *Econometrica* 52.2, p. 271.

Høgelund, Jan, Anders Holm, and James McIntosh (2010). “Does graded return-to-work improve sick-listed workers’ chance of returning to regular working hours?” en. In: *Journal of Health Economics* 29.1, pp. 158–169.

Holm, Anders et al. (2017). “Employment effects of active labor market programs for sick-listed workers”. In: *Journal of Health Economics* 52, pp. 33–44.

Huber, Martin, Michael Lechner, and Giovanni Mellace (2016). “The finite sample performance of estimators for Mediation Analysis under Sequential Conditional Independence”. en. In: *Journal of Business & Economic Statistics* 34.1, pp. 139–160.

Huber, Martin, Michael Lechner, and Conny Wunsch (2013). “The performance of estimators based on the propensity score”. en. In: *Journal of Econometrics* 175.1, pp. 1–21.

Huh, Keun and Robin C. Sickles (1994). “Estimation of the duration model by non-parametric Maximum Likelihood, Maximum Penalized Likelihood, and Probability Simulators”. In: *The Review of Economics and Statistics* 76.4, p. 683.

Jahn, Elke and Michael Rosholm (2013). “Is temporary agency employment a stepping stone for immigrants?” en. In: *Economics Letters* 118.1, pp. 225–228.

Kastoryano, Stephen and Bas Van der Klaauw (2011). “Dynamic evaluation of job search assistance”. In: *IZA Discussion Papers, No. 5424*.

Klaauw, Bas van der, Gerard J. van den Berg, and Jan C. van Ours (2004). “Punitive sanctions and the transition rate from welfare to work”. In: *Journal of Labor Economics* 22.1, pp. 211–241.

Klaauw, Bas van der and Jan C. van Ours (Mar. 2013). “Carrot and Stick: How Re-Employment Bonuses and Benefit Sanctions Affect Exit Rates from Welfare”. en. In: *Journal of Applied Econometrics* 28.2, pp. 275–296.

Kyyrä, Tomi (2010). “Partial unemployment insurance benefits and the transition rate to regular work”. en. In: *European Economic Review* 54.7, pp. 911–930.

Kyyrä, Tomi, Pierpaolo Parrotta, and Michael Rosholm (2013). “The effect of receiving supplementary UI benefits on unemployment duration”. en. In: *Labour Economics* 21, pp. 122–133.

Lalive, Rafael, Jan C. van Ours, and Josef Zweimüller (2005). “The effect of benefit sanctions on the duration of unemployment”. In: *Journal of the European Economic Association* 3.6, pp. 1386–1417.

— (2000). “The impact of active labor market programs and benefit entitlement rules on the duration of unemployment”. In: *IZA Discussion Papers, No. 149*.

LaLonde, Robert J. (1986). “Evaluating the econometric evaluations of training programs with experimental data”. In: *The American Economic Review*, pp. 604–620.

Lechner, Michael and Anthony Strittmatter (2017). "Practical procedures to deal with common support problems in matching estimation". en. In: *Econometric Reviews*, pp. 1–15.

Lechner, Michael and Conny Wunsch (2013). "Sensitivity of matching-based program evaluations to the availability of control variables". en. In: *Labour Economics* 21, pp. 111–121.

Lindeboom, Maarten, Ana Llena-Nozal, and Bas van der Klaauw (Dec. 2016). "Health shocks, disability and work". en. In: *Labour Economics* 43, pp. 186–200.

Lindsay, Bruce (1983). "The geometry of Mixture Likelihoods: A general Theory". In: *The Annals of Statistics* 11.1, pp. 86–94.

McVicar, Duncan, Julie Moschion, and Jan C. van Ours (Sept. 2018). "Early illicit drug use and the age of onset of homelessness". en. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Muller, Paul, Bas van der Klaauw, and Arjan Heyma (2017). "Comparing econometric methods to empirically evaluate Job-Search Assistance". In: *Unpublished manuscript*.

Ours, Jan C. van and Jenny Williams (2012). "The effects of cannabis use on physical and mental health". en. In: *Journal of Health Economics* 31.4, pp. 564–577.

— (2009). "Why parents worry: Initiation into cannabis use by youth and their educational attainment". en. In: *Journal of Health Economics* 28.1, pp. 132–142.

Ours, Jan C. van et al. (May 2013). "Cannabis use and suicidal ideation". en. In: *Journal of Health Economics* 32.3, pp. 524–537.

Palali, Ali and Jan C. van Ours (Dec. 2017). "Love Conquers all but Nicotine: Spousal Peer Effects on the Decision to Quit Smoking". en. In: *Health Economics* 26.12, pp. 1710–1727.

Richardson, Katarina and Gerard J. van den Berg (2013). "Duration dependence versus unobserved heterogeneity in treatment effects: Swedish labor market training and the transition rate to employment". en. In: *Journal of Applied Econometrics* 28.2, pp. 325–351.

