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

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

October 27, 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 for unemployed workers with information on participation in a training program. The real data is used to simulate placebo treatment durations using a large set of covariates. Then, we generate correlated unobserved heterogeneity by omitting some of the covariates when we estimate the ToE models. We estimate ToE models with a discrete support point distribution for the unobserved heterogeneity and compare different specifications of the model. One result is that the ToE model performs well, in particular, if time-varying covariates in the form of calendar-time variation is exploited. For the discrete support distribution, we find that both over-correcting for unobserved heterogeneity with to many points and under-correcting with too few points lead to large bias. Another result is that information criteria which penalizes parameter abundance are a very useful way to select the number of support points, but information criteria with little penalty should be avoided because they lead to problems with over-correction.

JEL-codes: C14; C15; C41; J64

[*] We thank Oskar Nordstrm Skans, participants at IFAU and conference participants at EEA and EALE conferences for useful suggestions. Vikström and Lombardi acknowledge support from FORTE.

[†] University of Bristol, IFAU-Uppsala, IZA, ZEW, CEPR and J–PAL.

[‡] Uppsala University, IFAU-Uppsala and Uppsala Center for Labor Studies.

[§] 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. They 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 is that the model allows both the exit rate from the initial state and the treatment rate to be affected by related 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 Abbring et al. (2005) on the effect of benefit sanctions on time in unemployment. Here, 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 Røed et al. (2006) on the effects of active labor market programs, Van Ours and Williams (2009, 2012) on cannabis use, Svarer and Verner (2008) on children and marital stability, Tatsiramos (2010) on job displacements, Svarer (2011) on crime, Jahn and Rosholm (2013) on temporary work, and Baert (2013) on overeducation.[1]

The ToE model is often implemented using a model with discrete support points. The idea is to approximate the unknown bivariate unobserved heterogeneity by means of a discrete distribution (Lindsay, 1983; Heckman and Singer, 1984). In practice, however, this could be implemented 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 an information criterion to select the number of support points. Sample size may also be important as estimation of the non-linear MPH model may be problematic with small samples. Another aspect is different sources of exogenous variation, for instance, in the form of multiple-spell data and time-varying characteristics. In this paper, we use a new simulation design based on actual data to evaluate these and related specifications issues for the implementation of the ToE model in practice.

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 treat-

[1] Moreover, Lalive et al. (2005), Kyyrä (2010) and Kyyrä et al. (2013) study additional aspects of unemployment insurance and re-employment rates. Additional examples include Arni et al. (2013) and Van den Berg and Vikström (2014), who both use the ToE approach study the relationship between benefit sanctions, unemployment durations and job quality. Besides these published studies, there are also many unpublished papers applying the Abbring and van den Berg (2003) approach.

ment effects under unconfoundedness.[2] The key idea is to use actual data for treated units to simulate placebo treatments for non-treated units and base the simulation on these placebo treatments. This ensures that the true effect is zero, the selection model is known and the unconfoundedness assumption holds by construction. The fact that real data is used instead of a data generating process chosen by the researcher makes it arguably more relevant for real applications.

Previous EMC simulations study estimators based on conditional independence assumptions. Here, we propose and implement a variant of the basic EMC approach, which allows us to study the ToE approach. Specifically, we use rich Swedish data for unemployed workers with information on participation in a training program (the treatment). For each worker, we create detailed background information in the same vein as Lechner and Wunsch (2013). A duration model for the time to the treatment using data for both treated and non-treated workers is used to simulate a placebo treatment duration for each non-treated worker. By construction, the effect of these placebo treatments that could start at any point in time is zero and the treatment assignment process is known. With these simulated data we estimate various ToE models, but the key is that we exclude subsets of the observed covariates that were used when generating the placebo treatments. This generates a bivariate duration model with correlated unobserved determinants, since the excluded covariates were used when generating the placebo treatments and the same covariates also affect the outcome duration (re-employment rate). 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 important research question for a very long time. Initial simulation evidence was 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 interesting conclusion is that it is possible to over-correct for unobserved heterogeneity, which leads to biased parameter estimates if the model includes too many support points. This is because an overly dispersed distribution of the unobserved heterogeneity lead to bias in other parts of the model to the model to the data. Gaure et al. (2007) use simulated data and examine a similar

[2] 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.

bivariate duration model as is used in this paper. One finding is that a discrete support points approach is generally very reliable if the sample is large and there is some exogenous variation in the hazard rates. In particular, they highlight that calendar time is a particulary 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.[3]

Our study adds to this evidence by using a simulation design based on actual data. First, by excluding covariates from the model, the estimated treatment effect is severely biased, but already with two support points, a large share of the bias is eliminated. Second, a substantial risk of over-correcting for unobserved heterogeneity is found. 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 increases in the number of support points.

Fourth, 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 and guard against under-correction by rejecting models without or with only weak correction for unobserved heterogeneity. Fourth, information criteria with very little penalty for parameter abundance, such as the ML criteria, should be avoided altogether because it tends to favor models with too many support points, and this leads to problems with over-correction. Fifth, in some specifications the less restrictive AIC criteria outperforms the more restrictive BIC and HQIC criteria, but vice versa occurs in other settings, so that no criteria are superior in all settings.

