Econometric Reviews

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/lecr20

Empirical Monte Carlo evidence on estimation of timing-of-events models

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

To cite this article: Stefano Lombardi, Gerard J. van den Berg & Johan Vikström (13 Sep 2024): Empirical Monte Carlo evidence on estimation of timing-of-events models, Econometric Reviews, DOI: 10.1080/07474938.2024.2390399

To link to this article: https://doi.org/10.1080/07474938.2024.2390399

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC

Published online: 13 Sep 2024.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at https://www.tandfonline.com/action/journalInformation?journalCode=lecr20

ECONOMETRIC REVIEWS
2024, VOL. 00, NO. 0, 1–29
https://doi.org/10.1080/07474938.2024.2390399

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Empirical Monte Carlo evidence on estimation of timing-of-events models

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

[a]VATT Institute for Economic Research Helsinki, IZA, IFAU and UCLS at Uppsala University, Helsinki, Finland; [b]University of Groningen, University Medical Center Groningen, IFAU, IZA, ZEW, CEPR and J–PAL, Groningen, Netherlands; [c]IFAU and Uppsala University, Uppsala, Sweden

ABSTRACT

This article builds on the Empirical Monte Carlo simulation approach 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 dynamic selection models for training participation. 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. A comparison with other duration models shows that a Stratified Cox model performs well with abundant multiple spells but less well when multiple spells are uncommon. The standard Cox regression model performs poorly in all configurations as it is unable to account for unobserved heterogeneity.

ARTICLE HISTORY
4 May 2024
15 July 2024

KEYWORDS
Duration analysis; employment; matching; propensity score; training; unemployment

MATHEMATICS SUBJECT CLASSIFICATION 2020:
C14, C15, C41, J64

1. Introduction

In many empirical applications, researchers are interested in identifying the effect of a treatment given during a spell in a state of interest on the rate of leaving that state. Whenever systematic unobserved confounders cannot be ruled out, a leading approach in this setting is the Timing-of-Events (ToE) model developed by Abbring and van den Berg (2003), who specify a bivariate Mixed Proportional Hazard (MPH) model and establish conditions under which all its components, including the treatment effect, are non parametrically identified. Due to its flexibility in allowing for unobserved confounders, the ToE approach has been applied in many different settings.[1]

CONTACT Gerard van den Berg gerard.van.den.berg@rug.nl University of Groningen, Netherlands.

[1]An early example is Abbring, van den Berg, and van Ours (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). Examples include Crépon et al. (2018), Richardson and van den Berg (2013), Holm et al. (2017), Bergemann, Pohlan, and Uhlendorff (2017) on labor market policies; van Ours and Williams (2009) on cannabis use; van den Berg and Gupta (2015), Lindeboom, Llena-Nozal, and van der Klaauw (2016) on health settings; Bijwaard, Schluter, and Wahba (2014) on migration; Jahn and Rosholm (2013) on temporary work; and Baert, Cockx, and Verhaest (2013) on overeducation.

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Despite its widespread use in the treatment evaluation literature, the ToE approach relies on taking several implementation decisions that if not properly addressed can lead to severe bias in the estimated model parameters. A primary decision concerns the unknown bivariate unobserved heterogeneity distribution, which in the literature is often approximated by way of a discrete distribution (Heckman and Singer, 1984; Lindsay, 1983). When estimating the model, this approximation can be implemented in several ways. One approach is to pre-specify a (relatively low) number of support points and increase this number until the numerical estimation routine indicates that the support points converge or their associated probabilities vanish, or until computational problems arise. Alternatively, information criteria can be used to select the number of support points. Sample size may also be a relevant factor, since the estimation of non linear MPH models with many parameters may be problematic when the sample size is small. In addition, time-varying covariates may make results less dependent on functional-form assumptions (van den Berg, 2001).

In this article, we use a new simulation design based on real 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) originally proposed by Huber, Lechner, and Wunsch (2013) to compare different methods for estimating treatment effects under unconfoundedness.[2] The key idea is to use actual data on treated units to simulate placebo treatments for non treated units. This ensures that the true simulated treatment effect is zero, that the selection model is known, and that the unconfoundedness assumption holds by construction. The relevance of this simulation design for real-life applications relies on the fact that the simulations are based on real data rather than on an arbitrarily chosen data generating process.

All previous EMC implementations have examined estimators based on conditional independence assumptions. We implement a variant of the original EMC approach that enables us to study the estimation of the ToE model, which is cast in a duration framework. In our simulation design, we take advantage of rich administrative data on Swedish job seekers, with detailed information on participation in a training program (the treatment). We start by using data on actual treated and non treated units to estimate an auxiliary duration model for the duration until treatment under the assumption that all systematic determinants of the treatment assignment are captured by a comprehensive set of covariates. Next, we simulate placebo treatment dates for each non treated unit using this estimated selection model. By construction, the effect of the placebo treatments is zero and the treatment assignment process is known. With the simulated data we then estimate alternative ToE models by omitting subsets of the variables that were previously used to simulate the placebo treatment dates. Since the excluded variables are 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, that is, the ToE setting.

Our simulations lead to several conclusions. When omitting a large number of variables from the model without controlling for unobserved heterogeneity, the estimated placebo treatment effect is far from the true zero effect, that is, the estimated treatment effect is characterized by substantial bias. However, two support points for the unobserved heterogeneity term are already able to eliminate a large share of the bias. We also find a risk of over-correcting for the unobserved heterogeneity: when using too many support points, the average bias is more than twice as large as when using few support points, and the variance of the estimated treatment effect increases in the number of support points.

We further find that information criteria are useful for selecting an appropriate number of support points of the discrete bivariate unobserved heterogeneity distribution of the ToE model. 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

[2] Other studies using the EMC simulation design include Huber, Lechner, and Mellace (2016) on the performance of parametric and semi-parametric estimators used in mediation analysis: Frölich, Huber, and Wiesenfarth (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. (2020) consider inference methods for matching and weighting methods.

abundance and guard against under-correction by rejecting models with few or no correction for the 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, leading to over-correction problems. We additionally show that the use of time-varying covariates (such as the unemployment rate in the local labor market measured at monthly intervals) helps reducing the bias.

We also provide additional insights by comparing the ToE model with other commonly used duration models. A frequently used estimation approach is the standard Cox proportional hazard regression model, estimated by using partial maximum likelihood. We confirm that this model is unable to adjust for the correlated unobserved heterogeneity that we generate with our empirical Monte Carlo design. A common approach to adjust for unobserved heterogeneity is to exploit information on individuals with multiple spells and conduct a stratified analysis. Such stratified analysis relies on the assumption that the unobserved heterogeneity is constant across individual's spells. Common violations of this assumption are that the unobserved characteristics may change over time and the outcome (exit and treatment) of previous spells may affect future spells. We show that such a Stratified Cox model performs well for data with frequent multiple spells, but less well when multiple spells are uncommon.

As mentioned, when adopting the ToE approach, a central question concerns how to specify the unobserved heterogeneity distribution. In the literature, initial simulation evidence for MPH models was provided by Heckman and Singer (1984), Ridder (1987), Huh and Sickles (1994), and Baker and Melino (2000). Gaure, Røed, and Zhang (2007) examine a discrete-time bivariate duration model and analyze if the use of a discrete unobserved heterogeneity distribution is able to uncover the treatment effect of interest. They find that the discrete support-points approach is generally reliable if the sample is large and time-varying covariates are used. Moreover, they find that pre-specifying a low number of support points for the unobserved heterogeneity or deviations from other model assumptions may lead to substantial bias of the treatment effect. We contribute to this literature by using a simulation design that is based on real data rather than on artificial simulations. This allows us to assess whether ToE models are able to identify treatment effects by using data similar to those used in typical applications. In addition, we exclude different blocks of covariates and study how the ToE approach performs with different types of unobserved heterogeneity. We also examine how the model performs under different degrees of correlation between the observed covariates and the unobserved heterogeneity.

Using our simulation design, we also examine how the choice of covariates affects the estimated bias. This relates to a literature that uses experimental data to examine the relevance of different sets of covariates (Dehejia and Wahba, 1999, 2002; Smith and Todd, 2005).[3] It also relates to studies that use rich survey data to examine characteristics that are often not recorded in administrative data[4], and to Lechner and Wunsch (2013), who implement an EMC approach with German administrative data to examine the relative importance of adjusting for different types of variables. Using Swedish data, we examine whether the results in Lechner and Wunsch (2013) carry over to other countries and programs, while also analyzing the relevance of additional covariates not considered by the authors. Specifically, since we model treatment durations, and since previous durations may capture aspects related to how long one stays unemployed in the current spell in a more natural way than non duration history variables, we include previous employment and unemployment durations in our set of covariates. We also use information on parental income to proxy for more general skills, and we examine the importance of time-varying covariates such as local business cycle conditions. One of our findings is that short-term labor market history variables are particularly important to adjust for, and that adjusting for employment

[3] 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, Troske, and Gorislavsky (2007) highlight the importance of socio-demographic characteristics and pre-treatment outcomes.

[4] For example, Caliendo, Mahlstedt, and Mitnik (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.

history is relatively more important than adjusting for unemployment, earnings, and out-of-labor-force history. We also find that, once controlling for short-term labor market history variables, further adjusting for long-term labor market histories becomes redundant.

Finally, it is useful to discuss our approach in light of Advani, Kitagawa, and Słoczyński (2019), who point out some limitations of the original EMC approaches that were developed to compare different estimators based on the unconfoundedness assumption. Notably, the authors show that rather modest model misspecifications 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 article, we do not allow for deviations of the proportionality assumptions in the MPH specifications, and we simulate treatment durations and estimate ToE models assuming that the ToE functional form is correct. The proportionality assumption is in line with the vast empirical literature based on the ToE approach in the past decades (see, e.g., the references above); maintaining it throughout our analysis has the benefit of keeping the simulation exercises computationally tractable, while allowing us to focus on other specification issues that, as we argued before, are key when adopting the ToE approach.

We also acknowledge that there are interesting topics for future research that are beyond the scope of this article. One direction would be to apply the generative adversarial networks approach in Athey et al. (2024) to conduct further simulations related to the ToE approach. The critique raised by Advani, Kitagawa, and Słoczyński (2019) 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, Kitagawa, and Słoczyński (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 (our data contains 2.6 million unemployment spells) and our likelihood-based inference already requires Swedish national supercomputing resources. Therefore, in our view the application of bootstrap procedures would be beyond the scope of this article.

The article 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. In Section 4, we describe 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. Sections 5 and 6 present the EMC simulation results for the ToE model and for other duration models, respectively. Section 7 concludes.

2. The timing-of-events model

This section presents the ToE approach introduced by Abbring and van den Berg (2003). The authors 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).[5] 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

[5] In the simulation, we will also exploit time-varying covariates, but for presentation reasons this is suppressed in the notation below.

[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.

in the exit duration and the treatment duration, respectively. Also, let G denote the joint distribution of $V_e, V_p|x$ in the inflow into unemployment.

Abbring and van den Berg (2003) show that all components of this model, including the treatment effect, δ, and the unobserved heterogeneity distribution, G, are identified under the following assumptions. The first assumption is no-anticipation, which requires that future treatments do not 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. 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. A second assumption is that X and V_e, V_p are independently distributed, implying that the observed characteristics are uncorrelated with the unobserved ones. A third assumption is the proportional hazard structure (MPH model). We discuss these assumptions in more detail when we describe our simulation design. We refer to Abbring and van den Berg (2003) for further details and for additional regularity conditions that are required to identify the model parameters.

The ToE model is semi-parametric, in the sense that given the MPH structure, the model does not rely on any other parametric assumptions, but it 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. 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.

3. Simulation approach

3.1. The basic idea

The idea behind EMC designs is to produce simulations 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 originally proposed by Huber, Lechner, and Wunsch (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 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, even if the simulations are based on real data. By construction, the true effect of the placebo treatments is zero. Then, the authors use the resulting simulated data to analyze the performance of various CIA-based estimators.

We modify this simulation design in some key dimensions in order to use the EMC approach to study the ToE model and other commonly used duration models. We use rich Swedish administrative register data and survey data of jobseekers, with information on participation in a major labor market training program.[7] 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

[7] 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.

information for each unit. We use this data to generate placebo treatments, but instead of simulating binary treatment indicators as Huber, Lechner, and Wunsch (2013) do, we estimate a hazard model for the treatment duration and use the estimated selection model to simulate placebo treatment durations at the daily level. As in 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. Since the excluded variables affect both the time in unemployment (the outcome) and, by construction, the treatment duration, the data is simulated according to a bivariate duration model with correlated unobserved determinants.

The simulated data is used for various simulation exercises. We mainly examine the ToE model and study specification of the unobserved heterogeneity with and without information criteria, we let the sample size vary, and assess the relevance of using time-varying covariates. We further explore the importance of excluding different types of covariates, the correlation between the observed and unobserved variables in the model, and model misspecifications when the simulations are based on a non multiplicative baseline hazard. We mainly focus on the estimation of the treatment effect (bias and variance), but we also study whether the model is able to recover the true unobseved heterogenity distribution of the treatment process.

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 the simulation design. In particular, the 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 used in the ToE estimation.[8] To explore this possibility, we leave out blocks of variables that are either highly or mildly correlated with the observable characteristics. It turns out that the degree of correlation between the observed and unobserved factors is relatively unimportant. We also explore other potential misspecifications that may be important in practice. Using our simulation design we explore the consequences of ignoring interaction terms between the variables in the model and consequences of using a ToE model when the true baseline hazard is non multiplicative. Last, in order to model the treatment selection process we use a duration model without embedded unobserved heterogeneity. This means that although to estimate the selection process we use an extremely rich set of variables that mimics the information available to caseworkers when assigning treatments, the model may be misspecified if there are omitted characteristics.

3.2. The relevance of different covariates

The analysis of the ToE model specification is the main contribution of our article. However, by leaving out different blocks of covariates, we can also evaluate the relevance of different observable characteristics 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 observable characteristics and compare the magnitude of the bias across the resulting specifications.

This analysis benefits from the availability of rich Swedish data. We first follow Lechner and Wunsch (2013), who use German data to create variables that have been shown to be important for the selection process and have been used in various CIA-based evaluations of active labor market programs. We use Swedish databases to re-construct covariates similar to those in Lechner and Wunsch (2013), but we also include additional ones not used by the authors. 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

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

related to how long one stays unemployed in a better way than non duration history variables. 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 address this, we use parental income. In the literature on the determinants and returns to education, parental income has been related to background family skills and traits and has shown to affect the investment in education (see e.g. Chevalier et al. (2013) and Jensen, Lindemann, and Weiss (2023)). Parental income has not been used in studies on unemployment durations because, in general, employment register data do not include such a variable (see Mazzotta (2010) and Farace, Mazzotta, and Parisi (2014) for rare examples of studies relating parental income to offspring unemployment durations). However, it is conceivable that parental income can provide an informal insurance mechanism against low income in unemployment and hence may affect participation in training and returns to training. Third, since we model treatment durations, certain time-varying covariates, such as business cycle conditions, may be important, especially for longer unemployment spells. Last, another difference compared Lechner and Wunsch (2013) is that we consider a duration outcome framework.[9]

3.3. Comparison of different duration models

Our analysis is centered around the ToE model specification, but we also compare the ToE performance with those of two other approaches for inference on duration outcome variables. The first comparison is made with a Cox proportional hazard model estimated using partial maximum likelihood (Cox, 1972). The model assumes that there is no unobserved heterogeneity in the hazard rate. Since this assumption is likely violated in empirical applications and does not hold by construction in our simulation design, the treatment effect and other model parameters might be severely biased (van den Berg, 2001). At the same time, the approach is flexible and computationally simple and can be seen as providing descriptive evidence of relations between variables in the data.

A second comparison is made with a Stratified Cox approach, involving the semiparametric stratified partial-likelihood estimation of a generalization of the Cox model. The model allows for unobserved fixed effects and for full interactions between unobserved heterogeneity and duration dependence in the hazard rate. The approach takes advantage of the fact that for a subset of individuals in our sample we have multiple spells, allowing us to *stratify* at the individual level. If the individual unobserved heterogeneity is constant across spells, this procedure should adjust for any unobserved heterogeneity even if it is correlated with the treatment. This offers a tractable way to estimate the treatment effect. However, the assumption of constant unobserved heterogeneity may be problematic for several reasons. Unobserved characteristics may change over time, in particular if we consider spells several years apart. Moreover, the outcome (exit to job and treatment) of previous spells may affect future spells, therefore creating non constant unobserved heterogeneity. A stratified Cox analysis also discards information for individuals with only one spell. Together, this suggests that the type of data at hand may affect how well a stratified analysis compares to a ToE analysis.

In Section 6, we compare the three approaches across different sampling designs.

3.4. The training program

One often-studied treatment that job seekers are assigned to is labor market training. This motivates our use of data on the Swedish vocational training program AMU (Arbetsmarknadsutbildning). The program and the type of administrative data that we use resemble those of other countries. The main

[9]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.

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, Forslund, and Liljeberg (2008), Richardson and van den Berg (2013), and van den Berg and Vikström (2022). These studies also describe the training program in detail.

3.5. 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 covariates similar to those in Lechner and Wunsch (2013).[10] 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). Unemployment Insurance (UI) records provide information on UI eligibility. The data from the Public Employment Service is also used to construct unemployment history variables and information on the regional unemployment rate. Earnings records and data on welfare participation are used to construct employment, out-of-labor force and earnings histories. We construct both short-run history (last two years) and more long-run history (last ten years). All these characteristics should capture key aspects of the workers employment and earnings history.

We also compute previous unemployment and employment durations, the idea being that previous durations may capture the duration in the current spell in a better way than the above-mentioned employment history variables. We measure 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, using the Swedish multi-generational register (linking children to parents) and income registers for the parents. Finally, we explore time-varying covariates through the monthly local unemployment rate in the region (Sweden has 21 regions).

The outcome considered in this article 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.

[10]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.

All covariates that are used in the analyses are summarized in Table A1 in Appendix. The statistics in the table show that immigrants from outside Europe, males, married and 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.

3.6. Simulation details

3.6.1. 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:[11]

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

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 A1 in Appendix. The model estimates in the same table 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 the 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 Eq. (2) to simulate the placebo times to treatment for each non treated, T_p, 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 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.[12]

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, Lechner, and Wunsch (2013).

[11] 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 the model does not impose a proportional hazard structure it may not be compatible with any ToE model.

[12] The actual distribution for the integrated hazard will depend on the specification of the selection model in Eq. (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_p^0 for $t \in [0, t_1)$ and θ_p^1 for $t > t_1$:

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

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).

3.6.2. 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 Root-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.

3.7. Implementation of the bivariate duration model

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

3.7.1. 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. 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, and 548 days. In the section below, 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 Eq. (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.[13] 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_{em}, α_{pm}, x_i, and the two outcome indicators, y^e_{ij} and y^p_{ij} which equals one if the spell part j ends with a transition to employment or treatment, respectively, and zero otherwise.

Then, with approximately continuous durations, $L_i(v)$ is:

$$L_i(v) = \prod_{c_{ij} \in C_i} \Big[\exp\left(-l_{ij}\theta_e(t|x_i, D_{it}, v_e)\right) \times \theta_e(t|x_i, D_{it}, v_e)^{y^e_{ij}} \times \\ \times \exp\left(-l_{ij}\theta_p(t|x_i, v_p)\right) \times \theta_p(t|x_i, v_p)^{y^p_{ij}} \Big], \tag{4}$$

with

$$\theta_e(t|x_i, D_{it}, v_e) = \lambda_e(t)\ \exp(x_i'\beta_e)\ \exp(\delta D_{it})\ v_e$$
$$\theta_p(t|x_i, v_p) = \lambda_p(t)\ \exp(x_i'\beta_p)\ v_p$$

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:

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

[13] In this case, the vector of observables in Eq. (3.7) below can simply be specified as x_{it}.

3.7.2. Search algorithm

To estimate the discrete support points, we use the iterative search algorithm in Gaure, Røed, and Zhang (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, Ferrier, and Rogers, 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.[14] 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 procedurem we obtain $\overline{W}$ maximum likelihood estimates: $\{\hat{\vartheta}_W,\ \mathcal{L}(\hat{\vartheta}_W)\}_{W=1}^{\overline{W}}$.

3.7.3. 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.[15]

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

In this section, we assess the relevance of the individual different types of covariates in a specific way, by leaving out various blocks of covariates and by comparing the size of the bias – the difference between the estimated treatment effect and the true zero effect of the placebo treatments – across proportional hazard (PH) specifications for the exit rate out of unemployment. All covariates are a subset of those used to generate the placebo treatments. For each specification, the full sample of placebo treated and placebo non treated units is used to estimate a parametric PH model. Here, the baseline hazard is specified in the same way as for the model used to simulate the placebo treatments. The main results are given in Table 1. Below Panel A, each subsequent panel of the table starts with the covariates from the proceeding

[14] 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.

[15] We follow Gaure, Røed, and Zhang (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.

Table 1. Estimated bias of the treatment effect when controlling for different blocks of covariates.

	Est.	SE
Panel A: Baseline		
Baseline socioeconomic 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.

panels and then adds additional information, so that the model is extended sequentially by adding blocks of covariates one by one. This will, for instance, reveal the relevance of adding long-term labor market histories on top of short-term history and socio-economic characteristics. Table A1 in Appendix lists all covariates in the blocks.

Panel A of Table 1 starts with the "baseline" model with a set of baseline socio-economic characteristics, which returns a positive and sizable bias of 6.9%. That is, the estimated treatment effect is 0.069 when the true effect of these placebo treatments is equal to zero. Adding spatial and temporal dummies and the local unemployment rate slightly reduces this bias to 6.2%. 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. All these 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). This indicates 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 with the vocational training program.

We next consider individual variables and examine what specific aspects of previous employment and unemployment are the most important to adjust for. Table A2 in Appendix 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 each covariate measuring past employment singlehandedly captures a large part of the bias, so that adjusting for previous employment rates is relatively more important than adjusting for previous employment durations. 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 the past employment status.[16,17]

Next, let us return to Table 1. Here, Panel C shows that adding information on long-term labor market history (last ten years) on top of short-term history (last 2 years) has a 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, Troske, and Gorislavsky (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 with the previous literature is that we find that adjusting for short-term employment history suffices to obtain a 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 shows that 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. To the extent that parental income is able to proxy for general unobserved skills, this indicates that labor market histories are also able to capture more general unobserved skills.[18]

5. Specification of the ToE model

We now study aspects related to the specification of the ToE model. Our 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, and mean squared error (MSE) of the placebo estimates.

5.1. Baseline results

Table 2 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, regional indicators, and unemployment rate (the covariates in Panels A and B of Table A1). Here, we control for time-fixed regional unemployment rate (measured as the month of inflow into unemployment). Later, in Table 3, 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

[16]Panels D–F of Table A2 in Appendix 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. We also tried to additionally include past employment and unemployment durations more flexibly, by either specifying them on logarithmic- and quadratic-scale or by including information from the previous two spells. The bias is only slightly reduced compared to the information reported in Table A2 and all patterns are qualitatively unaffected.

[17]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, 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.

[18]This is consistent with the results in Caliendo, Mahlstedt, and Mitnik (2017), who finds that once controlling for rich observables of the type that we include here, additional (usually unobserved) characteristics measuring personality traits and preferences become redundant.

Table 2. Bias, standard error (SE), and MSE of the estimated treatment effect for a pre-specified number of support points, and average number of support points by model selection criterium.

	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 criterium									
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: Bootstrapped bias, standard error, and MSE (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.

without unobserved heterogeneity (one support point) leads to large bias (6.0%).[19] 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%.[20] For three or more support points, the average bias is even larger and keeps increasing in the same direction when adding additional support points. 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).

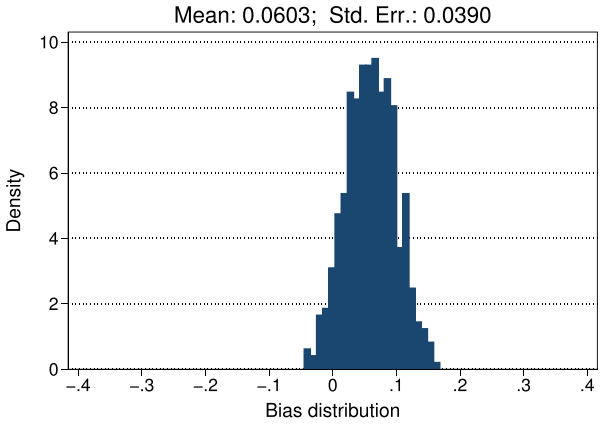
One explanation for the finding that the bias in the estimated treatment effect increases when using more than two support points is that specifications with many support points tend to over-correct for unobserved heterogeneity.[21] This pattern contradicts the intuition that one should adjust for unobserved heterogeneity in the most flexible way in order to avoid bias due to unaccounted unobserved heterogeneity.

To better understand this over-correction pattern, Fig. 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. Moreover,

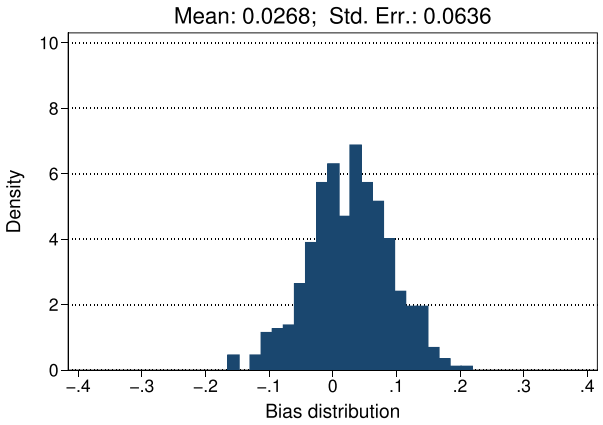
[19] This is roughly the same bias as in the corresponding model estimated with the full sample in Panel A of Table 1. The minor difference is due to sampling variation since here we report the average bias from random drawings, whereas estimates in Table 1 are obtained from the full set of placebo treated and non treated observations.

[20] 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, Røed, and Zhang, 2007).

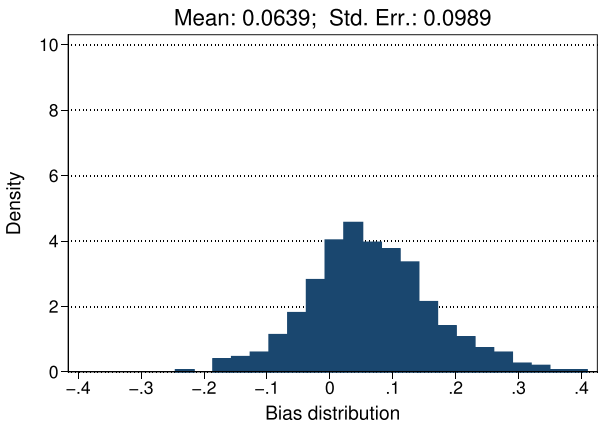
[21] There may be an analogy to estimation of models with non parametric components, where over-fitting prevents a bias in the model dimension that is fitted non parametrically but may have the unintended consequence that other model parameters are less precisely estimated.



(a) 1 support point



(b) 2 support points



(c) 6 support points

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

the entire distribution of the estimates shifts to the right (larger positive bias), which shows that the increased bias is not due to a few extreme estimates.

In sum, our simulation results suggest that both under- and over-correction are important issues 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.

5.2. Information criteria

Panel B of Table 2 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 the over-correction problem mentioned above. This pattern is confirmed in all simulation settings presented below. Therefore, from this first set of results we conclude that 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 information criteria are more encouraging. All three criteria select models with rather few unobserved heterogeneity 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 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 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. 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, in which case the more restrictive BIC and HQIC criteria are preferable. Thus, using all three criteria and reporting a range of estimates as a robustness check appears to be a reasonable approach.

In this section, our main interest is to provide information on the alternative specification choices. However, Table 2 also provides insights on the overall idea of using ToE models to adjust for unobserved heterogeneity. 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 2). This holds even though the only source of exogenous variation derives from time-fixed observed covariates.

5.3. Sample size

In Columns 4–6 and 7–9 of Table 2, the sample size is increased to 40,000 and 160,000 observations, respectively. The results with both sample sizes confirm that two support points are associated with the lowest bias. However, now 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%, whereas with 40,000 observations the bias increases from 2.2% to 3.7%. This shows that over-correction issues are mainly a problem with small sample sizes. This makes sense in the

light of the general principle that estimation imprecision due to overfitting becomes less of a problem when the sample size increases (provided that the number of parameters remains constant).

Note that what constitutes a small sample size in general differs across applications and relates 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. It is reassuring that with larger sample sizes the differences between the alternative information criteria are relatively small. 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

Next, we vary the unobserved heterogeneity by excluding various sets of covariates when estimating 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 socioeconomic characteristics in Panel A of Table A1) and less heterogeneity by excluding less covariates (earnings history in Panel F of Table A1). The resulting bias is 9.5% and 4.0%, respectively, which can be compared to the bias of 6.2% in the baseline setting.

Columns 1–3 of Table A3 in Appendix display results for the model with more extensive unobserved heterogeneity. As in the baseline setting, the ToE model is able to adjust for a large share of the bias induced by the unobserved heterogeneity. For instance, with a sample size of 10,000, the bias for the specification without unobserved heterogeneity (not displayed) is 9.4%, but it drops to 2–3% when we specify the mass points of the unobserved heterogeneity distribution according to the AIC, BIC, or HQIC criteria (Panel A). As before, these more restrictive criteria return the lowest bias, whereas the ML criterion selects a model with too many support points.

When using 40,000 observations, the bias is smaller for the AIC criterion than for BIC and HQIC criteria. However, when we create less substantial unobserved heterogeneity by excluding fewer covariates (Columns 4–6), the average bias is lower for the more restrictive BIC and HQIC criteria than for AIC. We conclude that none of the information criteria is superior in all settings.

5.5. Misspecifications of the model

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 taking out various different blocks of covariates from the model, where we compare settings with contrasting degrees of correlation between the covariates used in the ToE model and the excluded covariates. Specifically, we consider three settings: one with a strongly positive, one with a mildly positive and one with a negative correlation.[22] To facilitate the comparisons, 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 A4 in Appendix 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. The fact that this result holds regardless of the

[22]To compute the correlation, we use the parameter estimates of the selection model that includes all the covariates, as reported in Table A1 in Appendix. We take covariate values at the onset of the unemployment spell. For each spell, the parameter estimates and the excluded covariates can be used to calculate a single index. This linear predictor equals V_p in the simulations. We correlate this with a single index based on the included covariates used in the selection model. This produces a scalar that measures the degree of correlation between the observed and unobserved covariates in the model.

degree of correlation between the observed and the unobserved variables is reassuring: even when the variables taken out of the model are strongly related to those left in the ToE model, the relative performance of the information criteria is not affected. We obtain similar results for sample size of 40,000 (Panel B of Table A4).

An alternative approach to excluding different sets of covariates is to include interaction terms between covariates when estimating the true selection model and subsequently omit these interactions when we estimate the ToE model. This creates a different source of misspecification since the omitted interaction terms create another type of correlation between the covariates and the unobserved heterogeneity. We start from the baseline selection model, which includes all the covariates in Table A1. We then add interactions between the socio-economic characteristics and either the short-run employment history variables or the short-run unemployment history variables (see Panels A, C, and D of Table A1). This exercise generates two additional selection models which we use to simulate new placebo treatments. We then sample spells and estimate the ToE model as before. The results are given in Columns 5–12 of Table A5, which also reports the baseline results with no interaction in Columns 1–4. As before, there is no clear ranking between the AIC, BIC, and HQIC criteria, and all three perform better than a model without unobserved heterogeneity. Here, the ML approach performs similar to the three information criteria in terms of the bias but worse when looking at the MSE.

We also consider violations of the multiplicative hazard rate assumption embedded in the ToE model functional form. To this aim, we allow for separate baseline hazards for males and females in the true selection model, which creates non multiplicativity as the baseline hazard is now partly determined by the observed covariates in the model. By ignoring this non multiplicativity when estimating the ToE models, we examine this possible misspecification. Besides gender, we also consider separate baseline hazards by age reclassified into the categories of Table A1. The results in Table A6 are similar to those previously obtained when ignoring the interaction terms: we find no clear ranking of the information criteria but all three perform better than a model without unobserved heterogeneity.

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 also be assessed by inspecting to what extent the estimated discrete distributions for the unobserved heterogeneity approximates the true one. To examine 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. On the other hand, since we do not simulate the outcome durations, the exact composition of V_e is unknown.

For each actual treated and control unit, we use the coefficients of the estimated selection model reported in Table A1 in Appendix to compute the linear predictor of the variables left out from the model, corresponding to the V_p term. Then, in Table A7 in Appendix, we compare the first two moments of this true unobserved heterogeneity (Panel A) with the corresponding moments for the estimated unobserved heterogeneity from the ToE models (Panels B–C). Panel B shows that a larger number of support points tend to overestimate the dispersion of the unobserved heterogeneity, whereas the mean of the unobserved heterogeneity distribution tends to be slightly underestimated. Finally, Panel C indicates that the ML criterion returns an unobserved heterogeneity distribution with a variance that is too large when compared to the true one, whereas for the more restrictive information criteria (AIC, BIC, and HQIC) the variance is too small.[23]

[23] 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.

5.7. Time-varying covariates

We now use additional variation in the form of time-varying covariates (local unemployment rate). The idea is that time-varying covariates generate exogenous shifts in the hazard rates that help to recover the influence of the unobserved heterogeneity and separate it from the treatment effect. This is because current factors have an immediate impact on the exit rate, while 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).

The time-varying covariate that we use is time-varying unemployment rate measured at the monthly level for each region, referred to as the local unemployment rate. The same covariate was included in the selection model to simulate the placebo treatments. The results from this exercise are presented in Table 3. 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

Table 3. Bias, standard error (SE), and MSE of the estimated treatment effect with *time-varying local unemployment rate*, by model selection criterium 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, standard error and MSE (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.

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, supporting the idea of using time-varying covariates when estimating ToE models. Note that this result holds even though we 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, which produced substantial bias in the model without unobserved heterogeneity.

6. Alternative duration models

We now compare the approach for the ToE model to approaches based on two other commonly used duration models: the Cox PH model and the Stratified Cox model, as explained in Section 3.3 above. We use the data from our analyses above with placebo treated and non treated, where the placebo treatments are generated using the full set of covariates in Table A1. In each replication, we first select a sample of 10,000 spells (or 40,000). For the estimation of the ToE model and the Cox PH model, we directly use this sample of of 10,000 spells. For the Stratified Cox model, we must use multiple-spell data. We proceed as follows. First, for comparison reasons, we start with the same 10,000 spells. Out of the corresponding set of non repeated individuals, we select those with at least two spells (among the full set of simulated durations, not just in the sampled 10,000 spells). Since individuals with only one spell are discarded and the selected people have at least two but possibly more spells, the final sample size of each replication is in general different from 10,000. Overall, this mimics a realistic sampling scenario where researchers do not have access to the same number of spells when performing multiple-spell vs. single-spell analyses. We repeat this procedure in each replication and estimate the three models using the selected samples. For simplicity, we only present the AIC criterion results for the ToE model, but we obtain similar results when using the other criteria.

In Columns 1–4 of Table 4 we present results when using the same time window as in the main analysis (i.e., the full sampling period 2002–2011). As expected, the Cox PH model leads to biased estimates as it fails to adjust for the correlated unobserved heterogeneity created in our simulation setting. As before, the ToE model removes a substantial share of this bias. Also, as expected, the Stratified

Table 4. Treatment effect bias for different Cox model specifications.

	Spells sampled											
	All (2002-2011)				2006-2009				2006-2007			
	MP (1)	Bias (2)	SE (3)	MSE (4)	MP (5)	Bias (6)	SE (7)	MSE (8)	MP (9)	Bias (10)	SE (11)	MSE (12)
Panel A: 10,000 observations												
ToE (AIC)	2.72	0.022	0.065	0.0047	2.65	0.037	0.067	0.0059	2.74	0.014	0.073	0.0056
Cox model	–	0.118	0.035	0.0164	–	0.132	0.037	0.0201	–	0.097	0.039	0.0394
Stratified Cox	–	-0.032	0.062	0.0087	–	-0.023	0.107	0.0235	–	-0.038	0.159	0.0520
Panel B: 40,000 observations												
ToE (AIC)	2.94	0.025	0.031	0.0016	3.03	0.033	0.032	0.0021	3.19	0.009	0.035	0.0013
Cox model	–	0.120	0.018	0.0150	–	0.131	0.017	0.0176	–	0.097	0.019	0.0100
Stratified Cox	–	-0.031	0.032	0.0030	–	-0.021	0.048	0.0051	–	-0.045	0.081	0.0150

Notes: Estimated bias, standard error (SE), and mean squared error (MSE) of the placebo treatment effect on the hazard rate for time in unemployment. All model specifications include socio–economic characteristics, inflow year dummies, regional indicators, and local unemployment rate. The alternative runs use different sampling windows for estimating the ToE and Cox specifications: full set of spells (inflow years 2002–2011), medium time window (2006–2009), short time window (2006–2007). Simulations using 500 replications with random drawings from the full sample of placebo treated and placebo non treated. At each draw, both ToE and non Stratified Cox models use 10,000 spells (one per person). Stratified Cox models use the subset of spells of people that have at least 2 spells in the given sampling window.

Cox model reduces the bias. In particular, the magnitude of the bias is similar to that obtained when estimating the ToE model, but its sign is reversed. There may be several reasons for this result. In particular, if the unobserved heterogeneity is not constant across spells, this can drive the bias in different directions. For instance, unobserved characteristics may change over time and the outcome (exit and treatment) of previous spells may affect future spells.

We next consider different sampling windows. We first narrow the sampling from 2002–2011 to 2006–2009 and then to 2006–2007. These tighter sampling windows mean that the data will include relatively fewer individuals with multiple-spells, but also that the average time between the multiple spells will be shorter than for the full sampling window. This may affect how the stratified Cox analysis performs in relation to the other two estimation approaches. The results in Table 4 show that tightening the sampling windows has relatively little systematic impact on the ToE and Cox PH results. But tighter windows have more substantial impact on the Stratified Cox results, especially regarding the MSE: going from the full sampling period to the tightest sampling period increases the MSE by more than five times. An explanation for this result could be that a tighter sampling window implies less multiple spells. We conclude that a Stratified Cox approach is a viable alternative to account for unobserved heterogeneity if the data includes many individuals with multiple spells, especially since a stratified Cox analysis is easier and less computationally demanding than a ToE estimation with complex unobserved heterogeneity. However, a stratified Cox analysis is less useful if multiple spells are uncommon.

7. Conclusions

We modify a recently proposed simulation technique, the Empirical Monte Carlo approach, to evaluate the Timing-of-Events model for dynamic treatment evaluation with selective assignment. Our analyses provide several guidelines on how to specify and estimate ToE models in practice.

Information criteria are a reliable way to specify the number of support points to mimic the unobserved heterogeneity distribution in the model, provided that the criteria include a substantial penalty for parameter abundance. Information criteria with a small penalty for abundance, such as the ML criterion, should be avoided. Three criteria that perform well are the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the Hannan-Quinn information criterion (HQIC). We find that when specifying the unobserved heterogeneity distribution of the model, all three criteria protect both against the inclusion of spurious support points (which avoids over-correction problems) and against the inclusion of too few support points. None of these three criteria dominates the others across all settings.

Overall, we find that the ToE approach is able to adjust for substantial unobserved heterogeneity generated by omitting relevant and diverse covariates. As long as an appropriate information criterion is used, this finding is robust across alternative specifications. Moreover, adding time-varying covariates (such as the local unemployment rate) further improves the performance of the ToE estimator.

When comparing the ToE model with other commonly used duration models, we find that a standard Cox PH analysis, as expected, performs poorly in configurations characterized by correlated unobserved heterogeneity. On the other hand, the Stratified Cox model, which allows for unobserved heterogeneity by exploiting information from individuals who have multiple spells, performs well for data with frequent multiple spells but less well when multiple spells are uncommon.

We also examine which types of observable covariates are important confounders when evaluating labor market programs. We find that it is helpful to control for short-term labor market histories, whereas controlling for long-term labor market histories appears to be less important. Moreover, controlling for features of the short-term employment history appears to be more effective than controlling for features of the short-term unemployment history. One interesting topic for further research is to reconcile the use of of labor market history indicators with that of the ToE framework and its identifying assumptions.

A different topic for further research would be to develop formal data-driven methods for determining the optimal level of correction for unobserved heterogeneity. For example, it is conceivable that

cross-validation may determine the number of support points as a function of their predictive power. A potential obstacle for a formal statistical underpinning is that the number of support points is necessarily discrete.

Funding

We thank the Editor Esfandiar Maasoumi, an anonymous Referee, 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).

References

Abbring, J. H., Van Den Berg, G. J. (2003). The nonparametric identification of treatment effects in duration models. *Econometrica* 71(5):1491–1517. doi:10.1111/1468-0262.00456

Abbring, J. H., Van Den Berg, G. J., Ours, J. C. (2005). The effect of unemployment insurance sanctions on the transition rate from unemployment to employment. *The Economic Journal* 115(505):602–630. doi:10.1111/j.1468-0297.2005.01011.x

Advani, A., Kitagawa, T., Słoczyński, T. (2019). Mostly harmless simulations? using monte carlo studies for estimator selection. *Journal of Applied Econometrics* 34(6):893–910. doi:10.1002/jae.2724

Athey, S., Imbens, G. W., Metzger, J., Munro, E. (2024). Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics* 240(2):105076. doi:10.1016/j.jeconom.2020.09.013

Baert, S., Cockx, B., Verhaest, D. (2013). Overeducation at the start of the career: Stepping stone or trap?. *Labour Economics* 25:123–140. doi:10.1016/j.labeco.2013.04.013

Baker, M., Melino, A. (2000). Duration dependence and nonparametric heterogeneity: A monte carlo study. *Journal of Econometrics* 96(2):357–393. doi:10.1016/S0304-4076(99)00064-0

Bergemann, A., Pohlan, L., Uhlendorff, A. (2017). The impact of participation in job creation schemes in turbulent times. *Labour Economics* 47:182–201. doi:10.1016/j.labeco.2017.05.007

Bijwaard, G. E., Schluter, C., Wahba, J. (2014). The impact of labor market dynamics on the return migration of immigrants. *Review of Economics and Statistics* 96(3):483–494. doi:10.1162/REST_a_00389

Bodory, H., Camponovo, L., Huber, M., Lechner, M. (2020). The finite sample performance of inference methods for propensity score matching and weighting estimators. *Journal of Business & Economic Statistics* 38(1):183–200. doi:10.1080/07350015.2018.1476247

Caliendo, M., Mahlstedt, R., 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. doi:10.1016/j.labeco.2017.02.001

Chevalier, A., Harmon, C., O'Sullivan, V., Walker, I. (2013). The impact of parental income and education on the schooling of their children. *IZA Journal of Labor Economics*, 2(8)

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 34(2):187–202. doi:10.1111/j.2517-6161.1972.tb00899.x

Crépon, B., Ferracci, M., Jolivet, G., 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. doi:10.1002/jae.2621

De Luna, X., Forslund, A., 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)*, 1, IFAU working paper.

Dehejia, R. H., Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94(448):1053–1062. doi:10.1080/01621459.1999.10473858

Dehejia, R. H., Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics* 84(1):151–161. doi:10.1162/003465302317331982

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

Farace, S., Mazzotta, F., Parisi, L. (2014). *Characteristics of parents and the unemployment duration of offspring: Evidence from Italy*. In: Malo, M., Sciulli, D., eds., *Disadvantaged Workers*, Berlin, Germany: Springer, pp. 149–179.

Frölich, M., Huber, M., Wiesenfarth, M. (2017). The finite sample performance of semi- AND non-parametric estimators for treatment effects and policy evaluation. *Computational Statistics & Data Analysis* 115:91–102. doi:10.1016/j.csda.2017.05.007

Gaure, S., Røed, K., Zhang, T. (2007). Time and causality: A monte carlo assessment of the timing-of-events approach. *Journal of Econometrics* 141(2):1159–1195. doi:10.1016/j.jeconom.2007.01.015

Goffe, W. L., Ferrier, G. D., Rogers, J. (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics* 60(1-2):65–99. doi:10.1016/0304-4076(94)90038-8
Gørgens, TUE. (2006). Semiparametric estimation of single-index hazard functions without proportional hazards. *The Econometrics Journal* 9(1):1–22. doi:10.1111/j.1368-423X.2006.00174.x
Harkman, A., Johansson, A. (1999). *Training or subsidized jobs-what works? Working paper*, Solna: AMS.
Heckman, J., Ichimura, H., Smith, J., Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica* 66(5):1017. doi:10.2307/2999630
Heckman, J., Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 52(2):271. doi:10.2307/1911491
Heckman, J. J., 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. doi:10.1111/1468-0297.00451
Holm, A., Høgelund, JAN., Gørtz, M., Rasmussen, K. S., Houlberg, H. S. B. (2017). Employment effects of active labor market programs for sick-listed workers. *Journal of Health Economics* 52:33–44. doi:10.1016/j.jhealeco.2017.01.006 28182999
Huber, M., Lechner, M., Mellace, G. (2016). The finite sample performance of estimators for mediation analysis under sequential conditional independence. *Journal of Business & Economic Statistics* 34(1):139–160. doi:10.1080/07350015.2015.1017644
Huber, M., Lechner, M., Wunsch, C. (2013). The performance of estimators based on the propensity score. *Journal of Econometrics* 175(1):1–21. doi:10.1016/j.jeconom.2012.11.006
Huh, K., 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 doi:10.2307/2109770
Jahn, E., Rosholm, M. (2013). Is temporary agency employment a stepping stone for immigrants?. *Economics Letters* 118(1):225–228. doi:10.1016/j.econlet.2012.10.029
Jensen, S. S., Lindemann, K., Weiss, F. (2023). Parental job loss and the role of unemployment duration and income changes for children's education. *European Sociological Review* jcad068:1–17.
Lechner, M., Strittmatter, A. (2017). Practical procedures to deal with common support problems in matching estimation. *Econometric Reviews* 38(2):193–207.
Lechner, M., Wunsch, C. (2013). Sensitivity of matching-based program evaluations to the availability of control variables. *Labour Economics* 21:111–121. doi:10.1016/j.labeco.2013.01.004
Lindeboom, M., Llena-Nozal, ANA., Van Der Klaauw, BAS. (2016). Health shocks, disability and work. *Labour Economics* 43:186–200. doi:10.1016/j.labeco.2016.06.010
Lindsay, B. G. (1983). The geometry of mixture likelihoods: A general theory. *The Annals of Statistics*, 11(1):86–94.
Mazzotta, F. (2010). *The effect of parental background on youth duration of unemployment. CEnter for Labor and Political Economics workign paper no.* 113.
Mueser, P. R., Troske, K. R., Gorislavsky, A. (2007). Using state administrative data to measure program performance. *Review of Economics and Statistics* 89(4):761–783. doi:10.1162/rest.89.4.761
Richardson, K., 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. doi:10.1002/jae.2263
Ridder, G. (1987). The sensitivity of duration models to misspecified unobserved heterogeneity and duration dependence Report AE, University of Amsterdam.
Smith, J. A., 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). Duration models: Specification, identification and multiple durations. In *Handbook of Econometrics*, Vol. 5. Amsterdam, Netherlands: Elsevier, pp. 3381–3460.
Van Den Berg, G. J., 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. doi:10.1016/j.jhealeco.2014.02.004 25804346
Van Den Berg, G. J., Van Ours, J. C. (1994). Unemployment dynamics and duration dependence in France, the Netherlands and the united kingdom. *The Economic Journal* 104(423):432. doi:10.2307/2234762
Van Den Berg, G. J., Van Ours, J. C. (1996). Unemployment dynamics and duration dependence. *Journal of Labor Economics* 14(1):100–125. doi:10.1086/209805
Van Den Berg, G. J., Vikström, J. (2022). Long-run effects of dynamically assigned treatments: A new methodology and an evaluation of training effects on earnings. *Econometrica* 90(3):1337–1354. doi:10.3982/ECTA17522
Van Ours, J. C., Williams, J. (2009). Why parents worry: Initiation into cannabis use by youth and their educational attainment. *Journal of Health Economics* 28(1):132–142. doi:10.1016/j.jhealeco.2008.09.001 18926578
Vikström, J. (2017). Dynamic treatment assignment and evaluation of active labor market policies. *Labour Economics* 49:42–54. doi:10.1016/j.labeco.2017.09.003

Appendix: additional tables and figures

Table A1. Sample statistics and estimates for 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)
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)

Continued

Table A1. Continued

	Treated	Control	Selection model	
			Est.	Std. Er.
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)
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)

Continued

Table A1. Continued

	Treated	Control	Selection model	
			Est.	Std. Er.
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 A2. 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 A3. Bias, standard error, and MSE of the estimated treatment effect when *excluding different sets of covariates*, by model selection criterium 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, standard error, and MSE (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 A4. Bias, standard error (SE), and MSE of the estimated treatment effect when augmenting the baseline model with *covariates correlated in varying degrees* with those excluded from the ToE specifications.

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 2, 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 as explained in Section 5.5. 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 2.

Table A5. Treatment effect bias when the propensity score includes interactions between demographics and short-run history variables.

	No interactions (baseline)				Employment history				Unemployment history			
	MP (1)	Bias (2)	SE (3)	MSE (4)	MP (5)	Bias (6)	SE (7)	MSE (8)	MP (9)	Bias (10)	SE (11)	MSE (12)
Panel A: 10,000 observations												
1 MP	1.00	0.060	0.039	0.0052	1.00	0.071	0.038	0.0065	1.00	0.067	0.038	0,0060
ML	4.11	0.064	0.099	0.0139	5.58	0.031	0.067	0.0055	5.66	0.026	0.070	0.0056
AIC	2.14	0.032	0.076	0.0068	2.28	0.036	0.060	0.0049	2.29	0.031	0.061	0.0047
BIC	1.99	0.027	0.064	0.0048	1.99	0.041	0.055	0.0047	1.99	0.037	0.055	0.0044
HQIC	2.01	0.027	0.064	0.0048	2.02	0.041	0.056	0.0048	2.02	0.037	0.056	0.0045
Panel B: 40,000 observations												
1 MP	1.00	0.057	0.020	0.0037	1.00	0.068	0.020	0.0050	1.00	0.064	0.020	0.0045
ML	3.99	0.037	0.044	0.0033	5.45	0.024	0.033	0.0016	5.50	0.019	0.032	0.0014
AIC	2.21	0.024	0.036	0.0018	2.93	0.027	0.032	0.0017	2.94	0.023	0.031	0.0015
BIC	2.00	0.022	0.031	0.0014	2.00	0.040	0.026	0.0023	2.00	0.036	0.025	0.0019
HQIC	2.00	0.022	0.031	0.0014	2.61	0.030	0.031	0.0019	2.61	0.026	0.031	0.0016

Notes: Treatment effect bias from ToE baseline specifications as in Table 2 of the draft, when the simulated histories are obtained via alternative selection models. The bias and related measures are obtained across 500 replications of samples with 10,000 spells. In *No interactions (baseline)*, the selection model is specified including the covariates listed in Table A1 of the draft. In *Employment history*, we additionally include interactions between demographics (age squared, immigrant status, male, married, with children 0–3, education) and short-run employment history variables (Panel C of Table A1). In *Unemployment history* we add interactions between the demographics and short-run unemployment history variables (Panel D of Table A1).

Table A6. Treatment effect bias when the selection model has a non multiplicative hazard.

	Baseline				Non multiplicative $h_0(t)$: male				Non multiplicative $h_0(t)$: age			
	MP (1)	Bias (2)	SE (3)	MSE (4)	MP (5)	Bias (6)	SE (7)	MSE (8)	MP (9)	Bias (10)	SE (11)	MSE (12)
Panel A: 10,000 observations												
1 MP	1.00	0.060	0.039	0.0052	1.00	0.062	0.039	0.0054	1.00	0.062	0.039	0.0054
ML	4.11	0.064	0.099	0.0139	5.77	0.029	0.073	0.0062	5.76	0.029	0.073	0.0062
AIC	2.14	0.032	0.076	0.0068	2.37	0.030	0.063	0.0049	2.37	0.030	0.064	0.0050
BIC	1.99	0.027	0.064	0.0048	1.99	0.036	0.058	0.0047	1.99	0.036	0.058	0.0047
HQIC	2.01	0.027	0.064	0.0048	2.04	0.036	0.059	0.0048	2.04	0.036	0.059	0.0047
Panel B: 40,000 observations												
1 MP	1.00	0.057	0.020	0.0037	1.00	0.059	0.020	0.0039	1.00	0.059	0.020	0.0039
ML	3.99	0.037	0.044	0.0033	5.78	0.018	0.035	0.0015	5.79	0.018	0.035	0.0015
AIC	2.21	0.024	0.036	0.0018	2.99	0.021	0.032	0.0014	2.98	0.021	0.032	0.0015
BIC	2.00	0.022	0.031	0.0014	2.02	0.034	0.026	0.0019	2.02	0.034	0.026	0.0019
HQIC	2.00	0.022	0.031	0.0014	2.65	0.023	0.032	0.0016	2.64	0.023	0.032	0.0016

Notes: Treatment effect bias from ToE baseline specifications as in Table 2 of the draft, when the simulated histories are obtained via alternative selection models. The bias and related measures are obtained across 500 replications of samples with 10,000 spells. In Columns 1–4, the selection model is specified including the covariates listed in Table A1 of the draft. In Columns 5–8 we interact the baseline hazard splits by gender when generating the placebo treatments but ignore this when estimating the ToE model. In Columns 9–12 we instead interact the baseline hazard with age-group dummies (24–29, 30–34, 35–39, 40-44, 45–49, and 50–54).

Table A7. Comparison of the actual and the estimated distribution of unobserved heterogeneity V_p in the selection equation.

	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 deviation of the actual and the estimated distribution of the unobserved heterogeneity for the treatment duration. The actual distribution is based on the linear predictor of the covariates excluded from the ToE models. 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 the approximated unobserved heterogeneity distributions include the constant. The ToE model includes baseline socio-economic characteristics, inflow year dummies, regional indicators, and the local unemployment rate.