Ridder, Geert (1986). "The sensitivity of duration models to misspecified unobserved heterogeneity and duration dependence". In: *Unpublished manuscript*.

Røed, Knut and Oddbjørn Raaum (2006). "Do labour market programmes speed up the return to work?" In: *Oxford Bulletin of Economics and Statistics* 68.5, pp. 541–568.

Svarer, Michael (2008). "Crime and partnerships". In: *IZA Discussion Papers, No. 3543*.

Svarer, Michael and Mette Verner (2008). "Do children stabilize relationships in Denmark?" en. In: *Journal of Population Economics* 21.2, pp. 395–417.

Tatsiramos, Konstantinos (2010). "Job displacement and the transitions to re-employment and early retirement for non-employed older workers". en. In: *European Economic Review* 54.4, pp. 517–535.

Tables and Figures

Table 1: Sample statistics and estimates from the selection model using the full sample of actual treated and non-treated

	Treated	Control	Selection model	
	Mean	Mean	Est.	SE
Panel A: Baseline socio-economic characteristics				
Country of origin: Not Europe	0.26	0.19	0.2218***	(0.0470)
Age 25-29	0.23	0.26	0.1371***	(0.0528)
Age 30-34	0.20	0.20	0.1101**	(0.0500)
Age 40-44	0.16	0.15	-0.0255	(0.0525)
Age 45-49	0.12	0.11	-0.1221**	(0.0572)
Age 50-54	0.09	0.09	-0.3005***	(0.0645)
Male	0.67	0.51	0.4441***	(0.0389)
Married	0.35	0.34	-0.0242	(0.0381)
Children: At least one	0.42	0.43	0.0799*	(0.0430)
Children: No. of children in age 0-3	0.20	0.20	0.0762	(0.0503)
Education: Pre-high school	0.18	0.17	-0.0954	(0.1093)
Education: High school	0.57	0.50	0.1101	(0.1073)
Education: University College or higher	0.22	0.31	-0.0097	(0.1081)
Panel B: Inflow time and regional information				
Beginning of unemployment: June-August	0.26	0.30	-0.0107	(0.0360)
Inflow year: 2003-2005	0.29	0.35	-0.4033***	(0.0561)
Inflow year: 2006-2007	0.16	0.18	-0.1888***	(0.0644)
Inflow year: 2008-2009	0.26	0.18	-0.1336**	(0.0623)
Inflow year: 2010-2011	0.17	0.17	-0.2018***	(0.0736)
Region: Stockholm	0.13	0.21	-0.3254***	(0.0688)
Region: Gothenborg	0.14	0.16	-0.3273***	(0.0537)
Region: Skane	0.12	0.14	-0.3010***	(0.0556)
Region: Northern parts	0.20	0.15	0.1506***	(0.0486)
Region: Southern parts	0.14	0.12	0.0318	(0.0539)
Regional unemployment rate (inflow)	10.13	9.54	0.0284***	(0.0092)
Panel C: Previous unemployment duration				
Time unemployed in last spell	116.82	89.05	0.0001	(0.0001)
Missing time unemployed in last spell	0.52	0.51	0.0212	(0.0510)
Panel D: Short–term employment history (2 years)				
Employed 1 year before	0.59	0.59	0.0230	(0.0536)
Employed 2 years before	0.59	0.59	0.0073	(0.0533)
Months employed in last 6 months	3.36	3.54	-0.0032	(0.0166)
Months employed in last 24 months	12.73	13.50	0.0043	(0.0054)
Time since last employment if in last 24 months	2.31	2.42	-0.0068	(0.0061)
No employment in last 24 months	0.21	0.19	-0.1228	(0.0946)
Number of employers in last 24 months	1.64	1.79	0.0068	(0.0155)
Panel E: Short–term unemployment history (2 years)				
Days unemployed in last 6 months	19.34	14.76	0.0009	(0.0007)
Days unemployed in last 24 months	147.55	120.68	0.0002	(0.0002)
No unemployment in last 24 months	0.44	0.44	-0.0626	(0.0635)
Days since last unemployment if in last 24 months	14.76	14.77	0.0000	(0.0002)
Number of unemployment spells in last 24 months	0.81	0.88	0.0002	(0.0254)
Unemployed 6 months before	0.20	0.16	0.0083	(0.0646)

Continue to next page

Table 1 – continued from previous page

	Treated	Control	Selection model	
	Mean	Mean	Est.	SE
Unemployed 24 months before	0.24	0.22	-0.0381	(0.0514)
Any program in last 24 months	0.03	0.02	0.0650	(0.1240)
Panel F: Short–term welfare history (2 years)				
Welfare benefits -1 year	4837.20	3742.33	0.0167	(0.0268)
Welfare benefits -2 years	4208.16	3542.62	0.0030	(0.0335)
On welfare benefits -1 year	0.19	0.14	0.0134	(0.0709)
On welfare benefits -2 years	0.17	0.14	-0.0561	(0.0698)
Panel G: Earnings history (2 years)				
Earnings 1 year before	111493.04	110248.06	0.0197	(0.0357)
Earnings 2 years before	111593.34	110615.42	-0.0088	(0.0431)
Panel H: Long-term employment history (10 years)				
Months employed in last 10 years	58.15	62.93	-0.0016*	(0.0010)
Number of employers in last 10 years	4.69	5.12	0.0111**	(0.0055)
Cumulated earnings 5 years before	532569.32	530474.49	0.0609	(0.0516)
Panel I: Long-term unemployment history (10 years)				
Days unemployed in last 10 years	804.64	692.60	-0.0001*	(0.0000)
No unemployment in last 10 years	0.18	0.17	-0.0867	(0.0662)
Days since last unemployment if in last 10 years	248.79	290.84	-0.0000	(0.0000)
Number of unemployment spells in last 10 years	3.63	3.83	0.0091	(0.0078)
Average unemployment duration	96.91	90.24	-0.0001	(0.0001)
Duration of last unemployment spell	184.42	154.64	-0.0001	(0.0001)
Any program in last 10 years	0.15	0.12	0.0276	(0.0972)
Any program in last 4 years	0.07	0.05	0.0499	(0.1038)
Number of programs in last 10 years	0.19	0.15	0.0350	(0.0671)
Panel J: Long-term welfare history, out-of-labor-force (10 years)				
Yearly average welfare benefits last 4 years	4196.11	3533.19	-0.0190	(0.0498)
Yearly average welfare benefits last 10 years	3928.34	3447.39	-0.0683**	(0.0309)
No welfare benefits last 4 years	0.69	0.75	-0.0697	(0.0640)
No welfare benefits last 10 years	0.51	0.59	-0.0877*	(0.0465)
Panel K: Characteristics of the last job				
Wage	18733.10	18860.87	-0.0615***	(0.0236)
Wage missing	0.54	0.52	-0.0066	(0.1467)
Occupation:				
Manager	0.04	0.07	-0.2952*	(0.1684)
Requires higher education	0.04	0.06	-0.1059	(0.1629)
Clerk	0.05	0.05	0.0377	(0.1620)
Service, care	0.09	0.13	0.0091	(0.1551)
Mechanical, transport	0.13	0.07	0.2189	(0.1531)
Building, manufacturing	0.06	0.05	0.0824	(0.1610)
Elementary occupation	0.05	0.04	0.0207	(0.1625)
Panel L: Characteristics of the last firm				
Firm size	2532.16	3877.20	0.0000	(0.0000)
Age of firm	12.94	14.13	0.0009	(0.0040)
Average wage	21600.38	21517.14	-0.0044	(0.0221)
Wage missing	0.62	0.58	-0.0260	(0.2415)
Mean tenure of employees	3.44	3.69	-0.0040	(0.0102)
Age of employees	27.71	29.45	-0.0042	(0.0038)

Continue to next page

Table 1 – continued from previous page

	Treated	Control	Selection model	
	Mean	Mean	Est.	SE
Share of immigrants	0.13	0.13	-0.1970*	(0.1099)
Share of females	0.26	0.34	-0.4679***	(0.1038)
No previous firm	0.28	0.24	-0.3219	(0.3166)
Most common occupation:				
Manager	0.04	0.06	-0.0881	(0.2530)
Higher education	0.04	0.04	-0.0044	(0.2537)
Clerk	0.03	0.03	0.0769	(0.2574)
Service, care	0.10	0.17	0.0548	(0.2474)
Building, manufacturing	0.04	0.03	-0.0350	(0.2558)
Mechanical, transport	0.11	0.06	0.0559	(0.2473)
Elementary occupation	0.02	0.02	-0.0744	(0.2671)
Industry:				
Agriculture, fishing, mining	0.01	0.02	0.0258	(0.3060)
Manufacturing	0.20	0.11	0.3416	(0.2788)
Construction	0.05	0.07	-0.1136	(0.2849)
Trade, repair	0.07	0.08	-0.0242	(0.2805)
Accommodation	0.03	0.04	-0.1292	(0.2908)
Transport, storage	0.06	0.05	0.2818	(0.2824)
Financial, real estate	0.09	0.10	0.1211	(0.2795)
Human health, social work	0.08	0.14	0.0021	(0.2902)
Other - public sector	0.04	0.08	-0.0900	(0.2909)
Other	0.06	0.07	-0.0089	(0.2825)
Panel M: Unemployment insurance				
UI: Daily benefit level in SEK	388.15	274.86	0.2076***	(0.0506)
UI: Eligible	0.83	0.83	-0.0453	(0.0592)
UI: No benefit claim	0.36	0.54	0.1027	(0.1014)
UI 1 year before	13312.38	13192.71	0.0070	(0.0226)
UI 2 years before	13381.50	13162.67	0.0067	(0.0242)
Cumulated UI 5 years before	65486.46	63664.69	-0.0703**	(0.0306)
Panel N: Duration dependence				
Baseline hazard, part 2			0.2480***	(0.0840)
Baseline hazard, part 3			0.5564***	(0.0727)
Baseline hazard, part 4			0.6643***	(0.0755)
Baseline hazard, part 5			0.6481***	(0.0799)
Baseline hazard, part 6			0.7204***	(0.0741)
Baseline hazard, part 7			0.6542***	(0.0750)
Baseline hazard, part 8			0.2586***	(0.0704)

Notes: Columns 1–2 report sample averages for the full sample with actual treated and non-treated. Columns 3–4 estimates and standard errors from the corresponding selection model. *, ** and *** denote significance at the 10, 5 and 1 percent levels. All earnings and benefits are in SEK and inflation-adjusted.

Table 2: Bias of the effect of training with different sets of covariates

Included covariates	est.	se
Panel A: Baseline		
Baseline socio-economic characteristics	0.0841***	(0.00231)
Calendar time (inflow dummies)	0.1209***	(0.00229)
Region dummies	0.0995***	(0.00230)
Local unemployment rate	0.1193***	(0.00229)
All but socio-economic characteristics	0.1014***	(0.00230)
All the above	0.0749***	(0.00232)
Panel B: Baseline and		
Previous unemployment duration (last spell)	0.0702***	(0.00232)
Employment history (last 2 years)	-0.0145***	(0.00234)
Unemployment history (last 2 years)	0.0666***	(0.00232)
Earnings history (last 2 years)	0.0500***	(0.00233)
Welfare benefit history (last 2 years)	0.0535***	(0.00233)
All of the above	-0.0245***	(0.00234)
Panel C: Baseline, short-term history and		
Employment history (last 10 years)	-0.0262***	(0.00234)
Unmployment history (last 10 years)	-0.0308***	(0.00234)
Welfare benefit history (10 years)	-0.0223***	(0.00234)
All of the above	-0.0284***	(0.00234)
Panel D: Baseline, short-term history, long-term history and		
Last wage	-0.0311***	(0.00234)
Last occupation dummies	-0.0290***	(0.00235)
Firm characteristics (last job)	-0.0239***	(0.00235)
Unemployment benefits	0.0232***	(0.00234)
All of the above	0.0186***	(0.00236)

Notes: Estimated biases using the full sample with placebo treated and placebo non-treated adjusting for different sets of covariates. Hazard rate estimates for time in unemployment using a parametric proportional hazard model with piecewise constant baseline hazard (8 splits). *, ** and *** denote significance at the 10, 5 and 1 percent levels.

Table 3: Bias of the effect of training adjusting for short-term labor market histories

Included covariates	est.	se
Baseline	0.0749***	(0.00232)
Panel A: Short-term employment history (2 years) and baseline		
Earnings history: employed 1 year before	0.0264***	(0.00233)
Earnings history: employed 2 years before	0.0361***	(0.00233)
Months employed in the last 6m	0.0253***	(0.00233)
Months employed in the last 24m	0.0187***	(0.00233)
Time since last employment if in last 24m	0.0733***	(0.00232)
No employment in last 24m	0.0164***	(0.00233)
No. of employers in the last 24m	0.0562***	(0.00233)
All employment history (last 2 years)	-0.0109***	(0.00234)
Panel B: Short-term unemployment history (2 years) and baseline		
Days unemployed in last 6m	0.0767***	(0.00233)
Days unemployed in last 24m	0.0774***	(0.00232)
No unemployment in last 24m	0.0737***	(0.00232)
Days since last unemployment if in last 24m	0.0749***	(0.00232)
No. unemploymnt spells in last 24m	0.0676***	(0.00232)
Unemployed 6m before	0.0770***	(0.00232)
Unemployed 24m before	0.0734***	(0.00232)
Any program in last 24m	0.0751***	(0.00232)
All unemployment history (last 2 years)	0.0666***	(0.00232)

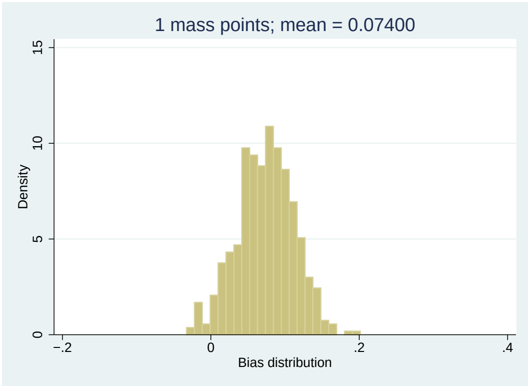
Notes: Estimated biases using the full sample with placebo treated and placebo non-treated adjusting for different sets of covariates. Hazard rate estimates for time in unemployment using a parametric proportional hazard model with piecewise constant baseline hazard (8 splits). The baseline model includes baseline socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

Table 4: Bias and variance of the estimated treatment effect for a pre-specified number of support points and support points according to model selection criteria. By sample size (1)