Sixth, we mainly focus on the specification choices, but the simulation results also indicate that the ToE models indeed are able to adjust for a large share of the bias due to unobserved heterogeneity. This holds in our baseline model, where the only source of "exogenous" variation is variation in the observed covariates. With time-varying covariates in the form calendar-time variation, the bias is further reduced and almost equal to zero. This holds even though model includes substantial heterogeneity since we omit a 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). Eight, our preliminary results suggest that multiple-spell data are not a magic bullet. One reason for this could be that the unobserved factors are not constant across spells, so that estimating a model that imposes this restriction may lead to obscure results.

[3] Another conclusion is that deviations from the model assumptions, such as violation of the MPH assumption, may cause serious problems.

Seventh, as a background to the main analyses, we evaluate the relevance of different set of covariates in a similar way as Lechner and Wunsch (2013). Here, we find that short-term labor market histories are particularly important to adjust for and adjusting for employment history is relatively more important than unemployment, earnings and welfare history (out-of-labor-force). We also find that adding information about long-term labor market history (last 10 years) on top of short-term history (last two years) is unimportant. The latter result for long-term histories are similar to the results in Lechner and Wunsch (2013).

This paper is also related to several other strands of the literature. Two recent papers, Kastoryano and van der Klaauw (2011), and Muller and van der Klaauw (2017) evaluate the ToE approach and other dynamic evaluation approaches by comparing the estimates from these approaches with an experimental benchmark in the spirit of Lalonde (1986). Both papers conclude that a regression discontinuity quasi-experimental approach, conventional matching methods and the ToE approach give quantitatively similar results.

The paper proceeds as follows: in the next section the Timing-of-Events model proposed by Abbring and van den Berg (2003) is described. Section 3 describes the simulation design and the data used in the simulations. Section 4 describes the estimated selection model that is used to simulate the placebo treatments and compares the bias when different sets of covariates are included in the model. In Section 4, we present our main simulation results when the variation in the observed covariates is the only exogenous variation. Section 5 presents estimates for the models with calendar-time variation and multiple-spell data. 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), which 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, $\theta_p(t|x, V_p)$, where t is the elapsed duration, follow the Mixed

Proportional Hazard (MPH) form:[4]

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

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 under certain 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.[5] 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). Abbring and van den Berg (2003) also impose several regularity conditions. We return to these assumptions when we describe our simulation design.

In that sense, identification is semi-parametric, since given the MPH structure, the ToE approach does not rely any other parametric assumptions. Unlike many other methods, 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 exit to from the initial state. If treatment is closely followed by an exit from the initial state, no matter the time before the treatment, 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. But, 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 creates variation in the hazard rates across units. Strictly speaking, this is the only variation that is needed for identification.

But, 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 ad-

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

[5] 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.

ditional exogenous variation (see e.g., 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 unobserved heterogeneity. More specifically, current calendar-time factors has an immediate impact on the exit rate, while 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). Another source of exogenous variation is multiple-spell data. If the unobserved factors are constant across spells for a given unit, Abbring and van den Berg (2003) shows that the random-effects assumption (X and V should be independently distributed) can be relaxed and it is no longer necessary to have exogenous variation through the observed covariates, presumably leading to reliable identification. However, in cases with more complicated selection processes, for instance if the selection process in the current spell depends on outcome realizations in previous spells, this may no longer hold.

In this paper we examine all three types of exogenous variation. Initially, we simulate and estimate models with only exogenous variation through observed covariates. Later we also explore calender-time variation in the form of time-varying local unemployment rate and multiple-spell data.

3 Our 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 as in the typical Monte Carlo study. The argument is that real data is more closely linked to real applications with real outcomes and real covariates, providing 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 unemployed workers in Germany to compare the performance of alternative estimators of treatment effects under conditional independence. They proceed in several steps. Initially, the real data for 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. With the simulated data, Huber et al. (2013) perform various simulations.

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

With this simulated data based on real data we perform various simulations with the aim to evaluate several specification issues when estimating ToE models. This includes several issues raised by previous Monte Carlo simulations (Gaure et al. 2007; Baker and Melino (2000), such as, specification of the distribution of the unobserved heterogeneity and the baseline hazard. We also exclude different blocks of covariates with the aim to study 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 and multiple spell data.

One important reason why we use the Swedish unemployment spell data is that there are many examples of evaluations that estimate ToE models using unemployment spell data.[6] The fact that we use 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

[6] 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).

model to generate the placebo treatment durations measured at the daily level. But, we nevertheless estimate continuous time ToE models, implicitly assuming that the daily spell data is approximately continuous. There are several reasons for this. First, continuous time models is often used in the literature even with daily duration data. Second, Abbring and van den Berg (2003) establish identification results for continuous time data and it is unclear to what extent the identifications results carry over to discrete and interval-censored data. Third, Gaure et al. (2007) estimate both continuous and discrete time models and, as expected, conclude that the discrete time model outperforms the continuous time model if the data truly is discrete. Despite this our continuous time model deliver some optimistic results for the performance of the ToE model, also suggesting that a discrete time version may improve the performance of the model even further.

Since, we do not simulate the exits to job the true effect of the treatment is zero. By construction, the no-anticipation assumption holds, since the units cannot anticipate and react to placebo treatments. However, there are other ToE assumptions that may not hold in our simulation design. First, one ToE assumption is that X and V are independently distributed. This random effects assumption may not hold in our simulations, since the excluded covariates representing unobserved heterogeneity may be correlated with the covariates that we actually use in the ToE estimation. Second, since we do not model the outcome duration, the outcome hazard (re-employment rate) may not follow the MHP structure. Third, we use a very rich set of covariates, but if there are some omitted characteristics, our treatment selection model will be miss-specified since it will be estimated using a hazard model with no embedded unobserved heterogeneity.

All these three potential violations of the ToE assumption arise because we simulate with real data and real data, most likely, do not follow a MPH structure. But, one may argue that this is a benefit of our approach, because we explore estimation of the ToE model using arguably more realistic data.[7]

3.2 Data and the training program

Training program. One often studied treatment for unemployed workers is labor market training, and this is why use data on a Swedish vocational training program called AMU (Arbetsmarknadsutbildning). The program and the analyses data resemble administrative data used in other countries. The main purpose of the program, which

[7] Moreover, as shown in Abbring and van den Berg (2003), the random effects assumption can be relaxed with multiple spell data, which we also explore in our simulations.

typically lasts for around 6 months, is to improve the skills of the unemployed workers and thereby enhance their chances of obtaining 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). Theses paper also describe the program in more detail.[8]

Data sources and sampling. We combine data from several administrative registers and surveys. From the Swedish Public Employment Service we have daily unemployment and labor market program records for all unemployed workers in Sweden. With this data we construct spell data with information 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 2002–2011.[9]. We restrict the analyses to the prime-age population 25–55, because younger workers are subject to different labor market programs and to avoid early retirement decisions. We also exclude unemployed workers with a reported disability. Occasionally, some individuals re-enters unemployment within our observation window, creating multiple-spell data, which we will use in some of our simulations. In total, we have roughly 2.6 million spells, of which 3% involves participation in training, and the mean unemployment duration in the sample is 370 days. In case a job-seeker enters into training multiple times, we consider only the first exit to program.

Covariates and outcome measure. For each spell we construct detailed data on individual-level characteristics (summarized in Table 1). Here, 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 (e.g., firm size, industry and average worker characteristics). Data from the Public Employment Service is used to construct unemployment history. 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

[8] The basic rule is that you must be 25 to participate in the program. During the training, participants receive a training grant. Those who are entitled to unemployment insurance (UI) receive a grant equal to their UI benefits level, and for those not entitled to UI the grant is lower fixed at a certain amount. In all cases, training is free of charge.

[9] Any ongoing spells are right-censored on December 31, 2013

history variables we use both short-run history (last 2 years) and more long-run history (last 10 years). Altogether, this captures all aspect of the workers employment and earnings history in the last 2/10 years. From Unemployment Insurance (UI) records we obtain information on UI eligibility.[10]

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

Since, we focus on duration models we also add some additional variables not used by Lechner and Wunsch (2013). We use information on time in the last unemployment spell and an indicator for 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. We see 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 worker among the treated. All labor market history measures point in the same direction: the training participants have more previous unemployment and welfare dependance in the last 2 and 10 years, respectively.

3.3 Simulation details

Selection model. The first step of our 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:

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

[10] Due to differences between the Swedish data and the German data, there are some differences compared to the Lechner and Wunsch (2013) data: the classification of occupations differ, we lack some firm characteristic and have less information on UI claims and use information on welfare benefits to construct measures of out-of-labor-force. Finally, we do not use information on health, but note that we exclude disabled workers, because they have access to very different labor market programs.

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, respectively. The observed variables, X, include all the variables in Table 1. From the estimates of this model in Table 1 we see that the the daily treatment rate peaks after roughly 300 days. The also confirm the patterns from the sample statistics: immigrants, younger workers, males, high-school graduates, UI recipients are more likely to be treated. Short- and long-term unemployment and employment histories are also important determinants of treatment assignment.

After having estimated the selection model using the full population, 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 simulate the time to treatment, T_p, we perform a transition lottery for each unit, where the realised time to treatment is generated by comparing each units estimated treatment probability with random drawings from uniform distributions.[11] 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. During this procedure, we multiply $\hat{\theta}_p(t|x_i)$ by a constant γ and select γ such that we obtain a share of placebo treated around 20%. This assures that we have a fairly large number of treated even with small samples. A similar approach is taken by Huber et. al (2013).

This describes 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 we create another set of placebo treatments taking calendar-time into account by adding spell-varying local unemployment rate to (1).

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

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

3.4 Implementation of the bivariate duration model

We estimate a continuous-time ToE model with a MPH structure (equation 1) for the treatment and outcome hazards. The unknown distribution of the unobserved heterogeneity is approximated by a discrete distribution with discrete support points (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 take 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 $31, 61, 122, 183, 244, 365, 548, 2160$. Below we explain which covariates, X, we include in the model.

To set up $L_i(v_i)$ we split the spell into parts where all right-hand side variables in (1) are constant. Splits occur for every new spell-duration indicator and when the treatment status changes. Later on we allow for calendar-time variation with additional splits. Spell-part j for unit i is denoted by c_{ij} with 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 an outcome indicators, y_{sij}, which equal one if the spell part ends with a transition to state s and zero otherwise, with two possible states (exit and treatment). Then, with approximately continuous durations, $L_i(v_i)$ is:

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

with

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

By integrating $L_i(v_i)$ over $G(V)$ we obtain L_i. 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. To estimate the discrete support points we use the iterative search algorithm in Gaure et al. (2007).[12] For each replication we estimate models with up to $\overline{W}$ support points and then select the appropriate model using, for instance, a 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 components 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. Return the resulting $(\tilde{v}_W, \tilde{p}_W)$ 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*, we search for a new support point in the $[-3,\ 3]$ interval.[13] In this step all other parameters of the model is fixed, which is why we in *Step 3* perform a ML maximization over all parameters, including the new support point. At the end of the procedure we have $\overline{W}$ ML estimates: $\{\hat{\vartheta}_W,\ \mathcal{L}(\hat{\vartheta}_W)\}_{W=1}^{\overline{W}}$.

Information criteria. We use different approaches to discriminate between the $\overline{W}$ estimates. We report results where we pre-specify the number of support points (1–6 points). Another approach is to increase the number of support points until we see no further improvement of the likelihood. We label this the maximum likelihood (ML) criterion.[14] Moreover, some previous studies suggest that it is useful to use information criteria which penalizes parameter abundance (see e.g., Baker and Melino, 2000). Three criteria used in the literature are the information criteria are the milder Akaike information criterion (AIC) and the two more restrictive criteria Bayesian information criterion (BIC) and Hannan-Quinn information criterion (HQIC).[15] The latter two are

[12] 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).

[13] As starting values we use $v_W = 0.5$ and $p_W = \exp(-4)$. The simulated annealing is stopped once we find a support point with a likelihood improvement of at least 0.01 limiting the number of simulated annealing iterations to 800 in which case we use the best support points at this point. In most cases we find a likelihood improvement within the first 200 iterations.

[14] We restrict to improvements larger than 0.001.

[15] Baker and Melino (2000) propose to use either the BIC or HQIC criteria, while Gaure et al. (2007) also explore the AIC criterion.

more restrictive since they impose a larger penalty on parameter abundance. Formally, $AIC = \mathcal{L}(\hat{\vartheta}_W)$, $BIC = \mathcal{L}(\hat{\vartheta}_W) - 0.5k \cdot \ln N$ 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.[16] The criteria are calculated for each replication so that the selected number of support points may vary both across replications and criteria.

4 The selection model and different sets of covariates

Before proceeding to our simulation results we evaluate the relevance of different set of covariates in a similar way as Lechner and Wunsch (2013), with the aim to assess the relative important of different covariates in evaluations of training program for unemployed workers. 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, creating different types of unobserved heterogeneity depending on included covariates.[17,18] Note that have constructed similar covariates as in Lechner and Wunsch (2013), so that we can examine to what extent their results extend to other countries. However, one difference is that we perform our analyses within a duration outcome framework and, thus, evaluate the bias across specifications using a duration model. For each block of covariates we the full sample of placebo treated and placebo non-treated and estimate parametric proportional hazard (PH) model without adjusting for any unobserved heterogeneity and report the estimated bias from each specification (the estimated treatment effect).[19]

[16] Here, we follow Gaure et al. (2007), but one alternative is to use use N instead of N_c. But, our simulations indicate that this is of minor importance in practice.

[17] Lechner and Wunsch (2013) provide good arguments why conditional independence (CIA) should in their German setting when they use their full set of covariates. This allows us them to study which covariates you need to adjust for. Here, we are less ambitious. 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 we evaluate the relative importance of different types of covariates, without arguing that this in fact is all the covariates that you need.

[18] Before Lechner and Wunsch (2013) several other papers examine the importance of different covariates. Some studies use experimental data to compare with non-experimental estimates with different covariates (see e.g.,) and other studies use novel survey data to assess the importance of usually unobserved variables such as personality traits (see e.g.,).

[19] We have also estimated the bias using other duration models, including a cox regression model, leading to similar results.

Results. Table 2 reports the estimated biases of the effect of training when the model is extended sequentially by adding blocks of covariates.[20] In Panel A we start with a model with a set of baseline socio-economic characteristics, in which case we obtain a positive and sizeable bias of around 8.4% (hazard rate estimates from a PH model). This estimate as well as all other estimates in Table 2 are significant, which is due to very small standard errors as results of the large sample size. Panel A also stepwise includes controls for calendar time and regional conditions (regional dummies and local unemployment rate), but these covariates turn out to be relatively unimportant (bias is reduced from 8.4% to 7.5%). Here, the excluded covariates include, for instance, 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. Initially, we adjust for the duration of the last unemployment spell.[21] The idea is that previous unemployment durations may capture important aspects of unobserved heterogeneity. But, 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, while if we adjust for employment history the bias is close to zero. In fact, the sign of bias is even reversed and becomes 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 select unemployed workers with appropriate occupation history to the vocational training program studied in this paper.

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 two years) has minor impacted on the bias of the estimated treatment effect. The same holds when we in Panel D also adjust for various characteristics of the last job (e.g., previous wage and occupation) as well as detailed information about the last firm (e.g., industry and composition of worker).

Lechner and Wunsch (2013) also find that after controlling for the calendar time,

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

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

regional conditions and short-term labor market history, adding additional covariates, such as long-term labor market history is relatively unimportant and may not be worth investing additional resources to reduce them further. However, one difference compared to their study is that in our setting adjusting for short-term employment history is enough, while Lechner and Wunsch (2013) find that it is important to 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 short-term employment history covariates and add each covariates separately together with the baseline covariates, aiming to to understand which aspects of employment history that are important. The results from this exercise in Panel A of Table 3 show that several employment covariates single handedly capture a large part of the bias. For instance, only adjusting for months employed in the last six months before the unemployment spell reduces the bias from 7.5% to 2.5% and other employment history covariates have equally large impact on the estimated bias. But, 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 once we include all all short-term employment history variables. This suggest that the these covariates capture different aspects of the treatment selection. As comparison, Panel B of Table 3 report estimates from a similar analyses where we control for the short-term unemployment history variables one at a time. Here, we see that all these covariates have modest impact on the estimated bias.

5 Main simulation results

This section presents our main simulation results. We mainly focus on the estimation of the treatment effect, and examine to what extent the ToE model is able to adjust for the bias observed in the previous section and study which specification of the model that leads to the best results. The results are presented in the form of the average bias, the variance of the estimates and the 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 additional sources of exogenous variation in the form of multiple-spell data and calendar-time variation.

Baseline results. Table 4 reports results from our 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). Initially, 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. From the first row we see that the baseline model without any unobserved heterogeneity (1 support point) leads to large bias (7.4%).[22] 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%.[23,24] For three or more support points the average bias is increasing, and the bias goes in the same direction for every support that is added to model. 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 with too many support points is inline the results from Baker and Melino (2000), which argue that it is possible to over-correct for unobserved heterogeneity by including to many spurious support points. Simply, with too many support points one obtains an overly dispersed distribution of unobserved heterogeneity and to fit the data this is compensated by changes (bias) to the the treatment effect and presumably also to the duration dependence. This pattern contradicts the general intuition among applied researchers that you need to adjust for unobserved heterogeneity in the most flexible way 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, respectively. For one support point the estimates are centered around a bias of around 7% and the variance of 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. For six support point we see a further increase in the variance, but perhaps more importantly, it also shifts the entire distribution of the estimates to the right (larger positive bias). This means that the increased bias are not explained by a few a extreme estimates, so that the overly dispersed distribution

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

[23] 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 lead to biased spell-duration and covariate effects (see e.g., Gaure et al., 2007).

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

of unobserved heterogeneity has a more general effect for almost all replications.

Interestingly, the problem with over-correcting for unobserved heterogeneity did not occur to the same extent with the simulated data used by Gaure et al. (2007). They rather highlight that the main problem is under-correction with too few support points.[25] Our simulation results based on real data instead suggest 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 are important.

Information criteria. Next, Panel B of Table 4 provides simulation results when the distribution of the unobserved heterogeneity (number of support points) are specified using various information criteria. We also report the average number of support points that were selected for each criteria (Panel C). The ML criterion which increases the number of support point as long as the likelihood is improved lead to 4.38 support points on average. The bias and variance is large compared to simply pre-specifying a low number of support points, so that the ML criterion tend to select too many support points, leading to the problem with over-correction due to the inclusion of spurious support points. However, the results for AIC, BIC and HQIC are much more encouraging. All three criteria produce models with unobserved heterogeneity, but with rather few support points (for many samples models with only 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 the above problems with over-correction due to too many support points by penalizing the number of parameters in the discrete heterogeneity distribution. They, also guard against under-correction with 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, we see that 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. For variance we see slightly larger differences, with lower variance BIC and HQIC. This is because these more restrictive criteria tend to select fewer support points and we have seen that the variance of the estimated treatment effects is increasing in the number of support points. However, when we below vary the covariates that are excluded from the model (different unobserved

[25] 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 below we will also see that over-correction also is a problem with larger sample sizes.

heterogeneity) we find cases where BIC and HQIC tend to under-correct for unobserved heterogeneity, leading to larger bias than for the less restrictive AIC criteria.

Note, that we are mainly interested in providing background information on the various specification choices, however, Table 4 also provides some insights on the overall idea of using ToE models to adjust for unobserved heterogeneity. Here, we conclude 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 and multiple-spell data can eliminate the bias entirely.

Sample size. In Columns 4–9 of Table 4 we increase the sample size 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 point increases the bias from 1.7% to 3.8% and with 40,000 observations it goes from 1.9% to 2.5%. This pattern is confirmed by the simulations results in Table 5, where we increase the sample size 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 suggest that over-correction with too many support points mainly is a problem with small samples. However, note that what constitute a small sample size, most likely, will differ 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 results is that for larger samples size there are smaller differences between the ML criterion and the three other information criteria. For instance, with a sample size of 160,000 we see virtually no difference in the average bias between the four information criteria.

Excluded covariates. We next vary the unobserved heterogeneity. We generate unobserved heterogeneity by excluding sets of the covariates when estimating the ToE models. Thus, a direct way to vary the source of unobserved heterogeneity is to vary the observed covariates that we include in the ToE model. In the baseline simulations we include baseline socioeconomic characteristics, inflow time dummies and regional indicators in the ToE model. Here, we generate more unobserved heterogeneity by excluding additional covariates (exclude some socioeconomic characteristics) and less heterogeneity by excluding fewer covariates (also include previous earnings). Here, we use 10,000 observations. In Table 1 these models generate a bias of 10% and 5.0%, respectively, in the full placebo treatment sample of (see Panel and B). This 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, using that for the treatment hazard the true distribution of the unobserved heterogeneity is known in our EMC framework. A comparison with the baseline model shows that the model with extended unobserved heterogeneity leads to substantially more dispersion of the unobserved heterogeneity for the treatment rate (Figure 2), but we also see increased variance of for the exit rate unobserved heterogeneity (Figure 2).

Initially, Panel A of Table 6 shows that for the model with more extensive unobserved heterogeneity the average bias without unobserved heterogeneity is 10.1% but this drops to 1.6% with two support points. Again, showing that our 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, leading 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 other criteria perform much better. In this case with more extensive unobserved heterogeneity all three criteria on average select a larger number of support points (see Panel C) than for the baseline model with less unobserved heterogeneity. All three criteria also gives a lower bias than for 2 support points (the number with the lowest bias). For instance, the average bias with two support points is 1.6% while for AIC criterion the bias is 1.0%. This is because the AIC criterion on average favours 2.46 support points, i.e. selects models with 2 or 3 support points depending on the data in each replication.

In these simulations we also see 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 bias 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, leading to larger bias than for the less restrictive AIC criterion. However, for the specification where we exclude fewer covariates (Columns 4–7, Table 6) the pattern is different. Here, we see that 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 to the other, suggesting that it may be informative to report several effect estimates if the three information criteria lead to different number of support points.

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 we have seen that the variance is increasing in the number of support points. The lower variance also means that mse is lower for BIC and HQIC (Column 3). It 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 that criterions with very little penalty for parameter abundance, such as the ML criteria, should be avoided altogether, since it tends to favor models with too many support points leading to problems with over-correction for unobserved heterogeneity. Another findings is that neither AIC, BIC nor HQIC is superior in all settings. All three penalize parameter abundance and this protects against problems with over-correction due to spurious support points. But, in some cases the risk of under-correcting is relatively more important and this favours the less restrictive AIC criterion, but in other cases the risk of over-correcting is more important and this the more restrictive BIC and HQIC criteria. Thus, using all three and possibly report several effect estimates if the suggested number of support points differ seems like a reasonable approach.

Finally, let us compare with the simulation results in Gaure et al. (2007). They conclude that the most reliable criteria are ML and AIC because these criteria only impose no and weak penalty for parameter abundance, respectively. Another conclusions is that with small samples, a stronger penalty may be required (BIC, HQIC). To some extent our simulation results confirm their results, but one important differences is that in our simulations based on real data we find a substantial risk of over-correcting for unobserved heterogeneity. This leads to the conclusion that criteria with virtually no penalty for parameter abundance, such as the ML criteria, lead to substantial bias.

5.1 Exogenous variation

Identification of the ToE model requires variation in the observed exogenous covariates, since this produces ”exogenous” variation in the hazard rates. Essentially, this was the only source of exogenous variation that we exploited in the baseline simulations above. It resulted in several insights on how to specify the unobserved heterogeneity when estimating ToE models. We also saw that the ToE model were 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. Specifically, we exploit variation in the form calendar-time variation (time-varying local unemployment rate) and multiple-spell data.

Calender-time variation. Calendar-time variation should be useful for identification since it generates a shift in the hazard rates which helps ro recover the distribution of the unobserved heterogeneity. To this end, we first re-estimate our treatment selection model for the actual treated and actual non-treated also including time-varying local unemployment rate among the covariates. The estimates from this extended selection model are all very similar to the baseline selection model. As before, we then simulate placebo treated and non-treated and estimate ToE models but now also including the time-varying local unemployment rate in the set of covariates.

The results from this exercise are presented in Columns 1–3 of Table 7. From the results in the first row of Panel A we see that the bias without adjusting for unobserved heterogeneity (1 support point) is 6.9%. This is slightly smaller than the bias we obtained in Table 4 for the baseline simulation model. This is because we now also include time-varying local unemployment rate in both the selection model and the ToE models, and this leads to a slightly different selection process. As before, we then stepwise include additional support points (Panel A). These results to a large extent confirm the results from our baseline simulation model. First, if we under-correct for unobserved heterogeneity (no unobserved heterogeneity) this leads to sizeable bias and if we over-correct for unobserved heterogeneity the bias of the estimated treatment effect also is large (increasing bias with more than three support points) . Second, if we use the ML criterion we tend to select models with overly dispersed unobserved heterogeneity distribution, leading to substantial bias. Third, the three criteria with penalty for parameter abundance (AIC, BIC and HQIC) all perform well, as they lead to models for which the bias is low.

However, one important difference compared to our baseline simulations is that the average bias for the three information criteria now is substantially lower than in the baseline simulations, 0.4–1.2% compared to 1.6–1.7% in the baseline analyses. This confirms that exploiting time-varying covariates greatly helps the identification as it pushes the bias further towards zero and almost eliminate the bias completely. Note that this result holds even though we have generated quite 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, and this produced substantial bias in the model without unobserved heterogeneity. All this echo the results from Gaure et al. (2007), which conclude that a particularly useful source of identification is the existence of a common calendar-time factor in this exogenous variation, in which case they obtain very reliable estimates of both the treatment effect and the spell-duration effects. The only difference is that they use time-varying calendar-time dummies while we exploit time-varying local unemployment rate.

Multiple-spell data. The intuition why multiple-spell data should be important is rather straightforward. If the unobserved factors are constant across spells we can exploit variation across spells to adjust for unobserved heterogeneity. To be able to exploit multiple spells we sample workers instead of spells and retain all spells for the selected workers. We sample the number of workers such that we on average obtain roughly 10,000 spells in each replication, although the exact number of spells will vary across replications. The simulation results are given in Columns 4–6 of Table 7. Initially, note that for the model with one support point do not exploit that we have multiple spells for the same worker. This explains why the average bias is of roughly the same size as in the baseline simulations in Table 4.[26]

The results for multiple spell show that adjusting for multiple spell information has rather substantial impact on our estimates. Already with two support points the average bias moves towards zero and even changes sign, now indicating a negative bias. Our interpretation is that on average the training participants are a positively selected group of workers, most notably, they on average has more employment in the last years before the unemployment spells, possible because caseworkers wish to enroll workers who really good benefit from the training. However, these more skilled workers are selected for training during unemployment spells where they struggle more to find employment relative to during other unemployment spells, and this could explain why the bias is negative when do adjust for the multiple-spell dimension. We also see that the bias increases for each support point if go beyond two support points, again pointing to problems with over-correction for unobserved heterogeneity.

However, perhaps more interestingly, we see that all four information criteria perform rather poorly with multiple-spell data. All four suggest models with too many support points. This even holds for the most restrictive criteria (BIC and HQIC). One reason for this could be that the unobserved factors are not constant across spells,

[26] Note that bias is slightly higher with multiple spell data sets (8.3%) than for the baseline data (7.4%). The explanation is that we here sample workers instead of spells, and this means that we give relatively more weight to workers with many spells, because if they are selected all of their spells are included. If the selection process differ for workers with many and few spells this can explain the difference in the average bias between the two samples.

so that estimating a model that imposes this restriction may lead to obscure results. This could, for instance, be the case if the selection process in the current spell depends on outcome realizations in previous spells. Thus, multiple-spell day may not be the magic bullet that allows the researcher to make a easy adjustment for unobserved heterogeneity.

6 Conclusions

In this paper we modify a recently proposed simulation technique, the Empirical Monte Carlo approach, to evaluate the Timing-of-Events model. This method allows 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. We use our simulation protocol to examine key specification issues for the estimation of ToE models. A key issue is the specification of the discrete support points distribution for the unobserved heterogeneity. Our main conclusion is that information criteria are 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, but 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 against both over-correction for unobserved heterogeneity due spurious support points and under-correction due to insufficient correction for unobserved heterogeneity. However, none of these three criteria are superior in all setting, suggesting that that robustness analyses involving all three seems like a reasonable strategy in empirical ToE analyses.

Another key conclusion is that ”exogenous” variation in the form calendar-time variation is a very useful sources of identification, but ToE models only relying on variation in the observed covariates also tend to produce good results as long as an appropriate information criterion is used. This echo the results from Gaure et al. (2007). As a background to our analysis we have also evaluated the relevance of different set of covariates. Here, the main conclusion is that it is important to adjust for short-term labor market histories when evaluating labor market program for unemployed workers, while adding long-term labor market histories are unimportant. This is inline with the results in Lechner and Wunsch (2013).

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)

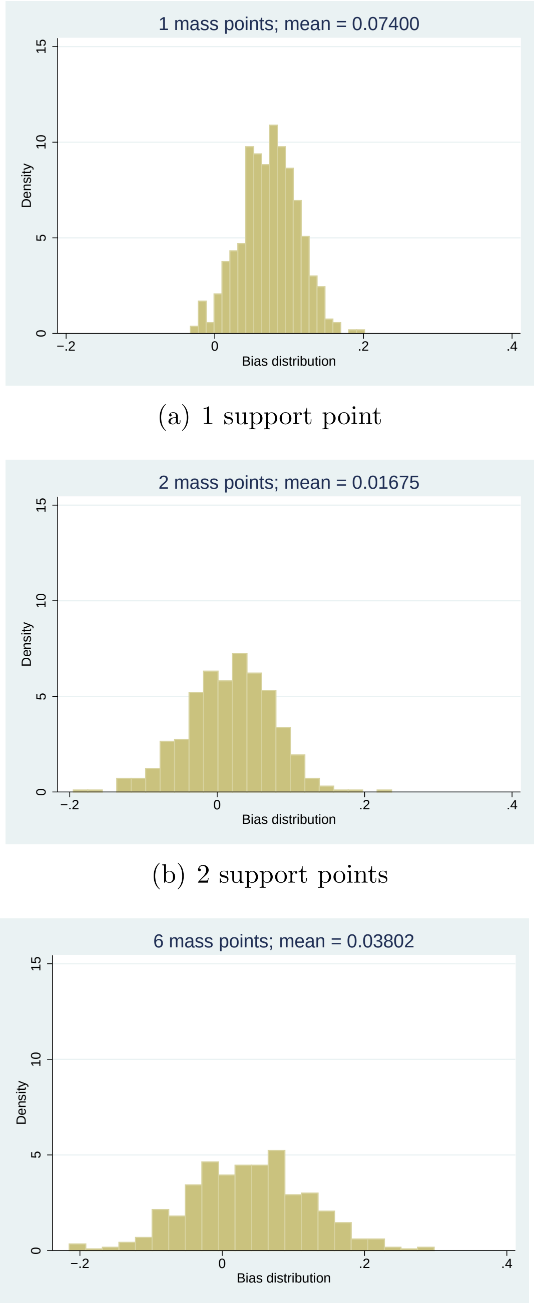
Notes: Estimated biases using the full sample with placebo treated and placebo non-treated adjusting for different sets of covariates. Hazard rate estimates for time in unemployment using a parametric proportional hazard model with piecewise constant baseline hazard (8 splits). The baseline model includes 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

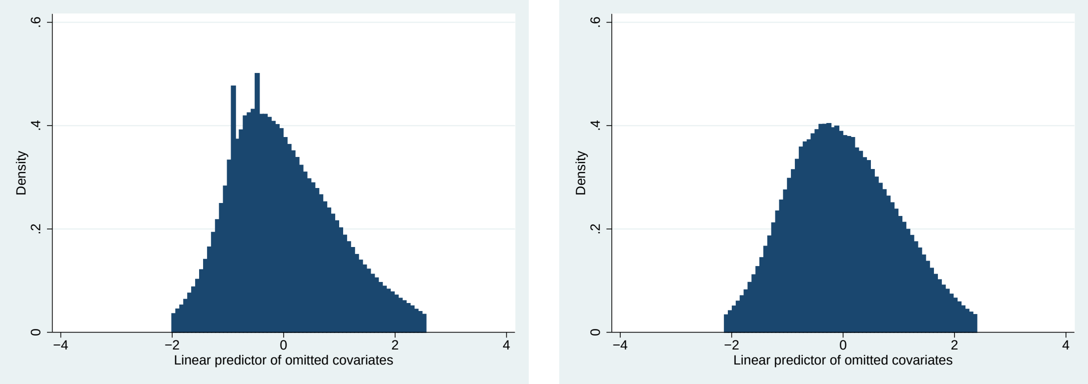
Note: Distribution of the estimated bias of the treatment effect from a ToE model with different specifications of the discrete support point distribution. Simulations using 500 replications with random drawings from the full sample with placebo treated and placebo non-treated. Hazard rate estimates 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 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 2: Distribution of the true unobserved heterogeneity for the *treatment rate*. Two sources of unobserved heterogeneity

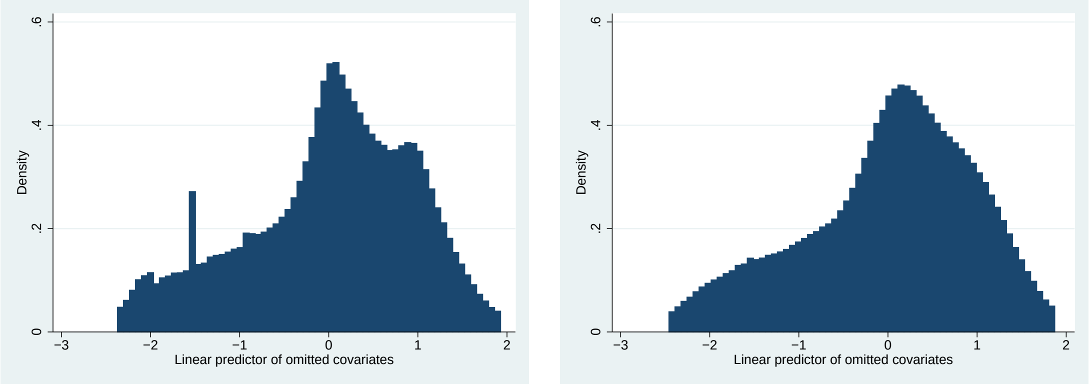


(a) Baseline covariates

(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

	Time-varying local unemployment rate			Multiple spells		
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	0.072	(0.036)	0.0065	0.083	(0.014)	0.0072
2	0.007	(0.055)	0.0031	-0.016	(0.016)	0.0005
3	0.025	(0.087)	0.0081	-0.034	(0.016)	0.0014
4	0.035	(0.096)	0.0105	-0.037	(0.016)	0.0016
5	0.042	(0.099)	0.0116	-0.037	(0.017)	0.0016
6	0.043	(0.100)	0.0119	-0.036	(0.017)	0.0015
Model selection criteria						
ML	0.043	(0.101)	0.0120	-0.036	(0.017)	0.0015
AIC	0.017	(0.074)	0.0058	-0.036	(0.016)	0.0016
BIC	0.007	(0.055)	0.0031	-0.037	(0.016)	0.0016
HQIC	0.008	(0.057)	0.0034	-0.037	(0.017)	0.0016
Average # support points, by selection criteria						
ML		4.70			6.00	
AIC		2.36			5.70	
BIC		2.00			4.59	
HQIC		2.02			5.22	

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). The baseline model include baseline socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate.