Lombardi, Stefano; Berg, van den Berg, Gerard J.; Vikström, Johan

Working Paper

Empirical Monte Carlo evidence on estimation of Timing-of-Events models

Working Paper, No. 2020:26

Provided in Cooperation with:
IFAU - Institute for Evaluation of Labour Market and Education Policy, Uppsala

Suggested Citation: Lombardi, Stefano; Berg, van den Berg, Gerard J.; Vikström, Johan (2020) : Empirical Monte Carlo evidence on estimation of Timing-of-Events models, Working Paper, No. 2020:26, Institute for Evaluation of Labour Market and Education Policy (IFAU), Uppsala

This Version is available at:
https://hdl.handle.net/10419/246035

WORKING PAPER 2020:26

Empirical Monte Carlo evidence on estimation of Timing-of-Events models

Stefano Lombardi
Gerard J. van den Berg
Johan Vikström

IFAU
INSTITUTE FOR EVALUATION OF LABOUR MARKET AND EDUCATION POLICY

The Institute for Evaluation of Labour Market and Education Policy (IFAU) is a research institute under the Swedish Ministry of Employment, situated in Uppsala.

IFAU's objective is to promote, support and carry out scientific evaluations. The assignment includes: the effects of labour market and educational policies, studies of the functioning of the labour market and the labour market effects of social insurance policies. IFAU shall also disseminate its results so that they become accessible to different interested parties in Sweden and abroad.

Papers published in the Working Paper Series should, according to the IFAU policy, have been discussed at seminars held at IFAU and at least one other academic forum, and have been read by one external and one internal referee. They need not, however, have undergone the standard scrutiny for publication in a scientific journal. The purpose of the Working Paper Series is to provide a factual basis for public policy and the public policy discussion.

More information about IFAU and the institute's publications can be found on the website www.ifau.se

ISSN 1651-1166

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

Stefano Lombardi[b], Gerard J. van den Berg[c] and Johan Vikström[d]

29th December, 2020

Abstract

This paper builds on the Empirical Monte Carlo simulation approach developed by Huber et al. (2013) to study the estimation of Timing-of-Events (ToE) models. We exploit rich Swedish data of unemployed job-seekers with information on participation in a training program to simulate placebo treatment durations. We first use these simulations to examine which covariates are key confounders to be included in selection models. The joint inclusion of specific short-term employment history indicators (notably, the share of time spent in employment), together with baseline socio-economic characteristics, regional and inflow timing information, is important to deal with selection bias. Next, we omit subsets of explanatory variables and estimate ToE models with discrete distributions for the ensuing systematic unobserved heterogeneity. In many cases the ToE approach provides accurate effect estimates, especially if time-varying variation in the unemployment rate of the local labor market is taken into account. However, assuming too many or too few support points for unobserved heterogeneity may lead to large biases. Information criteria, in particular those penalizing parameter abundance, are useful to select the number of support points.

Keywords: duration analysis, unemployment, propensity score, matching, training, employment
JEL-codes: C14, C15, C41, J64

[a]We thank Paul Muller, Oskar Nordström Skans, Helena Holmlund, seminar participants at the University of Bonn and IFAU and conference participants at the EEA and EALE for useful suggestions. Estimations were performed on supercomputing resources provided by the Swedish National Infrastructure for Computing (SNIC) at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). Funding from FORTE is gratefully acknowledged.
[b]VATT Institute for Economic Research Helsinki, IZA and UCLS at Uppsala University.
E-mail: stefano.lombardi@vatt.fi.
[c]University of Groningen, University Medical Center Groningen, IFAU, IZA, ZEW, CEPR and J–PAL. E-mail: gerard.van.den.berg@rug.nl.
[d]IFAU and Uppsala University.
E-mail: johan.vikstrom@ifau.uu.se.

Table of contents

1 Introduction

The Timing-of-Events (ToE) approach focuses on the effect of a treatment that may be given during a spell in a state of interest on the rate of leaving that state, when systematic unobserved confounders cannot be ruled out. Abbring and van den Berg (2003) specify a bivariate Mixed Proportional Hazard (MPH) model and establish conditions under which all parts of the model, including the treatment effect, are non-parametrically identified. The fact that this approach allows for unobserved confounders is one reason for which it has been applied in many settings.[1]

Several factors must be taken into account when using these models for empirical inference. In particular, in the literature, the unknown bivariate unobserved heterogeneity distribution is often approximated by way of a discrete distribution (Lindsay, 1983; Heckman and Singer, 1984), and in empirical settings this can be implemented in several ways. One is to pre-specify a (relatively low) number of support points and increase this number until the numerical estimation routine indicates that support points converge or their associated probabilities vanish, or until computational problems arise. Alternatively, one could use an information criterion to select the number of support points. Moreover, sample size may be a relevant factor, since estimation of (non-linear) MPH models with many parameters may be problematic with small samples. Time-varying covariates may make results less dependent on functional-form assumptions (van den Berg, 2001).

In this paper, we use a new simulation design based on actual data to evaluate these and related specification issues for the implementation of the ToE model in practice. To this end, we adapt the Empirical Monte Carlo design (EMC) proposed by Huber et al. (2013) and developed to compare different methods for estimating treatment effects under unconfoundedness.[2] The key idea is to use actual data on

[1] An early example is Abbring et al. (2005) who study the effect of benefit sanctions on the re-employment rate, with unobserved factors such as personal motivation potentially affecting both the time to a benefit sanction (treatment) and time in unemployment (outcome). Recent examples include Crépon et al. (2018), Richardson and van den Berg (2013), Caliendo et al. (2016), Busk (2016), Lindeboom et al. (2016), Holm et al. (2017), Bergemann et al. (2017) on labor market policies; van Ours and Williams (2009, 2012), McVicar et al. (2018) 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 et al. (2013) on overeducation.

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

treated units to simulate placebo treatments and then base the simulations on these placebo treatments. This ensures that the true 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 may make the simulations more relevant for real applications.

Previous EMC implementations have examined estimators based on conditional independence assumptions. The present paper proposes a variant of the original EMC approach, which enables us to study the estimation of the ToE model. In our simulation design, we take advantage of rich administrative data on Swedish job-seekers, with precise information on participation in a training program (the treatment). We use this detailed information on actual treated and non-treated units to estimate a descriptive duration model for the duration until treatment under the assumption that all systematic determinants of the treatment assignment are captured by the full set of observed covariates. Next, we simulate placebo treatment dates for each non-treated unit using the estimated model. By construction, the effect of these placebo treatments is zero and the treatment assignment process is known. With the simulated data we then estimate various ToE models, but leave out subsets of the variables used to simulate the placebo treatment dates. Since the excluded variables were used to generate the placebo treatments, and since in general they also affect the outcome duration (via the re-employment rate), we obtain a bivariate duration model with correlated unobserved determinants, i.e. the ToE setting. This new simulation design allows us to use real-life data to examine a number of model specification issues.

A first long-lasting question related to the specification of ToE models is how to best specify the distribution of unobserved heterogeneity. Initial simulation evidence for MPH models was provided by Heckman and Singer (1984), Ridder (1987), and Huh and Sickles (1994). More recently, Baker and Melino (2000) studied discrete duration models with unobserved heterogeneity and duration dependence. One of their conclusions is that model specifications that allow for too many support points over-correct for unobserved heterogeneity (through an overdispersed unobserved heterogeneity distribution), which leads to bias in all model components. Gaure et al. (2007) use simulated data to examine a bivariate duration model similar to the one analyzed in this paper. They find that a discrete support-points approach is generally reliable if the sample is large and there are time-varying covariates. Pre-specifying a low number of support points for unobserved heterogeneity, or deviations from the model assumptions, may cause substantial bias.

Our study adds to this evidence by using a simulation design based on real data, rather than on artificial simulations. This leads to several conclusions. If we leave out a large number of variables from the model without controlling for unobserved heterogeneity, the estimated effect of the placebo treatment is far from the true zero effect, i.e. we generate substantial bias. However, two support points are already able to eliminate a large share of the bias. We also find a risk of over-correcting for unobserved heterogeneity. With too many 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. The over-correction problem occurs because the estimated unobserved heterogeneity distribution is overdispersed, and to fit the data, the model compensates by generating biases in the treatment effect and duration dependence.

Our simulation results further show that information criteria are useful for selecting 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 without or with only a restricted correction for unobserved heterogeneity. On the other hand, information criteria with little penalty for parameter abundance, such as those solely based on the maximum likelihood (ML criterion), should be avoided. This is because they tend to favor models with too many support points, which leads to over-correction problems.

We mainly focus on the above-mentioned specification choices, but simulation results also indicate that the ToE model is generally able to adjust for a significant share of bias due to unobserved heterogeneity. Remarkably, this already holds in our baseline model in which the only source of variation is across cross-sectional units through time-fixed covariates. When we introduce more variation in the form of time-varying covariates (notably, the unemployment rate in the local labor market measured at monthly intervals), the bias is further reduced. The importance of time-varying covariates echoes the results in Gaure et al. (2007).

The results on how to specify the distribution of unobserved heterogeneity are not only relevant for ToE models but also for all kinds of selection models with random effects, including univariate duration models, general competing risks models, non-parametric maximum likelihood estimators for non-duration outcomes and

structural models with unobserved heterogeneity.[3]

As an additional contribution of the present paper, we address the relevance of different sets of covariates when measuring causal effects of active labor market programs. This is important for evaluations based on conditional independence (CIA) assumptions but is also important for identification strategies that allow for unobserved heterogeneity, as it helps to characterize the unobserved heterogeneity that needs to be taken into account. This contribution of our paper builds on a sizeable literature. Part of this uses experimental data to examine the relevance of different sets of covariates and the implications for the performance of non-experimental methods (Dehejia and Wahba, 1999, 2002; Smith and Todd, 2005).[4] Another part uses rich survey data to assess the importance of characteristics that are often not recorded in administrative data.[5]

In a related study, Lechner and Wunsch (2013) use data from Germany to examine the relevance of different covariates. Their starting point is to incorporate essentially all variables that are important for the selection process and have been used in various CIA-based evaluations of active labor market programs. This gives a flexible selection model that is used to simulate placebo treatments for the non-treated. Then, to assess the relative importance of different variables, they leave out alternative blocks of covariates and compare the size of the bias across specifications. We use our Swedish data in a similar way. Initially, we construct variables analogous to those in the German setting of Lechner and Wunsch (2013). This allows us to examine to what extent the results in Lechner and Wunsch (2013) carry over to other countries and programs. However, we also include additional covariates. First,

[3] Univariate duration models with unobserved heterogeneity have been used to study factors behind duration dependence in aggregate re-employment rates. The latter may be explained by individual-level duration dependence or dynamic sorting of unemployed with low exit probabilities into long-term unemployment (e.g., Abbring et al., 2001). In labor economics, competing risks models are used in studies of unemployment durations with competing exits to employment and non-employment (e.g., Narendranathan and Stewart, 1993) as well as exits to different types of jobs (Baert et al., 2013; Jahn and Rosholm, 2013). In health economics and epidemiology, two often studied competing risks are disease relapse and death (e.g., Gooley et al., 1999). Non-parametric maximum likelihood estimators have also been extensively used when modelling non-duration outcomes, for instance in consumer choice analysis (Briesch et al., 2010) and univariate or multinomial choice models with unobserved determinants (Ichimura and Thompson, 1998; Fox et al., 2012; Gautier and Kitamura, 2013).

[4] Heckman et al. (1998), Heckman and Smith (1999) and Dolton and Smith (2010) find that it is important to control for regional information and labor market history in a flexible way. Mueser et al. (2007) highlight the importance of socio-demographic characteristics and pre-treatment outcomes.

[5] For example, Caliendo et al. (2017) study the relevance of measures of personality traits, attitudes, expectations, social networks and intergenerational information. They find that such factors are indeed relevant elements in selection models, but they tend to become unimportant if the available information in the administrative data is sufficiently rich.

since we model treatment durations and not binary treatment indicators, we also include previous employment and unemployment durations in the set of covariates. This is because previous durations capture aspects related to how long one stays unemployed in the current spell in a more natural way than non-duration history variables. Second, to capture more general skills, we use information on parental income, which is a commonly used proxy for general unobserved skills. Third, time-varying covariates, such as local business cycle conditions, may play a role, especially for longer unemployment spells.

We find that short-term labor market history variables are particularly important to adjust for. Moreover, adjusting for employment history is relatively more important than adjusting for unemployment, earnings and welfare history (out-of-labor-force). We also find that adding information about long-term labor market history (last ten years) on top of controlling for short-term history (last two years) is unimportant. When comparing different short-term employment characteristics, we see that the short-term employment history (in particular, the employment rate) is important to control for, whereas the short-term unemployment history is relatively less important.

Taken together, the insights on how to best specify the ToE model under different types of unobserved heterogeneity and the study of the relevance of the different covariates included in the selection model offer practical guidance on how to choose among alternative identification strategies. When rich enough information is available to the researcher, CIA-based methods can be deemed appropriate. On the other hand, when less rich information is available, the ToE is able to approximate well different types of (substantial and complex) unobserved heterogeneity, especially in settings with time-varying covariates.

Finally, it is useful to discuss our approach in the light of a recent article by Advani et al. (2019) which points out some limitations of the original EMC approaches that were developed to compare different estimation methods for evaluation under unconfoundedness. Notably, it shows that rather modest misspecifications of a model may lead to incorrect EMC inference on what constitutes the best estimation approach for that model in a given empirical setting. Depending on the range of misspecification that is considered, this is potentially relevant for our study. Therefore, throughout the paper, we maintain the assumption that the ToE model is correct. In particular, we do not allow for deviations of the proportionality assumptions in the MPH specifications. This is in line with the vast empirical literature based on the ToE approach in the past decades (see, e.g., the references above). However, we

acknowledge that it is an interesting topic for future research to examine this issue more closely. The critique may also affect more specific assumptions of the empirical models that we estimate. For instance, in the presence of heterogeneous effects, a basic ToE model with a homogeneous effect is misspecified. Another finding in Advani et al. (2019) is that in modest sample sizes such as sizes below 8,000, bootstrap procedures often provide the most appealing approach to select the best estimator. However, our samples are substantially larger (from data containing 2.6 million unemployment spells) and our likelihood-based inference requires Swedish national supercomputing resources, so in our view the application of bootstrap procedures would be beyond the scope of our paper.

The paper proceeds as follows. Section 2 presents 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, and Section 4 describes the estimated selection model that is used to simulate the placebo treatments, and we compare the bias when different sets of covariates are included in the model. In Section 5, we present the EMC simulation results, and Section 6 concludes.

2 The Timing-of-Events model

This section presents the ToE approach as introduced by 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 realizations. The model includes 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:[6]

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

where t is the elapsed duration, $D(t)$ 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 in the exit duration and the treatment

[6] This is the most basic ToE model with time-constant and homogeneous treatment effect, 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.

duration, respectively. Also, let $G(V)$ 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 the following assumptions. The first assumption is no-anticipation, which means that future treatments are not allowed to 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.[7] 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 assumption is the proportional hazard structure (MPH model). We discuss these assumptions 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 ToE 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 the treatment is closely followed by an exit from the initial state, regardless of the time since 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 characteristics, x, 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 covariates that change with the elapsed duration, for instance due to business cycle variation or seasonal variation, are a useful source of variation (Gaure et al., 2007). The intuition is that such time-varying covariates shift the hazard rates, and this is informative on the influence of the unobserved heterogeneity. More specifically, current factors have an immediate impact on the exit rate, whereas past factors affect the current transition probabilities only through the selection process (for a more detailed discussion, see van den Berg and van Ours, 1994, 1996). We therefore examine both ToE models with time-invariant covariates only and specifications that include time-varying covariates.

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

3 Simulation approach

3.1 The basic idea

The idea behind EMC designs is to simulate by using real data, as opposed to using a data generating process entirely specified by the researcher 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 thus 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 jobseekers in Germany to compare the performance of alternative estimators of treatment effects under conditional independence. They proceed in the following way. They first use the real data on both treated and non-treated units 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 units 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 the simulations are based on real data. Moreover, by construction, the true effect of the placebo treatments is zero. Then, Huber et al. (2013) use the resulting simulated data to analyze the performance of various CIA-based estimators.

We tweak this simulation design in some key dimensions with the aim of using the EMC approach to study the ToE model. We use rich Swedish administrative register data and survey data of jobseekers, 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. Then, we use this data to generate placebo treatments, but we do this in a slightly different way than Huber et al. (2013). Instead of simulating binary treatment indicators as they do, we use a 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 omitting blocks of the covariates that were previously used in the true selection model to produce the placebo treatment durations. This leads to a bivariate duration model with correlated unobserved determinants, since the excluded variables affect both the time in unemployment (the outcome) and, by construction, the treatment duration.

The simulated data is used for various simulation exercises. We mainly focus on the estimation of the treatment effect. By construction, the true effect of the placebo treatments is zero, but since we leave out variables and generate correlated unobserved determinants, we introduce bias (estimated treatment effect non-zero). To evaluate important specification issues related to ToE models, we study the impact on the bias and the variance of the treatment effects estimates, but we also study other parts of the model. Some of these issues that we study were raised by previous Monte Carlo simulations studies (Gaure et al., 2007; Baker and Melino, 2000). This includes the specification of the unobserved heterogeneity distribution. However, we also study specification aspects that have not been studied before. One example is that we exclude different blocks of covariates, with the aim of studying how the ToE approach performs with different types of unobserved heterogeneity.

One important reason to use the Swedish unemployment spell data is that there are many examples of evaluations that estimate ToE models using this type of data (see Section 1). In addition, unemployment durations and labor market program entries are measured at the daily level. We treat the daily spell data as if it were continuous, and generate placebo treatment durations measured at the daily level by using a continuous-time selection model. Accordingly, we estimate continuous-time ToE models.

Next, let us relate our simulated data to the assumptions made in the ToE approach. 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) may not hold in our simulations, since the excluded variables representing unobserved heterogeneity may be correlated with the variables that were actually used in the ToE estimation.[8] To explore this, we leave out blocks of variables that are alternatively highly or mildly correlated with the observables. It turns out that the degree of correlation between the observed and unobserved factors is relatively unimportant. Second, a duration model without embedded unobserved heterogeneity is used to model the treatment selection process. This means that although we use an extremely rich set of variables to estimate the selection process, mimicking the information available to caseworkers when assigning treatments, the model may be misspecified if there are omitted characteristics.

[8] Likewise, indicators of past individual labor market outcomes included in the vector of covariates may be stochastically dependent on unobserved heterogeneity.

3.2 The relevance of different covariates

The analysis of the ToE model specification is the main contribution of our paper. However, by leaving out different blocks of covariates, we can also evaluate the relevance of different observables when measuring causal effects of active labor market programs. To this end, we use the simulated data with placebo treated and non-treated units, for which the "true" treatment effect is known to be zero. To assess the relative importance of different covariates, we leave out alternative blocks of observables and compare the bias size across the resulting specifications.

These analyses benefit from the rich Swedish data. We first follow Lechner and Wunsch (2013), who create variables that capture essentially all covariates claimed to be important for the selection process and used in various CIA-based evaluations of active labor market programs. Lechner and Wunsch use German data, and we use Swedish databases to re-construct similar covariates. However, we also include additional covariates not used by Lechner and Wunsch (2013). First, since we model treatment durations and not binary treatment indicators, we also include covariates that capture the duration aspect of employment and unemployment histories. The idea is that information on previous durations may capture aspects related to how long one stays unemployed in a better way than non-duration history variables. By comparing with other unemployment and employment history variables, such as the employment rate, we can see if indeed previous durations matter more for current duration outcomes.

Second, the covariates in Lechner and Wunsch (2013) reflect important aspects of labor market attachment, skills and benefit variables, but more general unobserved skills may also be relevant. To study this, we use parental income, a commonly used proxy for such general unobserved skills. Third, since we model treatment durations, certain time-varying covariates may be important factors. In particular, we consider business cycle conditions, which might change over time, especially for longer unemployment spells. Another difference compared Lechner and Wunsch (2013) is that here we consider a duration outcome framework, and use duration models to study the relevance of different blocks of covariates.

Note that this procedure holds under the assumption of CIA with the full set of covariates. Lechner and Wunsch (2013) provide good arguments as to why CIA should be valid in their German setting when they use their full set of covariates, and Vikström (2017) provides similar arguments for Sweden. This can of course always be questioned, for instance, because treatment selection is based on unobserved motivation and skills. Thus, we study the relevance of the different observed

covariates, keeping in mind that there may also be important information that is not included in our data.

3.3 The 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 type of 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 the jobseekers so as to enhance their chances of finding a job. Training courses include manufacturing, machine operator, office/warehouse work, health care, and computer skills. The basic eligibility criterion is to be at least 25 years 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 smaller. In all cases, training is free of charge.

Previous evaluations of the effects of the AMU training program on unemployment include Harkman and Johansson (1999), de Luna et al. (2008), Richardson and van den Berg (2013), and van den Berg and Vikström (2019). These papers describe the training program in great detail.

3.4 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 in Sweden. We use this information to construct spell data on the treatment duration (time to the 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. Any ongoing spells are right-censored on December 31, 2013.

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 patterns due to early retirement decisions of older workers. We also exclude disabled workers. In total, there are 2.6 million sampled spells, of which 3% involve training participation. The mean unemployment duration in the sample is 370 days. In case a job seeker enters into training multiple times, only the first instance is considered.

For each spell, we construct detailed information on individual-level characteristics. We start by constructing similar covariates as in the German data in Lech-

ner and Wunsch (2013).[9] The population register 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). From Unemployment Insurance (UI) records we obtain information on UI eligibility.

The data from the Public Employment Service is also used to construct unemployment history variables, and 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 two years) and more long-run history (last ten years). Altogether, this captures many aspects of the workers employment and earnings history in the last two or ten years.

As already mentioned, we also include additional covariates not used by Lechner and Wunsch (2013). These include previous unemployment and employment durations, the idea being that previous durations may capture the current ones in a better way than the above-mentioned employment history variables. To this aim, we construct time spent in the last employment spell, time in the last unemployment spell as well as indicators for no previous unemployment/employment spell. We also study the relevance of controlling for the mother's and father's income, under the assumption that parental income may capture general unobserved skills. Here, we exploit the Swedish multi-generational register (linking children to parents) together with income registers to create information on parental income (father and mother income, averaged over age 35-55 of the parent). Finally, we also explore time-varying covariates, and include the local unemployment rate in the region during each month as a time-varying covariate (Sweden has 21 regions).

The outcome considered in this paper is the re-employment 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.

All covariates that are used in the analyses are summarized in Table 1. The statistics in the table show that immigrants from outside Europe, males, married and

[9] There are some differences between the Swedish and German data. The classification of occupations differs, we lack some firm-level characteristics, and we have less information on UI claims. We also use welfare benefits transfers to construct measures of out-of-labor-force status.

the less educated jobseekers are over-represented among the training participants. Training participants also also more likely to be employed in firms with lower wages, and there are 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 two and ten years.

3.5 Simulation details

Selection model. The first step of the EMC design is to estimate the treatment selection model. We use a continuous-time parametric proportional hazard model for the treatment hazard, $\theta_p(t|x)$, at time, t, conditional on a set of covariates, x, which includes time-fixed covariates and time-varying monthly regional unemployment rate:[10]

$$\theta_p(t|x) = \lambda_p(t) \cdot \exp(x\beta_p). \tag{2}$$

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 for the m^{th} time interval. We use eight time intervals, with splits after 31, 61, 122, 183, 244, 365 and 548 days. The included covariates are listed in Table 1. The model estimates, also 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 by using the full population of actual treated and controls (i.e. the never treated), the treated units are discarded and play no further role in the simulations. Next, we use equation (2) to simulate the placebo times to treatment for each non-treated, T_s, which is generated according to (dropping x to simplify the notation):

$$\exp\left(-\int_0^{T_p} \theta_p(\tau)d\tau\right) = U, \tag{3}$$

where $U \sim \mathcal{U}[0,1]$. Since $\theta_p(t) > 0 \, \forall t$, the integrated hazard $\int_0^{T_p} \theta_p(\tau)d\tau$ is strictly

[10] Alternatively, one could use a semi-parametric single-index estimator for the hazard rate of $T_p|X$, for example the Gørgens (2006) estimator. However, this would be numerically cumbersome and since this does not impose a PH structure the resulting model may not be compatible with any ToE model.

increasing in T_p. By first randomly selecting U for each unit and then finding the unique solution to (3), we can retrieve T_p for each observation.[11]

Simulated treatments that occur after the actual exit from unemployment are ignored. Thus, the placebo treated units are those with a placebo treatment realized before the exit to job. During this procedure, $\hat{\theta}_p(t|x_i)$ is multiplied by a constant γ, which is selected such that the share of placebo treated is around 20%. This ensures that there is a fairly large number of treated units in each sample, even if the sample size is rather small. A similar approach is adopted by Huber et al. (2013).

Simulations. The placebo treatments are simulated for all non-treated units. Next, we draw random samples of size N from this full sample (independent draws with replacement). We set $N = 10{,}000$, 40,000 and 160,000 because ToE models are rarely estimated with 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, or from 40,000 to 160,000) should reduce the standard error by 50%. For each ToE specification we perform 500 replications.

3.6 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 support 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, $L_i(v)$, conditional on the vector of unobserved variables $v = (v_e, v_p)$. Then, the individual likelihood contribution, L_i, is obtained by integrating $L_i(v)$ over the distribution of the unobserved heterogeneity, $G(V)$. 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 $\alpha_{sm} = \mathbb{1}[t \in [t_{m-1}, t_m)]$, for $m = 1, \ldots, M$ cut-offs. We fix the cut-offs to $31, 61, 122, 183, 244, 365, 548, 2160$. In the

[11] The actual distribution for the integrated hazard will depend on the specification of the selection model in equation (2). In the simple case where all covariates are time-fixed and the placebo treatments are generated by using a proportional hazard model that has two piecewise constant parts, with θ_s^0 for $t \in [0, t_1)$ and θ_s^1 for $t > t_1$:

$$\exp\left(-\int_0^{T_s} \theta_s(\tau)d\tau\right) = \begin{cases} \exp\left(-\int_0^{T_s} \theta_s^0 d\tau\right) & \text{if } U > \exp\left(-\int_0^{t_1} \theta_s^0 d\tau\right) \\ \exp\left(-\int_0^{t_1} \theta_s^0 d\tau - \int_{t_1}^{T_s} \theta_s^1 d\tau\right) & \text{otherwise} \end{cases}$$

This can be easily extended to the case where the baseline hazard has more than two locally constant pieces and where X contains time-varying covariates (in both cases, the integrated hazard shifts in correspondence of changes in such covariates over calendar- or duration-time).

section we discuss the observed variables used in the model.

To set up $L_i(v)$, we split the spells into parts where all right-hand side variables in equation (1) are constant. Splits occur at each new spell-duration indicator and when the treatment status changes. In all baseline ToE specifications, the covariates specified are calendar-time constant. In additional specifications where the time-varying local unemployment rate is included, calendar-time variation leads 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 such possible states (employment and treatment). Then, with approximately continuous durations, $L_i(v)$ is:

$$L_i(v) = \prod_{c_{ij} \in C_i} \left[\exp\left(-l_{ij} \sum_{s \in S_{it}} \theta_s(t, x_i, D_{it}, v_s|\cdot) \right) \times \prod_{s \in S_{it}} \theta_s(t|\cdot)^{y_{sij}} \right], \tag{4}$$

with

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

L_i is obtained by integrating $L_i(v)$ 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:

$$\mathcal{L} = \sum_{i=1}^{N} \left(\sum_{w=1}^{W} p_w \ln L_i(v_w) \right) \equiv \sum_{i=1}^{N} L_i. \tag{5}$$

Search algorithm. To estimate the discrete support points, we use the iterative search algorithm in Gaure et al. (2007). 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 $\hat{\vartheta}_{W-1}$ 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.[12] 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 $\overline{W}$ maximum likelihood 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 (up to six points). An alternative approach is to increase the number of support points until there is no further improvement in the likelihood (ML criterion). It is defined as $ML = \mathcal{L}(\hat{\vartheta}_W)$, where only likelihood increases greater than 0.01 are considered. We also use information criteria that penalize parameter abundance. Specifically, the 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 in the estimation.[13]

All 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 all information criteria.

4 Available covariates and evaluations of ALMPs

We now evaluate the relevance of different types of covariates. Specifically, we leave out various blocks of covariates and compare the size of the bias – the difference between the estimated treatment effect and the true zero effect of the placebo treatments – across specifications. All covariates are a subset of those used to generate

[12] 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.

[13] We follow Gaure et al. (2007) and use the grand total number of spell parts. N can be alternatively used, but our simulations indicate that this is of minor importance in practice.

the placebo treatments. For each specification, the full sample of placebo treated and placebo non-treated units is used to estimate a parametric proportional hazard (PH) model. Here, the baseline hazard is specified in the same way as for the model used to simulate the placebo treatments.[14] Table 1 lists all covariates in each block.