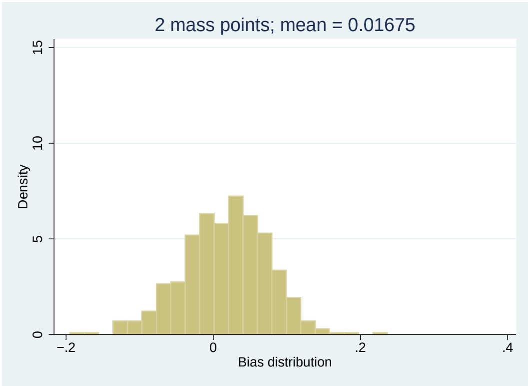
	Sample size								
	10,000			20,000			40,000		
Specification	bias (1)	se (2)	mse (3)	bias (4)	se (5)	mse (6)	bias (7)	se (8)	mse (9)
Panel A: Number of pre-specified support points									
1	0.074	(0.038)	0.0070	0.074	(0.028)	0.0084	0.075	(0.019)	0.0060
2	0.017	(0.058)	0.0036	0.019	(0.040)	0.0020	0.019	(0.029)	0.0012
3	0.029	(0.075)	0.0065	0.026	(0.052)	0.0034	0.022	(0.037)	0.0019
4	0.036	(0.083)	0.0081	0.028	(0.058)	0.0041	0.024	(0.040)	0.0022
5	0.037	(0.084)	0.0083	0.029	(0.057)	0.0041	0.025	(0.040)	0.0022
6	0.038	(0.084)	0.0084	0.029	(0.057)	0.0041	0.025	(0.040)	0.0022
Panel B: Model selection criteria									
ML	0.037	(0.080)	0.0079	0.025	(0.056)	0.0037	0.024	(0.039)	0.0022
AIC	0.017	(0.065)	0.0045	0.017	(0.044)	0.0022	0.021	(0.036)	0.0018
BIC	0.016	(0.056)	0.0034	0.017	(0.038)	0.0018	0.019	(0.030)	0.0012
HQIC	0.016	(0.056)	0.0034	0.017	(0.038)	0.0018	0.020	(0.032)	0.0014
Panel C: Average # support points, by selection criteria									
ML		4.38			4.28			4.28	
AIC		2.22			2.29			2.60	
BIC		2.00			2.00			2.00	
HQIC		2.01			2.01			2.03	

Notes: Estimated bias, variance and mean squared error of the treatment effect from a ToE model with different specifications of the discrete support point distribution. Simulations using 500 replications with random drawings from the full sample with placebo treated and placebo non-treated. Hazard rate estimates for time in unemployment. Each model uses a piecewise constant baseline hazard (8 splits) and the observed covariates include socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate.

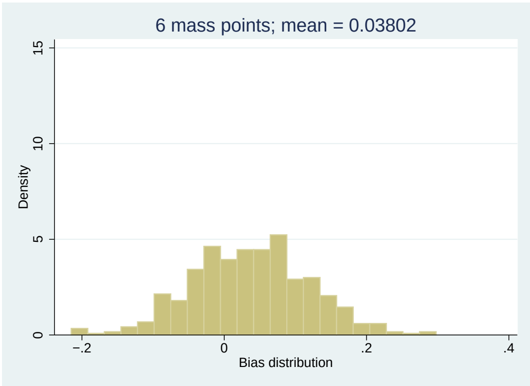
Figure 1: Distribution of the bias of the estimated treatment effect for a pre-specified number of support points. By number of support points



(a) 1 support point



(b) 2 support points



(c) 6 support points

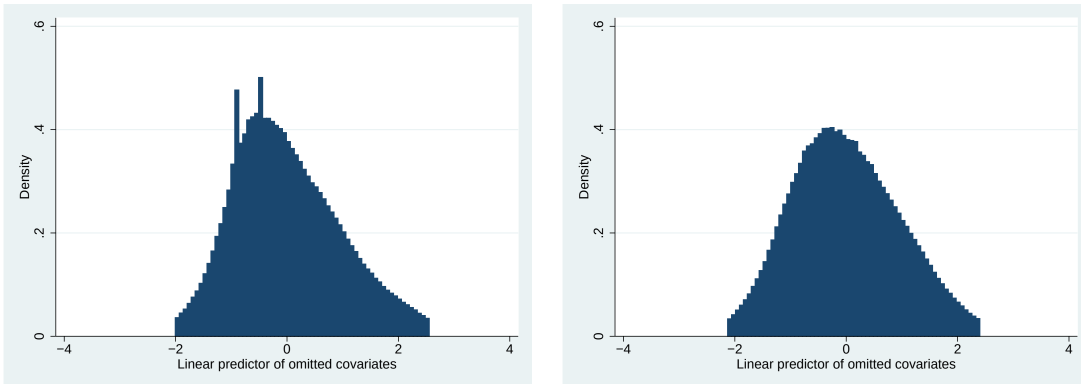
Note: Distribution of the estimated bias of the treatment effect from a ToE model with different specifications of the discrete support point distribution. Simulations using 500 replications with random drawings from the full sample with placebo treated and placebo non-treated. Hazard rate estimates for time in unemployment. Each model uses a piecewise constant baseline hazard (8 splits) and the observed covariates include socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate.

Table 5: Bias and variance of the estimated treatment effect for a pre-specified number of support points and support points according to model selection criteria. By sample size (2)

	Sample size					
	80,000			160,000		
Specification	bias (1)	se (2)	mse (3)	bias (4)	se (5)	mse (6)
Panel A: Number of pre-specified support points						
1	0.074	(0.014)	0.0057	0.074	(0.009)	0.0055
2	0.017	(0.020)	0.0007	0.016	(0.014)	0.0005
3	0.019	(0.024)	0.0009	0.017	(0.015)	0.0005
4	0.021	(0.027)	0.0012	0.019	(0.018)	0.0007
5	0.021	(0.027)	0.0012	0.019	(0.018)	0.0007
6	0.021	(0.026)	0.0011	0.019	(0.018)	0.0007
Panel B: Model selection criteria						
ML	0.021	(0.026)	0.0012	0.019	(0.018)	0.0007
AIC	0.020	(0.024)	0.0010	0.018	(0.018)	0.0007
BIC	0.017	(0.020)	0.0007	0.016	(0.013)	0.0005
HQIC	0.018	(0.021)	0.0007	0.017	(0.015)	0.0005
Panel C: Average # support points, by selection criteria						
ML		4.27			4.22	
AIC		2.90			3.35	
BIC		2.00			2.00	
HQIC		2.11			2.33	

Notes: Estimated bias, variance and mean squared error of the treatment effect from a ToE model with different specifications of the discrete support point distribution. Simulations using 200 replications with random drawings from the full sample with placebo treated and placebo non-treated. Hazard rate estimates for time in unemployment. Each model uses a piecewise constant baseline hazard (8 splits) and the observed covariates include socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate.

Figure 2: Distribution of the true unobserved heterogeneity for the *treatment rate*. Two sources of unobserved heterogeneity

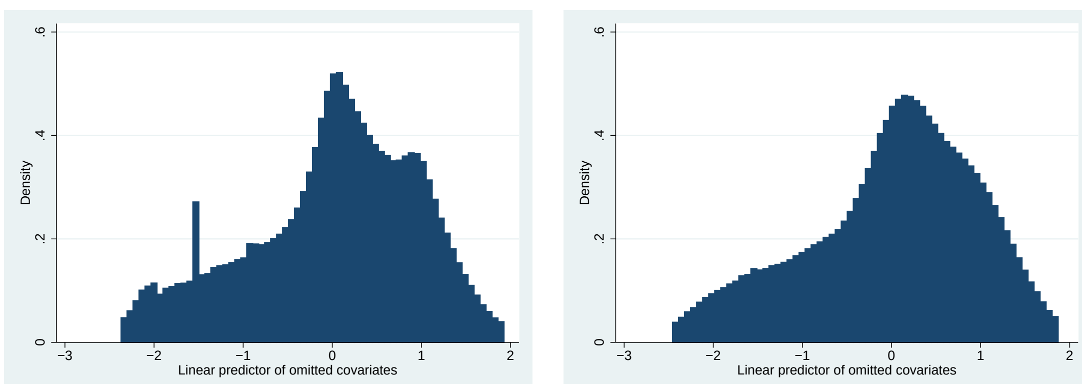


(a) Baseline covariates

(b) Reduced set of covariates

Note: Distributions for the full sample with placebo treated and placebo non-treated. Hazard rate estimates. The baseline model excludes covariates in Panels C-M of Table 1 and the reduced set of covariates in additional excludes baseline socio-economic characteristics.

Figure 3: Distribution of the true unobserved heterogeneity for the *exit rate*. Two sources of unobserved heterogeneity



(a) Baseline covariates

(b) Reduced set of covariates

Note: Distributions for the full sample with placebo treated and placebo non-treated. Hazard rate estimates. The baseline model excludes covariates in Panels C-M of Table 1 and the reduced set of covariates in additional excludes baseline socio-economic characteristics.

Table 6: Bias and variance of the estimated treatment effect when *excluding different sets of covariates*, by model selection criteria and sample size

	Exclude more covariates			Exclude fewer covariates		
Specification	bias (1)	se (2)	mse (3)	bias (4)	se (5)	mse (6)
Panel A: 10,000 observations						
Number of pre-specified support points						
1	0.101	(0.038)	0.0116	0.042	(0.040)	0.0033
2	0.016	(0.064)	0.0044	0.003	(0.052)	0.0027
3	0.023	(0.108)	0.0122	0.009	(0.088)	0.0078
4	0.029	(0.132)	0.0184	0.031	(0.102)	0.0113
5	0.047	(0.139)	0.0216	0.039	(0.101)	0.0117
6	0.054	(0.140)	0.0226	0.040	(0.101)	0.0117
Model selection criteria						
ML	0.054	(0.140)	0.0226	0.040	(0.101)	0.0117
AIC	0.012	(0.095)	0.0092	0.014	(0.097)	0.0096
BIC	0.016	(0.064)	0.0044	-0.002	(0.055)	0.0030
HQIC	0.016	(0.067)	0.0047	-0.002	(0.075)	0.0057
Average # support points, by selection criteria						
ML		4.84			5.17	
AIC		2.44			3.17	
BIC		2.00			2.24	
HQIC		2.01			2.73	

Notes: Estimated bias, variance and mean squared error of the treatment effect from a ToE model with different specifications of the discrete support point distribution. Simulations using 500 replications with random drawings from the full sample with placebo treated and placebo non-treated. Hazard rate estimates for time in unemployment. Each model uses a piecewise constant baseline hazard (8 splits). The baseline model include baseline socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate. The exclude more covariates model excludes baseline socio-economic characteristics and the exclude fewer covariates adds control for short-term earnings history.

Table 7: Bias and variance of the estimated treatment effect with *exogenous variation*, by model selection criteria and sample size

Specification	Time-varying local unemployment rate		
	bias (1)	se (2)	mse (3)
Panel A: 10,000 observations			
Number of pre-specified support points			
1	0.068	(0.037)	0.0061
2	0.004	(0.057)	0.0032
3	0.023	(0.085)	0.0077
4	0.033	(0.096)	0.0103
5	0.039	(0.096)	0.0108
6	0.041	(0.097)	0.0110
Model selection criteria			
ML	0.041	(0.097)	0.0110
AIC	0.015	(0.076)	0.0060
BIC	0.004	(0.057)	0.0032
HQIC	0.004	(0.058)	0.0033
Average # support points, by selection criteria			
ML		4,68	
AIC		2,34	
BIC		2,00	
HQIC		2,01	

Notes: Estimated bias, variance and mean squared error of the treatment effect from a ToE model with different specifications of the discrete support point distribution. Simulations using 500 replications with random drawings from the full sample with placebo treated and placebo non-treated. Hazard rate estimates for time in unemployment. Each model uses a piecewise constant baseline hazard (8 splits). The baseline model include baseline socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate.

Appendix A Additional Tables

Table A.1: Selection model for the hazard into training, time-varying local unemployment rate

	Est. (1)	SE (2)
Baseline hazard, part 2	0.2476	(0.0840)
Baseline hazard, part 3	0.5549	(0.0728)
Baseline hazard, part 4	0.6618	(0.0756)
Baseline hazard, part 5	0.6446	(0.0800)
Baseline hazard, part 6	0.7155	(0.0742)
Baseline hazard, part 7	0.6456	(0.0752)
Baseline hazard, part 8	0.2407	(0.0709)
Local unemployment rate	0.0194	(0.0086)
Country of origin: Not Europe	0.2206	(0.0470)
Age 25-29	0.1380	(0.0528)
Age 30-34	0.1110	(0.0500)
Age 40-44	-0.0245	(0.0526)
Age 45-49	-0.1211	(0.0573)
Age 50-54	-0.2977	(0.0645)
Male	0.4437	(0.0389)
Married	-0.0234	(0.0381)
Children: At least one	0.0794	(0.0430)
Children: No. of children in age 0-3	0.0758	(0.0503)
Education: Pre-high school	-0.0958	(0.1091)
Education: High school	0.1099	(0.1072)
Education: University College or higher	-0.0107	(0.1079)
Beginning of unemployment: June-August	-0.0062	(0.0359)
Inflow year: 2003-2005	-0.3929	(0.0559)
Inflow year: 2006-2007	-0.1803	(0.0643)
Inflow year: 2008-2009	-0.1541	(0.0648)
Inflow year: 2010-2011	-0.1639	(0.0714)
Region: Stockholm	-0.3666	(0.0673)
Region: Gothenborg	-0.3388	(0.0537)
Region: Skane	-0.3129	(0.0554)
Region: Northern parts	0.1664	(0.0482)
Region: Southern parts	0.0166	(0.0536)
Time unemployed in last spell	0.0001	(0.0001)
Missing time unemployed in last spell	0.0230	(0.0510)
UI: Daily benefit level in SEK	0.2105	(0.0506)
UI: Eligible	-0.0462	(0.0592)
UI: No benefit claim	0.1060	(0.1014)
Earnings history: UI 1 year before	0.0061	(0.0226)
Earnings history: UI 2 years before	0.0063	(0.0242)
Earnings history: Cumulated UI 5 years before	-0.0707	(0.0306)
Earnings history: Employed 1 year before	0.0234	(0.0536)
Earnings history: Employed 2 years before	0.0075	(0.0533)
Earnings history: Cumulated earnings 5 years before	0.0612	(0.0515)
Earnings history: Earnings 1 year before	0.0193	(0.0355)
Earnings history: Earnings 2 years before	-0.0091	(0.0431)
Months employed in last 6 months	-0.0036	(0.0166)
Months employed in last 24 months	0.0046	(0.0054)
Time since last employment if in last 24 months	-0.0066	(0.0061)
No employment in last 24 months	-0.1193	(0.0946)
Number of employers in last 24 months	0.0062	(0.0156)

Continue to next page

Table A.1 – continued from previous page

	Est. (1)	SE (2)
Days unemployed in last 6 months	0.0009	(0.0007)
Days unemployed in last 24 months	0.0002	(0.0002)
No unemployment in last 24 months	-0.0608	(0.0635)
Days since last unemployment if in last 24 months	0.0000	(0.0002)
Number of unemployment spells in last 24 months	0.0007	(0.0254)
Unemployed 6 months before	0.0092	(0.0646)
Unemployed 24 months before	-0.0392	(0.0514)
Any program in last 24 months	0.0662	(0.1241)
Welfare benefits -1 year	0.0174	(0.0268)
Welfare benefits -2 years	0.0033	(0.0336)
On welfare benefits -1 year	0.0113	(0.0709)
On welfare benefits -2 years	-0.0551	(0.0698)
Last employment, job: Wage	-0.0625	(0.0236)
Last employment, job: Wage missing	-0.0011	(0.1471)
Last employment, job: Manager	-0.2897	(0.1687)
Last employment, job: Higher education	-0.0990	(0.1632)
Last employment, job: Clerk	0.0431	(0.1623)
Last employment, job: Service, care	0.0144	(0.1554)
Last employment, job: Mechanical, transport	0.2256	(0.1534)
Last employment, job: Building, manufacturing	0.0894	(0.1613)
Last employment, job: Elementary	0.0265	(0.1628)
Months employed in last 10 years	-0.0016	(0.0010)
Number of employers in last 10 years	0.0113	(0.0055)
Days unemployed in last 10 years	-0.0001	(0.0000)
No unemployment in last 10 years	-0.0867	(0.0662)
Days since last unemployment if in last 10 years	-0.0000	(0.0000)
Number of unemployment spells in last 10 years	0.0090	(0.0078)
Average unemployment duration	-0.0001	(0.0001)
Duration of last unemployment spell	-0.0001	(0.0001)
Any program in last 10 years	0.0287	(0.0971)
Any program in last 4 years	0.0495	(0.1039)
Number of programs in last 10 years	0.0355	(0.0670)
Yearly average welfare benefits last 4 years	-0.0205	(0.0498)
Yearly average welfare benefits last 10 years	-0.0679	(0.0309)
No welfare benefits last 4 years	-0.0714	(0.0640)
No welfare benefits last 10 years	-0.0873	(0.0465)
Last employment, firm: firm size	0.0000	(0.0000)
Last employment, firm: age of firm	0.0010	(0.0040)
Last employment, firm: Firm ANST missing	-0.3224	(0.3150)
Last employment, firm average hourly wage	-0.0033	(0.0221)
Last employment, firm average hourly wage, missing	-0.0330	(0.2398)
Last employment, firm: mean tenure of employees	-0.0039	(0.0102)
Last employment, firm: mean age of employees	-0.0041	(0.0038)
Last employment, firm: share of immigrants	-0.1938	(0.1098)
Last employment, firm: share of females	-0.4688	(0.1039)
Last employment, modal occ: Manager	-0.0962	(0.2513)
Last employment, modal occ: Higher education	-0.0145	(0.2521)
Last employment, modal occ: Clerk	0.0714	(0.2557)
Last employment, modal occ: Service, care	0.0478	(0.2457)
Last employment, modal occ: Building, manufacturing	-0.0444	(0.2542)
Last employment, modal occ: Mechanical, transport	0.0465	(0.2455)
Last employment, modal occ: Elementary	-0.0827	(0.2655)
Last employment, industry: Agric, fishing, mining	0.0242	(0.3041)
Last employment, industry: Manufacturing	0.3331	(0.2771)

Continue to next page

Table A.1 – continued from previous page

	Est. (1)	SE (2)
Last employment, industry: Construction	-0.1203	(0.2832)
Last employment, industry: Trade, repair	-0.0321	(0.2787)
Last employment, industry: Accommodation	-0.1372	(0.2891)
Last employment, industry: Transport, storage	0.2756	(0.2806)
Last employment, industry: Financial, real estate	0.1101	(0.2777)
Last employment, industry: Human health, social work	-0.0026	(0.2884)
Last employment, industry: Other - public sector	-0.0970	(0.2892)
Last employment, industry: Other	-0.0101	(0.2807)
Constant	-9.3938	(0.4545)

Notes: Selection model for the hazard of transitioning from unemployment into training, estimated using the full sample of actual treated and controls, and later used to simulate placebo treated and controls. The specification consists of a complementary log-log hazard model, with time measured in days.