The main results are given in Table 2. In each panel of the table, we start with the covariates from the proceeding panels and add additional information to the covariates already in the model, so that the model is extended sequentially by adding blocks of covariates one by one. This will, for instance, reveal the relevance of adding information on long-term labor market history on top of the more basic covariates such as short-term history and baseline socio-economic characteristics.[15]

In Panel A, we start with a baseline model with a set of baseline socio-economic characteristics, which returns a positive and sizable bias of around 6.9%. That is, the estimated treatment effect is 0.069 when the true effect of these placebo treatments is equal to zero. Additionally controlling for calendar time (inflow year and month dummies) and regional information (regional dummies and local unemployment rate at inflow) reduces the bias from 6.9% to 6.2%.[16] Since the corresponding excluded covariates include short- and long-term labor market history, the positive bias means that training participants tend to have more favorable labor market histories.

Panel B compares the relevance of short-term employment, unemployment, earnings and welfare benefit histories. Here, we compare the relevance of entire blocks of covariates, while later we do so for individual variables, such as previous employment rates against employment durations. All blocks of short-term history covariates 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 status). If we adjust for unemployment history and earnings history, the bias drops to 5.0% and 4.0%, respectively, whereas if the model includes employment history the bias is much closer to zero. In fact, the sign of the bias is even reversed (slightly negative, -1.4%) when adjusting for short-term employment history. These results indicate that participants in labor market training are to a large extent selected based on their previous employment records. One explanation may be that caseworkers aim to select jobseekers with an occupational history aligned

[14] We have also estimated the bias using other duration models, including a Cox-model, leading to similar results.

[15] We add the covariates in a similar order as Lechner and Wunsch (2013), who argue that the order resembles the ease, likelihood and cost of obtaining the respective information.

[16] For completeness, we also report estimates when using these time and regional variables only, without including the baseline socio-economic characteristics. This leads to larger bias.

with the vocational training program.

We next examine what specific aspects of employment and unemployment that are the most important to adjust for. We control for either past employment duration, different measures of the share of time spent in employment (employment rate), employment status at a given point in time, or other history variables. A reason for this exercise is that we model treatment durations and not a binary treatment status. Accordingly, it may be the case that previous durations capture aspects of the ongoing unemployment spell in a better way than previous employment rates and employment status at a given point in time. Table 3 shows that information on previous employment duration reduces the bias considerably: from 6.2% in the baseline specification to 3.9% (Panel A). However, adding information on past employment rates or other short-term employment history variables reduces the bias even more, leading to biases of -0.04% and 0.2%, respectively (Panel B and C). In particular, Panel B shows that all covariates measuring past employment rate single-handedly capture a large part of the bias. We also note that the bias is positive or close to zero in all cases, so that the reversal of the bias sign that was observed in Panel B of Table 2 occurs only once all short-term employment variables are included together. That is, even if some short-term history variables are more relevant, they all capture different aspects of the selection process, so that adjusting for both previous employment durations and rates is important.

Panels D to F of Table 3 report estimates from a similar exercise where we control for the short-term unemployment history and duration variables one at a time. This confirms that unemployment history variables have a modest impact on the estimated bias compared to the employment history variables. All in all, this suggests that for training programs with emphasis on human capital accumulation, the most important characteristics to control for are those related to employment history.[17,18]

Next, let us return to Table 2. Here, Panel C shows that adding information on long-term labor market history (last ten years) on top of short-term history (last 2

[17] We also tried to additionally include past employment and unemployment durations more flexibly, by specifying them on logarithmic- and quadratic-scale, and by including information from the previous two spells. The bias is only slightly reduced compared to the information reported in Table 3, and qualitatively all patterns are unaffected.

[18] It may be argued that aspects of past unemployment experience are good indicators of the unobserved heterogeneity term V_e in the current spell. For example, in MPH duration models, the log mean individual duration is additive in V_e. This would suggest that inclusion of such aspects as covariates strongly reduces the bias. However, note that the actual bias in the estimated treatment effect also depends on the extent to which these aspects affect treatment assignment over and above the included determinants of the latter.

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, including additional covariates such as long-term labor market history is relatively unimportant. This is also consistent with the results in Heckman et al. (1998), Heckman and Smith (1999), Mueser et al. (2007), and Dolton and Smith (2010), who find that it is important to control for regional information, labor market history and pre-treatment outcomes. However, one difference compared to Lechner and Wunsch (2013) is that in this setting adjusting for short-term employment history is enough to obtain small bias, whereas 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).

Panel D examines the relevance of parental income, which we use to proxy for general unobserved skills. This may be important if unobserved skills are not captured by the covariates discussed so far, which are mainly related to labor market attachment. However, parents' income turns out to have limited impact on the bias, at least once we control for both short- and long-term labor market history variables. This indicates that labor market histories are also able to capture more general unobserved skills.[19]

5 Specification of ToE models

This section presents the main simulation results. The main focus is on the (placebo) treatment effects. 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 leads to the best results in terms of average bias, variance of the placebo estimates, and mean squared error (MSE).

5.1 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, inflow time dummies,

[19] This is consistent with the results in Caliendo et al. (2017), which finds that once one controls for rich observables of the type that we include here, additional (usually unobserved) characteristics measuring personality traits and preferences become redundant.

regional indicators and unemployment rate (the covariates in Panels A–B, Table 1). Here, we control for time-fixed regional unemployment rate (measured as the month of inflow into unemployment). Later, in Table 8, we estimate ToE models with time-varying regional unemployment rate.

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 unobserved heterogeneity (one support point) leads to large bias (6.0%).[20] This confirms that under-correcting for unobserved heterogeneity may lead to substantial bias. However, already with two support points the bias is reduced from 6.0% to 2.7%.[21] For three or more support points, the average bias is even larger and keeps increasing in the same direction when adding additional support points. In fact, with six support points the average bias (6.4%) is more than twice as large as the average bias with two support points (2.7%). Moreover, both the variance and the MSE increase in the number of support points (Columns 2–3).

The increased bias due to too many support points is consistent with the results from Baker and Melino (2000), which argue that specifications with too many (spurious) support points tend to 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, the model compensates this with changes (bias) in the treatment effect, and presumably also in the duration dependence. This pattern contradicts the general intuition that one should always adjust for unobserved heterogeneity in the most flexible way in order to avoid bias due to unaccounted unobserved heterogeneity.

To better understand the over-correction pattern, Figure 1 shows the distribution of the treatment effect estimates for one, two and six support points. With one support point, the estimates are centered around a bias of around 6% and the variance of the estimates is relatively low. With two support points the entire distribution shifts towards zero (although the average bias is non-zero), but the variance gets 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

[20] 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 obtained from the full set of placebo treated and non-treated observations.

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

the estimates shifts to the right (larger positive bias). This shows that the increased bias is not explained by a few extreme estimates.

Interestingly, the problem with over-correcting for unobserved heterogeneity does not occur to the same extent in the simulated data used by Gaure et al. (2007). They highlight that the main problem is under-correction with too few support points. Our simulation results that are based on real data, instead, suggest that both under- and over-correction are important problems when estimating ToE models. Thus, finding a way to select the appropriate number of support points appears to be important. We explore this in the next section.[22]

5.2 Information criteria

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 reports the average number of support points that are selected according to each criterion. The ML criterion, where the number of support points is increased as long as the likelihood is improved, leads to 4.11 support points on average. The bias and variance are large compared to simply pre-specifying two or three 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). This pattern is confirmed in all simulation settings presented below. As a result, criteria with little penalty for parameter abundance, such as the ML criterion, should be avoided altogether when selecting the number of mass points.

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 achieved when pre-specifying a low 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 so 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 (one 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

[22] In their main simulations, Gaure et al. (2007) find no evidence that too many support points over-correct for unobserved heterogeneity. However, when they reduce the sample size they also find evidence of some over-correction.

HQIC) lead to models with fewer support points, and the average bias is slightly lower than for the less restrictive AIC criterion. The variance is also slightly 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, later we will see that none of the three criteria 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, the opposite holds, and this favors the more restrictive BIC and HQIC criteria. Thus, using all three criteria and reporting several estimates as robustness check appears to be a reasonable approach.

The main interest here is in providing background information on the alternative specification choices. However, Table 4 also provides some insights on the overall idea of using ToE models to adjust for unobserved heterogeneity. In general, the table shows that the ToE approach corrects for a large share of the bias, which is reduced from 6.0% for the model without unobserved heterogeneity to around 2.7% when 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 derives from time-fixed observed covariates. In subsequent analyses, we explore whether additional sources of exogenous variation in the form of time-varying covariates further reduce the bias.

5.3 Sample size

In Columns 4–6 and 7–9 of Table 4, the sample size is increased to 40,000 and 160,000 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 two to six support points increases the bias from 2.7% to 6.4%, and with 40,000 observations, it increases from 2.2% to 3.7%. For the largest sample with 160,000 observations, the increase in the bias when going from two to six support points is even smaller. It shows that over-correction, due to too many support points, mainly is a problem with small sample sizes. However, note that what constitutes a small sample size most likely differs across applications. For instance, it might be related to the number of parameters in the model, the fraction of treated units, the number of exit states, and the variation in the observed variables.

Another result is that for larger sample sizes there are smaller differences between the different information criteria. For instance, with a sample size of 160,000, there are virtually no differences in the average bias between the four criteria.

5.4 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 socio-economic characteristics, inflow time dummies and regional information. Here, we generate more unobserved heterogeneity by excluding additional covariates (all the socio-economic characteristics in Panel A of Table 1) and less heterogeneity by excluding fewer covariates (earnings history in Panel F of Table 1). Table 3 shows that these models generate a bias of 9.5% and 4.0%, respectively, in the full sample of placebo treated and controls (Panels A and B). These values can be compared to the bias of 6.2% in the baseline setting.

Columns 1–3 of Table 5 report the results for the model with more extensive unobserved heterogeneity. Again, the ToE model adjusts for a large share of the bias due to unobserved heterogeneity. For instance, with a sample size of 10,000, the bias for the specification without unobserved heterogeneity is 9.4%, but it drops to 2–3% when we adjust for unobserved heterogeneity using the AIC, BIC or HQIC criteria (Panel A). As before, these more restrictive criteria return the lowest bias, whereas the ML criterion leads to a model with too many support points.[23] Again, this is consistent with previous results.

Overall, the specification with less substantial unobserved heterogeneity, obtained by excluding fewer covariates, produces similar patterns (Columns 4–6 of Table 5). The main difference concerns the relative performance of the AIC, BIC and HQIC criteria. Consider the results for a sample size of 40,000. With more extensive unobserved heterogeneity (Columns 1–3), the bias for the AIC criterion is 0.9%, whereas it is 1.8% and 1.9% for the BIC and HQIC criteria, respectively. This suggests that the more restrictive information criteria (BIC and HQIC) may under-correct for the substantial unobserved heterogeneity by favoring models with too few support points, and this leads to larger bias. This pattern is reversed when we create less substantial unobserved heterogeneity by excluding fewer covariates (Columns 4–6). Here, the average bias is lower for the more restrictive BIC and HQIC criteria than for AIC. This is because for this specification, there likely is a

[23] We obtain similar results with 40,000 observations, but here the difference between the ML criterion and the other criteria is smaller.

larger risk of over-correcting for unobserved heterogeneity, favoring criteria with a larger penalty for parameter abundance. From all this, we conclude that neither one of the information criteria is superior in all settings.

5.5 Degree of correlation between *X* and *V*

Since we use single-spell data, identification of the ToE model requires independence between the included covariates and the unobserved heterogeneity (random effects assumption). This may not hold in our setting, since we create unobserved heterogeneity by leaving out certain blocks of covariates, and these excluded covariates may be correlated with those that we include when we estimate the ToE model. We therefore perform additional simulation exercises leaving out different blocks covariates from the model. We consider three settings with strongly positive, mildly positive and negative correlation between the covariates used in the ToE model and the excluded covariates, respectively.[24] We select covariates to include in the model such that the starting bias, corresponding to the specifications with one support point (no unobserved heterogeneity), is similar across the alternative degrees of correlation (between 4.4% and 4.8%).

Panel A of Table 6 shows the simulation results with samples of size 10,000. Overall, the information criteria perform similarly as before. The ML criterion selects a larger number of support points which leads to larger bias, and the AIC, BIC and HQIC criteria select more parsimonious models characterized by lower bias than for the ML criterion. Importantly, this holds regardless of the degree of correlation between the observed and the unobserved variables. This is reassuring: even when the variables left out from the model are largely related with those left in the ToE model, the relative performance of the information criteria does not appear to be affected. We obtain similar results when drawing samples of size 40,000 (Panel B of Table 6).

5.6 Estimation of the unobserved heterogeneity distribution

So far we have focused on the treatment effect, but the overall performance of the ToE model can be also checked by inspecting to what extent the estimated discrete distributions for the unobserved heterogeneity approximates the true one. To exam-

[24] To compute the correlation, we use the estimates from the selection model with all covariates reported in Table 1. Then, for each cross-sectional unit, the estimated parameters are used to compute the linear predictor of the excluded covariates. This linear predictor equals V in the simulations. Finally, we correlate this with the observed covariates used in the model (by using the linear predictor of all included covariates). This produces one measure of the correlation between the observed and unobserved covariates in the model.

ine this, we focus on the unobserved heterogeneity for the treatment duration, T_p. For this duration, the true unobserved heterogeneity, V_p, is known since we generate it by leaving out certain blocks of covariates. However, since we do not simulate the outcome durations, the exact composition of V_e is unknown.

Specifically, for each actual treated and control unit, we use the coefficients of the estimated selection model reported in Table 1 to compute the linear predictor of the variables left out from the model. This linear predictor corresponds to V_p in the model. We compare the first two moments of this true unobserved heterogeneity with the corresponding moments for the estimated unobserved heterogeneity from the ToE models (with samples of size 10,000).

The results from this exercise are shown in Table 7. The table reports results for the true unobserved heterogeneity (Panel A) and the estimated unobserved heterogeneity (Panels B–C). Panel B shows that a larger number of support points tend to overestimate the dispersion of the unobserved heterogeneity. The mean of the unobserved heterogeneity distribution tends to be slightly underestimated. Panel C indicates that the ML criterion returns an unobserved heterogeneity with too large variance when compared to the true variance, whereas for the more restrictive information criteria (AIC, BIC and HQIC) the variance is too small. However, overall, the ToE model appears to approximate well the true underlying unobserved heterogeneity distribution of the selection model.[25]

5.7 Time-varying covariates

Identification of the ToE model relies on variation in the observed exogenous covariates. This was the only source of exogenous variation exploited in the baseline simulations above. One result was that the ToE model adjusts for a large part of the selection due to unobserved heterogeneity, but it did not eliminate the bias entirely. We now examine if an additional variation in the form of time-varying covariates (local unemployment rate) can further reduce the bias. The idea is that time-varying covariates should be useful for identification since they generate exogenous shifts in the hazard rates that help to recover the distribution of the unobserved heterogeneity. Specifically, the time-varying covariate used is time-varying unemployment rate measured at the monthly level for each county (län). We refer to it as local unemployment rate. This time-varying covariate was included in the selection model to

[25] Note that all information criteria select the number of support points based on the joint assessment of the treatment and outcome equations. This complicates the interpretation of whether a given model fits the unobserved heterogeneity in the best way, since as mentioned we do not know the true unobserved heterogeneity distribution for the outcome equation.

simulate the placebo treatments. Here, the samples are of size 10,000.

The results from this exercise are presented in Table 8. The first row of Panel A shows that the bias without adjusting for unobserved heterogeneity (one support point) is 5.6%. As before, additional support points are then stepwise included (Panel A). The results confirm what was found in the baseline simulations: both under-correcting and over-correcting for unobserved heterogeneity leads to bias; the ML criterion tends to select models with an overly-dispersed unobserved heterogeneity, whereas the three criteria that penalize parameter abundance (AIC, BIC and HQIC) all perform well.

One important difference compared to the baseline simulations is that the average bias for the BIC and HQIC are now closer to zero. This confirms that exploiting time-varying covariates greatly helps identifying the model parameters. 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 attributes of the last job. This produced substantial bias in the model without unobserved heterogeneity. The importance of variation induced by time-varying covariates 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 modified a recently proposed simulation technique, the Empirical Monte Carlo approach, to evaluate the Timing-of-Events model. It has resulted in several conclusions on how to specify ToE models. Our simulations show that information criteria are a reliable way to specify the number of support points that approximate the unobserved heterogeneity distribution of the model. This result holds as long as the criteria include a substantial penalty for parameter abundance. Information criteria with little penalty for parameter abundance, such as the ML criterion, 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 adjustment for unobserved heterogeneity. But, none of the three criteria is superior in all settings.

Another result is that the ToE model is able to adjust for substantial unobserved heterogeneity generated by omitting large numbers of relevant and diverse covariates. The model is also able to approximate well the true underlying unobserved heterogeneity distribution of the treatment equation. As long as an appropriate information criterion is used, these patterns are robust across alternative specifications. Adding time-varying covariates (local unemployment rate) on top of time-invariant covariates, improves the performance of the ToE estimator.

We have also examined which observed covariates that are important confounders when evaluating labor market programs. Here, one conclusion is that it is important to adjust for short-term labor market histories, whereas adding long-term labor market histories appears to be less important. Controlling for short-term employment histories appears to be more effective than controlling for short-term unemployment histories. We also conclude that variables measuring the share of time spent in employment in the near past are valuable. Other types of short-term employment history variables, such as previous employment durations, also turn out to be important, but relatively less so.

References

Abbring, J. H. and van den Berg, G. J. (2003). The nonparametric identification of treatment effects in duration models. *Econometrica*, 71(5):1491–1517.

Abbring, J. H., van den Berg, G. J., and van Ours, J. C. (2001). Business cycles and compositional variation in U.S. unemployment. *Journal of Business and Economic Statistics*, 19(4):436–448.

Abbring, J. H., van Ours, J. C., and van den Berg, G. J. (2005). The effect of unemployment insurance sanctions on the transition rate from unemployment to employment. *The Economic Journal*, 115(505):602–630.

Advani, A., Kitagawa, T., and Słoczyński, T. (2019). Mostly harmless simulations? Using Monte Carlo studies for estimator selection. *Journal of Applied Econometrics*, 34(6):893–910.

Baert, S., Cockx, B., and Verhaest, D. (2013). Overeducation at the start of the career: Stepping stone or trap? *Labour Economics*, 25:123–140.

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

Bergemann, A., Pohlan, L., and Uhlendorff, A. (2017). The impact of participation in job creation schemes in turbulent times. *Labour Economics*, 47:182–201.

Bijwaard, G. E., Schluter, C., and Wahba, J. (2014). The impact of labor market dynamics on the return migration of immigrants. *Review of Economics and Statistics*, 96(3):483–494.

Bodory, H., Camponovo, L., Huber, M., and Lechner, M. (2016). The finite sample performance of inference methods for propensity score matching and weighting estimators. IZA Discussion Papers, No. 9706.

Briesch, R. A., Chintagunta, P. K., and Matzkin, R. L. (2010). Nonparametric discrete choice models with unobserved heterogeneity. *Journal of Business and Economic Statistics*, 28(2):291–307.

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

Caliendo, M., Künn, S., and Uhlendorff, A. (2016). Earnings exemptions for unemployed workers: The relationship between marginal employment, unemployment duration and job quality. *Labour Economics*, 42:177–193.

Caliendo, M., Mahlstedt, R., and Mitnik, O. A. (2017). Unobservable, but unimportant? The relevance of usually unobserved variables for the evaluation of labor market policies. *Labour Economics*, 46:14–25.

Crépon, B., Ferracci, M., Jolivet, G., and van den Berg, G. J. (2018). Information shocks and the empirical evaluation of training programs during unemployment spells. *Journal of Applied Econometrics*, 33(4):594–616.

de Luna, X., Forslund, A., and Liljeberg, L. (2008). Effekter av yrkesinriktad arbetsmarknadsutbildning för deltagare under perioden 2002-04 (Effects of vocational labor market training for participants in the period 2002–04). IFAU working paper, 2008:1.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.

Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1):151–161.

Dolton, P. and Smith, J. A. (2010). The impact of the UK New Deal for lone parents on benefit receipt. IZA Discussion Paper, No. 5491.

Fox, J. T., Kim, K. i., Ryan, S. P., and Bajari, P. (2012). The random coefficients logit model is identified. *Journal of Econometrics*, 166(2):204–212.

Frölich, M., Huber, M., and Wiesenfarth, M. (2017). The finite sample performance of semi- and non-parametric estimators for treatment effects and policy evaluation. *Computational Statistics and Data Analysis*, 115:91–102.

Gaure, S., Røed, K., and Zhang, T. (2007). Time and causality: A Monte Carlo assessment of the timing-of-events approach. *Journal of Econometrics*, 141(2):1159–1195.

Gautier, E. and Kitamura, Y. (2013). Nonparametric estimation in random coefficients binary choice models. *Econometrica*, 81(2):581–607.

Goffe, W. L., Ferrier, G. D., and Rogers, J. (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics*, 60(1-2):65–99.

Gooley, T. A., Leisenring, W., Crowley, J., and Storer, B. E. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine*, 18:695–706.

Gørgens, T. (2006). Semiparametric estimation of single-index hazard functions without proportional hazards. *The Econometrics Journal*, 9(1):1–22.

Harkman, A. and Johansson, A. (1999). Training or subsidized jobs–what works? Working paper, AMS, Solna.

Heckman, J. J., Ichimura, H., Smith, J. A., and Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66(5):1017–1098.

Heckman, J. J. and Singer, B. (1984). A Method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52(2):271–320.

Heckman, J. J. and Smith, J. A. (1999). The pre-programme earnings dip and the determinants of participation in a social programme: Implications for simple programme evaluation strategies. *The Economic Journal*, 109(457):313–348.

Holm, A., Høgelund, J., Gørtz, M., Rasmussen, K. S., and Houlberg, H. S. B. (2017). Employment effects of active labor market programs for sick-listed workers. *Journal of Health Economics*, 52:33–44.

Huber, M., Lechner, M., and Mellace, G. (2016). The finite sample performance of estimators for mediation analysis under sequential conditional independence. *Journal of Business and Economic Statistics*, 34(1):139–160.

Huber, M., Lechner, M., and Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics*, 175(1):1–21.

Huh, K. and Sickles, R. C. (1994). Estimation of the duration model by nonparametric maximum likelihood, maximum penalized likelihood, and probability simulators. *The Review of Economics and Statistics*, 76(4):683–694.

Ichimura, H. and Thompson, T. S. (1998). Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *Journal of Econometrics*, 86(2):269–295.

Jahn, E. and Rosholm, M. (2013). Is temporary agency employment a stepping stone for immigrants? *Economics Letters*, 118(1):225–228.

Lechner, M. and Strittmatter, A. (2017). Practical procedures to deal with common support problems in matching estimation. *Econometric Reviews*, pages 1–15.

Lechner, M. and Wunsch, C. (2013). Sensitivity of matching-based program evaluations to the availability of control variables. *Labour Economics*, 21:111–121.

Lindeboom, M., Llena-Nozal, A., and van der Klaauw, B. (2016). Health shocks, disability and work. *Labour Economics*, 43:186–200.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: A general theory. *The Annals of Statistics*, 11(1):86–94.

McVicar, D., Moschion, J., and van Ours, J. C. (2018). Early illicit drug use and the age of onset of homelessness. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(1):345–372.

Mueser, P. R., Troske, K. R., and Gorislavsky, A. (2007). Using state administrative data to measure program performance. *Review of Economics and Statistics*, 89(4):761–783.

Narendranathan, W. and Stewart, B., M. (1993). Modeling the probability of leaving unemployment: competing risks models with flexible base-line hazards. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 42(1):63–83.

Palali, A. and van Ours, J. C. (2017). Love conquers all but nicotine: spousal peer effects on the decision to quit smoking. *Health Economics*, 26(12):1710–1727.

Richardson, K. and van den Berg, G. J. (2013). Duration dependence versus unobserved heterogeneity in treatment effects: Swedish labor market training and the transition rate to employment. *Journal of Applied Econometrics*, 28(2):325–351.

Ridder, G. (1987). The sensitivity of duration models to misspecified unobserved heterogeneity and duration dependence. Unpublished manuscript.

Smith, J. A. and Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1-2):305–353.

van den Berg, G. J. (2001). Chapter 55 Duration models: Specification, identification and multiple durations. In *Handbook of Econometrics*, volume 5, pages 3381–3460. Elsevier.

van den Berg, G. J. and Gupta, S. (2015). The role of marriage in the causal pathway from economic conditions early in life to mortality. *Journal of Health Economics*, 40:141–158.

van den Berg, G. J. and van Ours, J. C. (1994). Unemployment dynamics and duration dependence in France, The Netherlands and the United Kingdom. *The Economic Journal*, 104(423):432.

van den Berg, G. J. and van Ours, J. C. (1996). Unemployment dynamics and duration dependence. *Journal of Labor Economics*, 14(1):100–125.

van den Berg, G. J. and Vikström, J. (2019). Long-run effects of dynamically assigned treatments. IFAU working paper, 2019:18.

van Ours, J. C. and Williams, J. (2009). Why parents worry: Initiation into cannabis use by youth and their educational attainment. *Journal of Health Economics*, 28(1):132–142.

van Ours, J. C. and Williams, J. (2012). The effects of cannabis use on physical and mental health. *Journal of Health Economics*, 31(4):564–577.

van Ours, J. C., Williams, J., Fergusson, D., and Horwood, L. J. (2013). Cannabis use and suicidal ideation. *Journal of Health Economics*, 32(3):524–537.

Vikström, J. (2017). Dynamic treatment assignment and evaluation of active labor market policies. *Labour Economics*, 49:42–54.

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	
			Est.	Std. Er.
Number of observations	76,302	2,564,561	2,640,863	
Panel A: Baseline socio-economic characteristics				
Country of origin: Not Europe	0.20	0.16	0.0910***	(0.0120)
Age 25-29	0.23	0.26	0.1366***	(0.0126)
Age 30-34	0.20	0.20	0.1188***	(0.0117)
Age 40-44	0.16	0.15	-0.0363***	(0.0123)
Age 45-49	0.12	0.11	-0.1441***	(0.0137)
Age 50-54	0.09	0.09	-0.3510***	(0.0160)
Male	0.67	0.51	0.4719***	(0.0091)
Married	0.35	0.34	0.0017	(0.0089)
Children: At least one	0.43	0.43	0.1265***	(0.0100)
Children: No. of children in age 0-3	0.20	0.20	0.0565***	(0.0116)
Education: Pre-high school	0.18	0.17	-0.1432***	(0.0253)
Education: High school	0.57	0.50	0.0624**	(0.0248)
Education: University College or higher	0.22	0.31	-0.0490**	(0.0250)
Panel B: Inflow time and regional information				
Beginning of unemployment: June-August	0.26	0.30	-0.0135	(0.0084)
Inflow year: 2003-2005	0.30	0.35	-0.3952***	(0.0217)
Inflow year: 2006-2007	0.16	0.18	-0.2562***	(0.0230)
Inflow year: 2008-2009	0.23	0.18	-0.3304***	(0.0233)
Inflow year: 2010-2011	0.18	0.17	-0.2455***	(0.0240)
Region: Stockholm	0.13	0.21	-0.3412***	(0.0158)
Region: Gothenborg	0.13	0.16	-0.3634***	(0.0127)
Region: Skane	0.12	0.14	-0.2910***	(0.0129)
Region: Northern parts	0.21	0.15	0.1647***	(0.0112)
Region: Southern parts	0.14	0.12	0.0111	(0.0126)
Monthly regional unemployment rate	10.54	9.77	0.0234***	(0.0021)
Panel C: Short–term employment history (2 years) and employment duration				
Time employed in last spell	859.82	831.20	0.0000	(0.0000)
Missing time employed in last spell	0.20	0.17	0.0493***	(0.0150)
Months employed in last 6 months	3.37	3.54	-0.0003	(0.0039)
Months employed in last 24 months	12.79	13.50	0.0040***	(0.0013)
No employment in last 24 months	0.22	0.19	-0.1354***	(0.0250)
Time since last employment if in last 24 months	2.31	2.42	-0.0069***	(0.0015)
Number of employers in last 24 months	1.66	1.79	0.0115***	(0.0035)
Employed 1 year before	0.59	0.59	0.0353***	(0.0122)
Employed 2 years before	0.59	0.59	0.0207*	(0.0122)
Panel D: Short–term unemployment history (2 years) and unemployment duration				
Time unemployed in last spell	107.11	89.43	0.0000	(0.0000)
Missing time unemployed in last spell	0.53	0.51	0.0213*	(0.0130)
Days unemployed in last 6 months	18.94	14.79	0.0008***	(0.0002)
Days unemployed in last 24 months	143.53	120.87	0.0003***	(0.0000)
No unemployment in last 24 months	0.44	0.44	-0.0511***	(0.0150)
Days since last unempl. if in last 24 months	15.12	14.76	0.0001	(0.0001)
Number of unempl. spells in last 24 months	0.82	0.88	0.0033	(0.0060)

Continue to next page

Table 1 – continued from previous page

	Treated	Control	Selection model	
			Est.	Std. Err.
Unemployed 6 months before	0.20	0.16	0.0171	(0.0151)
Unemployed 24 months before	0.24	0.22	-0.0327***	(0.0121)
Any program in last 24 months	0.03	0.02	0.0579**	(0.0291)
Panel E: Short–term welfare history (2 years)				
Welfare benefits -1 year	4928.00	3742.27	0.0318***	(0.0078)
Welfare benefits -2 years	4258.73	3542.66	0.0075	(0.0095)
On welfare benefits -1 year	0.19	0.14	0.0028	(0.0166)
On welfare benefits -2 years	0.17	0.14	-0.0720***	(0.0163)
Panel F: Earnings history (2 years)				
Earnings 1 year before	111684.78	110247.91	0.0095*	(0.0055)
Earnings 2 years before	111858.48	110612.95	-0.0157*	(0.0094)
Panel G: Long-term employment history (10 years)				
Months employed in last 10 years	58.19	62.91	-0.0022***	(0.0002)
Number of employers in last 10 years	4.72	5.12	0.0119***	(0.0012)
Cumulated earnings 5 years before	533484.45	530466.42	0.0629***	(0.0114)
Panel H: Long-term unemployment history (10 years)				
Days unemployed in last 10 years	788.31	693.41	-0.0001***	(0.0000)
No unemployment in last 10 years	0.18	0.17	-0.0890***	(0.0158)
Days since last unemployment if in last 10 years	256.77	290.49	-0.0000***	(0.0000)
Number of unemployment spells in last 10 years	3.63	3.83	0.0074***	(0.0018)
Average unemployment duration	95.31	90.15	-0.0001***	(0.0000)
Duration of last unemployment spell	180.26	154.83	-0.0001***	(0.0000)
Any program in last 10 years	0.15	0.12	0.0348	(0.0227)
Any program in last 4 years	0.06	0.05	0.0509**	(0.0243)
Number of programs in last 10 years	0.19	0.15	0.0342**	(0.0157)
Panel I: Long-term welfare history, out-of-labor-force (10 years)				
Yearly average welfare benefits last 4 years	4239.77	3533.38	-0.0213	(0.0142)
Yearly average welfare benefits last 10 years	3918.49	3448.42	-0.0828***	(0.0086)
No welfare benefits last 4 years	0.69	0.75	-0.0824***	(0.0150)
No welfare benefits last 10 years	0.51	0.59	-0.0946***	(0.0109)
Panel J: Characteristics of the last job				
Wage	18733.31	18860.58	-0.0597***	(0.0052)
Wage missing	0.54	0.52	-0.0215	(0.0337)
Occupation:				
Manager	0.04	0.07	-0.3102***	(0.0388)
Requires higher education	0.04	0.06	-0.1240***	(0.0375)
Clerk	0.04	0.05	-0.0037	(0.0374)
Service, care	0.09	0.13	-0.0047	(0.0357)
Mechanical, transport	0.13	0.07	0.2107***	(0.0352)
Building, manufacturing	0.06	0.05	0.0597	(0.0371)
Elementary occupation	0.05	0.05	-0.0044	(0.0375)
Panel K: Characteristics of the last firm				
Firm size	2523.01	3873.70	0.0000**	(0.0000)
Age of firm	12.95	14.13	0.0006	(0.0009)
Average wage	21588.62	21517.77	0.0007	(0.0048)
Wage missing	0.62	0.58	-0.0459	(0.0541)

Continue to next page

Table 1 – continued from previous page

	Treated	Control	Selection model	
			Est.	Std. Err.
Mean tenure of employees	3.43	3.68	-0.0029	(0.0024)
Age of employees	27.74	29.44	-0.0033***	(0.0009)
Share of immigrants	0.12	0.13	-0.1709***	(0.0255)
Share of females	0.26	0.34	-0.4736***	(0.0236)
No previous firm	0.28	0.24	-0.4104***	(0.0428)
Most common occupation:				
Manager	0.04	0.06	-0.1260**	(0.0571)
Higher education	0.04	0.04	-0.0294	(0.0572)
Clerk	0.03	0.03	0.0633	(0.0579)
Service, care	0.10	0.17	0.0396	(0.0554)
Building, manufacturing	0.04	0.03	-0.0574	(0.0574)
Mechanical, transport	0.11	0.06	0.0581	(0.0554)
Elementary occupation	0.02	0.02	-0.0817	(0.0602)
Industry:				
Agriculture, fishing, mining	0.01	0.01	-0.0906**	(0.0406)
Manufacturing	0.17	0.10	0.2257***	(0.0253)
Construction	0.05	0.06	-0.2065***	(0.0292)
Trade, repair	0.06	0.07	-0.1552***	(0.0270)
Accommodation	0.02	0.03	-0.2239***	(0.0336)
Transport, storage	0.06	0.04	0.1663***	(0.0278)
Financial, real estate	0.08	0.08	-0.0127	(0.0265)
Human health, social work	0.06	0.12	-0.1581***	(0.0298)
Other - public sector	0.04	0.08	-0.2254***	(0.0308)
Other	0.06	0.07	-0.1207***	(0.0277)
Panel L: Unemployment insurance				
UI: Daily benefit level in SEK	384.11	277.33	0.2316***	(0.0118)
UI: Eligible	0.84	0.83	-0.0134	(0.0136)
UI: No benefit claim	0.37	0.54	0.2181***	(0.0238)
UI 1 year before	12712.71	13211.32	-0.0086	(0.0054)
UI 2 years before	12779.13	13181.89	0.0056	(0.0059)
Cumulated UI 5 years before	62624.69	63758.25	-0.0929***	(0.0075)
Panel M: Parents' previous income				
Mother's past income (age 35-55)	659.10	772.63	-0.0061	(0.0052)
Father's past income (age 35-55)	856.04	1039.85	-0.0505***	(0.0055)
Missing mother's past income	0.39	0.34	0.0185	(0.0138)
Missing father's past income	0.47	0.42	-0.0517***	(0.0137)
Panel N: Duration dependence				
Baseline hazard, part 2			0.2653***	(0.0186)
Baseline hazard, part 3			0.5528***	(0.0161)
Baseline hazard, part 4			0.6408***	(0.0169)
Baseline hazard, part 5			0.6466***	(0.0178)
Baseline hazard, part 6			0.6843***	(0.0166)
Baseline hazard, part 7			0.5186***	(0.0171)
Baseline hazard, part 8			-0.0601***	(0.0162)

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: Estimated bias of the treatment effect when controlling for different blocks of covariates

	Est.	SE
Panel A: Baseline		
Baseline socio-economic characteristics	0.0693***	(0.00241)
Calendar time (inflow dummies)	0.1107***	(0.00239)
Region dummies	0.0912***	(0.00240)
Local unemployment rate	0.1174***	(0.00239)
All the above	0.0616***	(0.00243)
Panel B: Baseline and:		
Employment history (last 2 years) and duration	-0.0144***	(0.00244)
Unemployment history (last 2 years) and duration	0.0503***	(0.00243)
Earnings history (last 2 years)	0.0401***	(0.00243)
Welfare benefit history (last 2 years)	0.0469***	(0.00243)
All of the above	-0.0228***	(0.00244)
Panel C: Baseline, short-term history and:		
Employment history (last 10 years)	-0.0239***	(0.00244)
Unemployment history (last 10 years)	-0.0289***	(0.00244)
Welfare benefit history (10 years)	-0.0190***	(0.00244)
All of the above	-0.0241***	(0.00244)
Panel D: Baseline, short-term history, long-term history and:		
Last wage	-0.0266***	(0.00244)
Last occupation dummies	-0.0246***	(0.00244)
Firm characteristics (last job)	-0.0228***	(0.00245)
Unemployment benefits	0.0153***	(0.00244)
Parents income	-0.0231***	(0.00244)
All of the above	0.0090***	(0.00246)

Notes: Estimated biases using the full sample of placebo treated and non-treated with control for for different blocks of covariates. The number of observations is 2,564,561. 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: Estimated bias of the treatment effect when controlling for different short-term labor market history variables

	Est.	SE
Baseline	0.0616***	(0.00243)
Panel A: Employment duration		
Time employed in last spell	0.0394***	(0.00243)
Panel B: Short-term employment rates (2 years)		
Months employed in last 6 months	0.0168***	(0.00243)
Months employed in last 24 months	0.0091***	(0.00243)
No employment in last 24 months	0.0121***	(0.00243)
All variables	-0.0004	(0.00244)
Panel C: Other short-term employment history (2 years)		
Employed 1 year before	0.0160***	(0.00243)
Employed 2 years before	0.0265***	(0.00243)
Time since last employment if in last 24 months	0.0598***	(0.00243)
Number of employers in last 24 months	0.0427***	(0.00243)
All variables	0.0022	(0.00243)
Panel D: Unemployment duration		
Time unemployed in last spell	0.0547***	(0.00243)
Panel E: Short-term unemployment rates (2 years)		
Days unemployed in last 6 months	0.0632***	(0.00243)
Days unemployed in last 24 months	0.0616***	(0.00243)
No unemployment in last 24 months	0.0611***	(0.00243)
All variables	0.0564***	(0.00243)
Panel F: Other short-term unemployment history (2 years)		
Days since last unemployment if in last 24 months	0.0616***	(0.00243)
Number of unemployment spells in last 24 months	0.0560***	(0.00243)
Unemployed 6 months before	0.0632***	(0.00243)
Unemployed 24 months before	0.0590***	(0.00243)
Any program in last 24 months	0.0618***	(0.00243)
All variables	0.0539***	(0.00243)

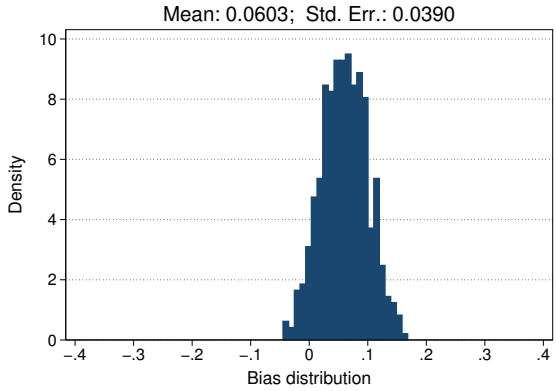
Notes: All models also include the baseline covariates (socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate). Estimated biases using the full sample of placebo treated and non-treated with control for for different blocks of covariates. The number of observations is 2,564,561. 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 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

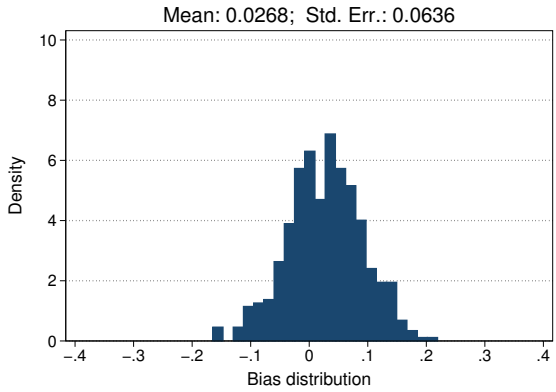
	Sample size								
	10,000			40,000			160,000		
	Bias (1)	SE (2)	MSE (3)	Bias (4)	SE (5)	MSE (6)	Bias (7)	SE (8)	MSE (9)
Panel A: Number of pre-specified support points									
1	0.060	(0.039)	0.0052	0.057	(0.020)	0.0037	0.058	(0.009)	0.0034
2	0.027	(0.064)	0.0048	0.022	(0.031)	0.0014	0.023	(0.014)	0.0007
3	0.046	(0.089)	0.0101	0.030	(0.042)	0.0026	0.028	(0.019)	0.0011
4	0.057	(0.098)	0.0128	0.035	(0.043)	0.0031	0.032	(0.021)	0.0015
5	0.062	(0.097)	0.0133	0.037	(0.044)	0.0033	0.033	(0.021)	0.0015
6	0.064	(0.099)	0.0138	0.037	(0.044)	0.0033	0.033	(0.021)	0.0015
Panel B: Model selection criteria									
ML	0.064	(0.099)	0.0139	0.037	(0.044)	0.0033	0.033	(0.021)	0.0015
AIC	0.032	(0.076)	0.0068	0.024	(0.036)	0.0018	0.026	(0.018)	0.0010
BIC	0.027	(0.064)	0.0048	0.022	(0.031)	0.0014	0.023	(0.014)	0.0007
HQIC	0.027	(0.064)	0.0048	0.022	(0.031)	0.0014	0.023	(0.014)	0.0007
Panel C: Average # support points, by selection criteria									
ML		4.11			3.99			4.10	
AIC		2.14			2.21			2.53	
BIC		1.99			2.00			2.00	
HQIC		2.01			2.00			2.04	

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.

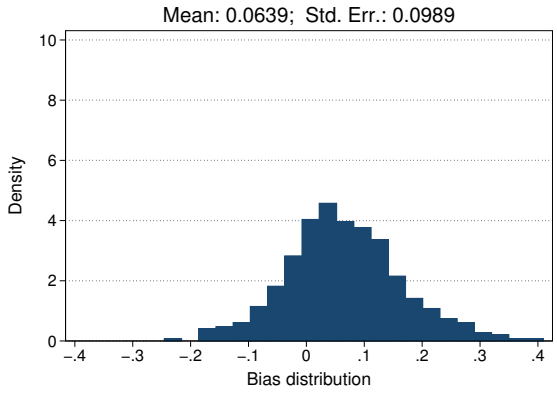
Figure 1: Distribution of the bias of the estimated treatment effect for a pre-specified number of support points, by number of support points



(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 10,000 random drawings from the full sample of 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 when *excluding different sets of covariates*, by model selection criteria and sample size

	Exclude more covariates			Exclude fewer covariates		
	Bias (1)	SE (2)	MSE (3)	Bias (4)	SE (5)	MSE (6)
Panel A: 10,000 observations						
ML	0.091	(0.162)	0.0344	0.073	(0.122)	0.0201
AIC	0.029	(0.010)	0.0108	0.035	(0.114)	0.0142
BIC	0.024	(0.067)	0.0051	0.005	(0.063)	0.0039
HQIC	0.024	(0.068)	0.0052	0.013	(0.091)	0.0085
Average # support points, by selection criteria						
ML		4.78			5.20	
AIC		2.34			3.12	
BIC		2.00			2.20	
HQIC		2.01			2.62	
Panel B: 40,000 observations						
ML	0.025	(0.068)	0.0053	0.049	(0.060)	0.0060
AIC	0.009	(0.049)	0.0025	0.029	(0.062)	0.0047
BIC	0.019	(0.034)	0.0015	0.005	(0.039)	0.0016
HQIC	0.018	(0.036)	0.0016	0.010	(0.050)	0.0026
Average # support points, by selection criteria						
ML		4.88			5.59	
AIC		2.65			4.22	
BIC		2.00			3.16	
HQIC		2.04			3.62	

Notes: The "exclude more covariates" model excludes baseline socio-economic characteristics and the "exclude fewer covariates" adds control for short-term earnings history from the baseline model which includes baseline socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate. 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).

Table 6: Bias and variance of the estimated treatment effect when augmenting the baseline model with *covariates more or less correlated* with those left in the error term

Degree of correlation	Positive			Small positive			Negative		
	Bias (1)	SE (2)	MSE (3)	Bias (4)	SE (5)	MSE (6)	Bias (7)	SE (8)	MSE (9)
Correlation		0.278			0.049			-0.257	
Panel A: 10,000 observations									
ML	0.063	(0.093)	0.0127	0.063	(0.100)	0.0140	0.044	(0.099)	0.0119
AIC	0.035	(0.076)	0.0070	0.033	(0.087)	0.0087	0.021	(0.081)	0.0070
BIC	0.027	(0.060)	0.0043	0.028	(0.070)	0.0057	0.019	(0.065)	0.0046
HQIC	0.027	(0.060)	0.0043	0.029	(0.071)	0.0059	0.017	(0.066)	0.0046
Average # support points, by selection criteria									
ML		4.19			4.48			4.27	
AIC		2.17			2.28			2.20	
BIC		2.00			1.99			1.95	
HQIC		2.01			2.01			2.01	
Panel B: 40,000 observations									
ML	0.042	(0.041)	0.0034	0.036	(0.047)	0.0035	0.019	(0.046)	0.0025
AIC	0.025	(0.036)	0.0019	0.025	(0.045)	0.0026	0.011	(0.039)	0.0016
BIC	0.022	(0.029)	0.0013	0.024	(0.034)	0.0018	0.013	(0.032)	0.0012
HQIC	0.022	(0.030)	0.0014	0.024	(0.035)	0.0018	0.013	(0.032)	0.0012
Average # support points, by selection criteria									
ML		3.99			4.62			4.34	
AIC		2.24			2.62			2.28	
BIC		2.00			2.00			2.00	
HQIC		2.01			2.04			2.01	

Notes: The three model specifications correspond to the baseline model of Table 4 augmented with Welfare benefit history (last 2 years), Previous firm most common occupation dummies and Last occupation dummies, for the positive correlation, small positive correlation and negative correlation specifications, respectively. Correlation coefficients computed from the outcome model using all actual treated and control units, by correlating the linear predictor of the covariates included in the model with the linear predictor of all covariates left in the error term. 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 set as for Table 4.

Table 7: Comparison between the actual and the estimated distribution of the unobserved heterogeneity for the treatment duration

	Mean $\exp(V_p)$	SE $\exp(V_p)$
Panel A: Actual distribution		
	0.00056	0.00023
Panel B: Estimated using a fixed number of support points		
2	0.00047	0.00003
3	0.00047	0.00020
4	0.00046	0.00023
5	0.00047	0.00027
6	0.00047	0.00031
Panel C: Estimated using section criteria		
ML	0.00047	0.00030
AIC	0.00047	0.00003
BIC	0.00047	0.00010
HQIC	0.00047	0.00003

Notes: Mean and standard error of the actual and the estimated distribution of the unobserved heterogeneity for the treatment duration. The actual distribution is based on linear predictor of the covariates left in the error term. The estimated distribution is based on the estimated discrete distributions from the ToE models (averaged across 500 replications, each with a sample of 10,000 units). Both the actual and approximated unobserved heterogeneity distributions include the constant. The ToE model includes baseline socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate.

Table 8: Bias and variance of the estimated treatment effect with *time-varying local unemployment rate*, by model selection criteria and sample size

Specification	Time-varying unemployment rate		
	Bias (1)	SE (2)	MSE (3)
Panel A: 10,000 observations			
Number of pre-specified support points			
1	0.056	(0.039)	0.0046
2	0.016	(0.066)	0.0046
3	0.056	(0.100)	0.0132
4	0.074	(0.109)	0.0174
5	0.082	(0.108)	0.0185
6	0.084	(0.109)	0.0189
Model selection criteria			
ML	0.084	(0.109)	0.0189
AIC	0.033	(0.090)	0.0093
BIC	0.016	(0.066)	0.0046
HQIC	0.017	(0.069)	0.0051
Average # support points, by selection criteria			
ML		4.46	
AIC		2.25	
BIC		1.99	
HQIC		2.01	
Panel B: 40,000 observations			
Number of pre-specified support points			
1	0.053	(0.020)	0.0032
2	0.010	(0.032)	0.0012
3	0.036	(0.053)	0.0040
4	0.052	(0.055)	0.0057
5	0.056	(0.053)	0.0060
6	0.057	(0.053)	0.0060
Model selection criteria			
ML	0.057	(0.053)	0.0060
AIC	0.026	(0.050)	0.0032
BIC	0.010	(0.032)	0.0012
HQIC	0.011	(0.035)	0.0014
Average # support points, by selection criteria			
ML		4.69	
AIC		2.40	
BIC		2.00	
HQIC		2.01	

Notes: Simulations with 10,000 observations. 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 ToE model also includes baseline socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate.