DISSERTATION SERIES 2019:4

Essays on event history analysis and the effects of social programs on individuals and firms

Stefano Lombardi

IFAU
INSTITUTE FOR
EVALUATION OF
LABOUR MARKET AND
EDUCATION POLICY

The Institute for Evaluation of Labour Market and Education Policy (IFAU) is a research institute under the Swedish Ministry of Employment, situated in Uppsala. IFAU's objective is to promote, support and carry out scientific evaluations. The assignment includes: the effects of labour market and educational policies, studies of the functioning of the labour market and the labour market effects of social insurance policies. IFAU shall also disseminate its results so that they become accessible to different interested parties in Sweden and abroad.

IFAU also provides funding for research projects within its areas of interest. The deadline for applications is October 1 each year. Since the researchers at IFAU are mainly economists, researchers from other disciplines are encouraged to apply for funding.

IFAU is run by a Director-General. The institute has a scientific council, consisting of a chairman, the Director-General and five other members. Among other things, the scientific council proposes a decision for the allocation of research grants. A reference group including representatives for employer organizations and trade unions, as well as the ministries and authorities concerned is also connected to the institute.

Postal address: P O Box 513, 751 20 Uppsala
Visiting address: Kyrkogårdsgatan 6, Uppsala
Phone: +46 18 471 70 70
Fax: +46 18 471 70 71
ifau@ifau.uu.se
www.ifau.se

This doctoral dissertation was defended for the degree of Ph.D. in Economics at the Department of Economics, Uppsala University, on August 30'th, 2019. Essay 1 has been published by IFAU as Working paper 2019:22 and Swedish report 2019:23.

ISSN 1651-4149

Economic Studies 180

Stefano Lombardi

Essays on Event History Analysis and the Effects of Social Programs on Individuals and Firms

Stefano Lombardi

Essays on Event History Analysis and the Effects of Social Programs on Individuals and Firms

UPPSALA
UNIVERSITET

Department of Economics, Uppsala University

Visiting address:	Kyrkogårdsgatan 10, Uppsala, Sweden
Postal address:	Box 513, SE-751 20 Uppsala, Sweden
Telephone:	+46 18 471 00 00
Telefax:	+46 18 471 14 78
Internet:	http://www.nek.uu.se/

ECONOMICS AT UPPSALA UNIVERSITY

The Department of Economics at Uppsala University has a long history. The first chair in Economics in the Nordic countries was instituted at Uppsala University in 1741.

The main focus of research at the department has varied over the years but has typically been oriented towards policy-relevant applied economics, including both theoretical and empirical studies. The currently most active areas of research can be grouped into six categories:

- Labour economics
- Public economics
- Macroeconomics
- Microeconometrics
- Environmental economics
- Housing and urban economics

Additional information about research in progress and published reports is given in our project catalogue. The catalogue can be ordered directly from the Department of Economics.

Dissertation presented at Uppsala University to be publicly examined in Hörsal 2, Ekonomikum, Kyrkogårdsgatan 10, Uppsala, Friday, 30 August 2019 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Conny Wunsch (University of Basel).

Abstract

Lombardi, S. 2019. Essays on Event History Analysis and the Effects of Social Programs on Individuals and Firms. *Economic studies* 180. 150 pp. Uppsala: Department of Economics. ISBN 978-91-506-2769-5.

Essay I: This paper studies threat effects of unemployment insurance (UI) benefit sanctions on job exit rates. Using a difference-in-differences design, I exploit two reforms of the Swedish UI system that made monitoring and sanctions considerably stricter at different points in time for different jobseeker groups. I find that men and long-term unemployed respond to the stricter UI rules by finding jobs faster. I also estimate the effect of receiving a sanction on the job exit rates, and find significant sanction imposition effects. However, a decomposition exercise shows that these effects explain very little of the overall reform effects, which instead are driven the threat of sanction imposition.

Essay II (with Gerard J. van den Berg and Johan Vikström): We use an Empirical Monte Carlo design and rich administrative data to generate realistic placebo treatment durations. First, we highlight important confounders to be controlled for when estimating selection models. Next, we omit some of the covariates used to simulate placebo treatments, and we estimate Timing-of-Events models. The model is generally able to adjust for a large share of the resulting unobserved heterogeneity. However, we find that specifying too many or too few support points to approximate the unobserved heterogeneity distribution leads to large bias. Information criteria that penalize parameter abundance can help selecting the appropriate number of support points.

Essay III (with Oskar Nordström Skans and Johan Vikström): We study how targeted wage subsidies affect the performance of the recruiting firms. Using Swedish linked employer-employee data from 1998–2008, we show that the firms hiring through subsidies substantially outperform other recruiting firms, despite identical pre-treatment performance levels and trends in a wide set of key dimensions. The pattern is less clear from 2007 onwards, after a reform removed the involvement of caseworkers from the subsidy approval process. Our results suggest that targeted employment subsidies can have large positive effects on outcomes of the hiring firms, at least if the policy environment allows for pre-screening by caseworkers.

Essay IV (with Raffaella Piccarreta and Marco Bonetti): We propose different methods for comparing the ability of competing non-nested event history models to generate trajectories that are similar to the observed ones. We first introduce alternative criteria to compare pairwise dissimilarities between observed and simulated sequences. Next, we estimate two alternative multi-state models using data on family formation and childbearing decisions from the Dutch Fertility and Family Survey. We use the estimated models to simulate event histories and to illustrate the proposed comparison criteria.

Keywords: Labor Economics, Unemployment Insurance, Job Search, Monitoring and Sanctions, Policy Evaluation, Dynamic Treatment Evaluation, Duration Analysis, Firm performance, Employer-employee Match, Monte Carlo simulation

Stefano Lombardi, Department of Economics, Box 513, Uppsala University, SE-75120 Uppsala, Sweden. Units outside the University, Office of Labour Market Policy Evaluation, Box 513, Uppsala University, SE-75120 Uppsala, Sweden.

ISSN 0283-7668
ISBN 978-91-506-2769-5
urn:nbn:se:uu:diva-382742 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-382742)

Dedicato ai miei genitori,
e a “nonna sprint” Nivea Bastoni

Acknowledgments

I would like to thank the many people that in different ways contributed to this moment of joy: hurray, it is done!! At first sight, this might seem a quite redundant statement: after all, if you are reading this sentence, it must indeed be the case that it is done. Nonetheless, please let me enjoy this moment, since if I think at even relatively recent times in the past, I had quite mixed feelings about being able to graduate within the four and a half canonical years of the PhD program at Uppsala University (well, five, cough cough).

My first thanks go to Johan Vikström, my main supervisor, and to Oskar Nordström Skans, my co-supervisor. Without them, I would have been easily lost several times along the way. They were key figures for my development as an independent researcher, and I hope one day I can be a good supervisor as they both were to me. I also had the luck to work with both Johan and Oskar in two co-authored projects in this thesis. Their vast knowledge and experience have taught me so many things about what to do (and not to do) in research. Johan's effectiveness in dealing with all sort of problems in our projects taught me that there is always yet another way to tackle a seemingly unsolvable issue. Oskar's calm and ability to see things in perspective was fundamental throughout, and in particular at some key turning points of my journey. Thanks also for gently nudging me towards presenting in international conferences during the program, I learned a lot from these experiences, and now I even look forward to presenting in such (typically stressful) situations.

In what comes in the next pages, get ready to read about causes and effects, with much stress on the timing of focal events and on how they impact on other subsequent events: in short, life and the world that surrounds us. So it makes sense to apply this perspective (briefly, I promise) to what led me (and possibly you) here today. I will left-censor roughly the first 23 years of my life in the interest of space.

Between 2008 and 2010, I was an undergraduate student in economics at Bocconi University. Here I had the luck to attend some great introductory courses in statistics and methodology in social sciences. When the time of the thesis approached, I decided that I wanted to use some of the empirical methods that I had learned to study a real-world problem. Thanks to my parents, I managed to obtain patient data to analyze, and I approached Marco Bonetti for my thesis. It was a great experience, so great that I had him as supervisor of my master thesis as well, this time with also Raffaella Piccarreta (you can now quickly browse through this thesis and notice that they co-authored its last chapter). I thank both of them for having sparked my interest in event history analysis. I thank Raffaella, in particular, for answering last minute dramatic phone calls just before the thesis submission.

I would also like to thank Massimo Florio for being the first person to believe in my potential as a researcher, in times where I still had not received my MSc degree. During a stage at CSIL (Centre for Industrial Studies), he asked me to co-author a book chapter on risk assessment in the context of the cost-benefit analysis of large-scale projects. This would also eventually end up in a collaboration where I designed a risk analysis routine to run Monte Carlo simulations. Both were challenging and fun experiences that certainly contributed to continuing my studies in economics. It is also most likely not by chance that two of the four chapters in this thesis use simulation designs.

Soon after finishing my studies in economics, I was a trainee at the Joint Research Centre (JRC) of the European Commission. I have very fond memories of those times. Both Béatrice d'Hombres and Paolo Paruolo were excellent supervisors that really helped me to integrate in the work environment. Here I experienced the most meaningful event of my life, I met Cristina Bratu, also a trainee at the JRC. I thank Cristina for having started a new crucial phase of our life trajectories together, and for being so incredibly supportive and kind-hearted ever since. I am confident that without her I would not be here. In fact, I mean this literally. Together with other colleagues, we wrote a literature review on active labor market policies, and the names of Uppsala and IFAU were recurrently coming up. As a result, when we applied for PhD programs and were both selected by Uppsala University, we decided to join the economics department. I also thank Paolo for suggesting the Uppsala destination, I can still hear his words "they are good with duration models!" resonating.

Knowing that I would have liked to do research in policy evaluation and event history analysis, I immediately contacted Johan, who together with Oskar were to become my dream team of supervisors. I would also like to thank Gerard van den Berg for being part of the project with Johan and me, and Johan for making this possible. Having the chance to work with Gerard was a dream come true. I hope we will also have a chance to continue collaborating in the future!

During the PhD program, I spent a large part of my time at IFAU (Institute for Evaluation of Labour Market and Education Policy). I thank Olof Åslund for allowing me to be part of this big family, and Johan for proposing this arrangement. Even if I am sometimes a bit shy in social situations, I had a very nice and intellectually stimulating time at IFAU. I really hope for the future to keep strong ties with IFAU people, this is an amazingly good place to do policy evaluation and study labor- and education-related questions. Thanks in particular to my officemates Mathias, Dagmar, Gunnar, Anna and Lillit.

I would also like to thank all my fellow PhD students at Uppsala University. In my cohort, Arnie, Mathias, Daniel, Maria, Mohammad, Jonas, and very importantly Cristina shared the joys and the miseries of being PhD students. In particular, special thanks to Arnie and Anna for being such good friends to Cristina and me! I am also grateful to have been given the chance to teach econometrics exercise sessions in the PhD program. A special thanks to Luca Repetto, I hope one day to develop your excellent teaching skills and dedication! Thanks also to all the younger cohorts of PhD students that participated in the teaching sessions that I was responsible for. It was an incredibly rewarding experience. When I am sad, I read again (a selected sample of) your teaching evaluations, and the sun shines again.

I would also like to say thank you to the entire economics department. This was a great place when I arrived in Sweden five years ago, and I can see that it is getting better and better. It is not possible to single out all people that directly or indirectly, through their research, contributed to my work. The acknowledgments and the bibliographic references in the chapters of this thesis give a sense of how important it has been for me to be part of such a great research environment, thanks to you all! My thanks also include the entire administrative staff for its always excellent and efficient work.

I am also grateful for the possibility to spend one semester at Harvard University, thanks to funding from Handelsbanken and to the invitation of Richard Freeman. If Cristina and I had a great time in Boston, it is also thanks to Ellie, Michele and Axel, who have been good friends to us since then. Special thanks also to Marta Florio for coming to visit us regularly at our place. Marta is a very special friend of mine, and it was a very lucky coincidence that she started her post-doc at Harvard while Cristina and I were there. Marta, thanks for the discussions about life and research, it was a real pleasure to feed you in exchange for beers!

Many thanks to my brother Enrico for listening to my rants about causality and the need of evidence-based policies. I will keep pitching research ideas to you until we can start a project together. Thanks also to the group of my closest friends, *splendida cornice*. In particular, thanks to Francesca Falco and Rocco Polin, they are very special friends of mine, and they patiently had to listen about my research work many times. Thanks Fra for the many discussions about environmental policies! Thanks also to my very good friends Filippo Archi and Vincenzo Paci. They also bore the burden of having as a friend an economist excited about his work.

Lastly, but most importantly, I would like to thank and dedicate this thesis to my parents. No need to bother James Heckman to highlight the importance of early parental input for later outcomes in life. It is very obvious to me that today I would not be here without the education, the constant support, and the material means that I received from them. I would also like to say thank you to my grandmother Nivea, she was an extremely important figure in my childhood and adolescence, and this thesis is dedicated also to her.

A final word to all the people that know me best, I promise that this is the end of my studies!

Uppsala, May 2019

Stefano Lombardi

Contents

Introduction

This thesis touches upon different topics in the broad fields of applied labor economics, dynamic treatment evaluation and event history analysis. In this short introduction, I first describe the main underlying topics connecting the four chapters of the thesis. Then, I summarize one by one each of these chapters more in detail.

A first common element linking the first three chapters is the focus on policy evaluation in the context of unemployment insurance (UI) systems. Evaluating policies is of central concern to economists. It entails comparing outcomes under different regimes, but since we do not always observe the outcomes under all the policy regimes of interest, and since we never observe the same economic unit under two policy regimes at the same time, we need to build counterfactuals. In other words, we need to estimate outcomes relative to states of the world that are not observed.[1]

Unsurprisingly, since doing policy evaluation means building counterfactuals, taking a closer look at the counterfactual world considered helps to understand what type of policy parameter we can identify. The two sets of analyses in the first chapter of this thesis offer a useful illustration of this. In that chapter I am interested in estimating the effect (in a broad sense) of monitoring and sanctions on jobseekers' re-employment rates.[2] A first possibility, largely adopted in the literature, is to estimate the so-called *ex post* or *sanction imposition* effects. The idea is to compare the outcomes of the sanctioned unemployed individuals with those of the jobseekers that have not been sanctioned. In other words, since we will never observe the counterfactual world where sanctioned individuals are also not sanctioned, we use the jobseekers that are not sanctioned as the comparison group.

Importantly, since the comparison considered in this first example is within a system where sanctions exist throughout, the ex post effect that we can identify here does *not* inform us on what happens if a new sanction is introduced

[1] See Holland (1986), Neyman (1923) and Rubin (1978) for the treatment effect approach to policy evaluation, originally proposed in statistics and epidemiology and nowadays widely used in economics. Heckman and Vytlacil (2007), instead, present an econometric framework for the evaluation of social programs explicitly rooted in economic theory. For alternative views on the merits and shortcomings of the two approaches, see for instance Keane (2010) and Deaton (2010) (taking the *structuralists'* side), and Angrist and Pischke (2010) and Banerjee and Duflo (2009) (the *randomistas'* side).

[2] As explained below, in this context, sanctions are unemployment benefit cuts (i.e. monetary fines) imposed on the UI benefit recipients that do not comply with the UI rules, while monitoring consists of all procedures implemented to detect rule violations.

or made stricter in an existing policy setting. To study this, we need to build a different counterfactual. Instead of studying what happens within a given policy setting, we could for instance take advantage of a reform of the UI system, and contrast the outcomes of two jobseeker groups under two alternative policy regimes, one more lenient and one harsher in terms of sanction size or probability of being sanctioned. If one of the two groups is unaffected by the reform, it is possible to use the average pre- and post-reform outcomes of this comparison (or control) group to retrieve the fundamentally unobservable counterfactual average outcome for the treated group (namely, what would have happened to the jobseekers affected by the reform if they had not been subject to the harsher rules). I implement precisely such a research design in the first chapter of the thesis.

I also take advantage of a similar setup in the third chapter of this thesis, where I look at the effect of hiring long-term unemployed through wage subsidies on firm performance. In this case, the counterfactual outcomes are provided by observationally identical firms that also hire long-term unemployed people, but do so without using subsidies. An important aspect that is investigated in this chapter is whether the subsidized hires have a long-lasting effect on the firm performance measures considered. In other words, the temporal dimension is of central interest. After all, subsidies have a limited duration, and it might well be the case that any positive effects are short-lived. Or, it could be that those hired with the subsidies are quickly replaced once the financial incentives expire. The availability of repeated performance measures of the same firms over time allows us to study these and other key elements of this policy. However, note that in this particular research design I do not consider duration (time-to-event) outcomes, as I instead do in all other chapters.

A second underlying topic of this thesis is the use or the study of event history analysis methods. As mentioned, I do this in all chapters of the thesis except for the third chapter. There is a long tradition of using event history analysis (also known as survival analysis or duration analysis) in economics and other disciplines, such as statistics, epidemiology, demography, sociology, actuarial science, and engineering.[3] In economics, event history analysis is one of the leading approaches in dynamic treatment evaluation. To fix ideas, consider the estimation of sanction imposition effects done in the first chapter of this thesis. The focus is still on identifying policy parameters of interest. However, the outcome of interest is now characterized by the timing at which it occurs, and both the treatment and the outcome considered are duration variables, realizations of stochastic processes. This situation occurs naturally in many settings. In the sanctions example, each individual is followed over time, for the period since the inflow into unemployment and until exiting to job or the end of the observation time. The treatment (receiving a benefit sanction) is allowed to happen at any time after the start of the unemployment

[3] For an overview of duration analyses methods in economics, see van den Berg (2001).

spell. The question is to what extent sanction imposition affects the probability of exiting to a job. In general, by considering the treatment as a dynamic process, it is possible to account for non-trivial dynamics of policy assignment rules (e.g. treatment assignment related to past program participation, past unemployment duration, intermediate outcomes and time-varying covariates).

Whenever the outcome of interest is a duration variable (e.g. unemployment duration) or depends on a duration variable realization (e.g. wage of the post-unemployment job), it is well known that the identification of policy parameters becomes, other things equal, more complicated. This is because, as time passes by, the individuals still "at risk" of experiencing the outcome event are a selected sub-sample of the starting population, even in the presence of randomization of the treatment status at the inflow (Abbring and van den Berg, 2005). This feature is commonly known as dynamic selection.[4] In our example, those exiting earlier from the unemployment state are generally characterized by better (unobservable) characteristics, which means that over time the sample of survivors is negatively selected. The omission of such characteristics from the model leads to biased estimates of both the structural duration dependence (the baseline hazard) and the systematic component of the model. This has motivated a large literature on the identification of duration models in the presence of unobserved heterogeneity (see for instance Hausman and Woutersen, 2014, and references therein). In particular, identification requires imposing some structure on the process considered.

Mixed proportional hazard models (such as those used in Chapters 1 and 2) are a popular choice to disentangle structural duration dependence captured by the baseline hazard from the dynamic selection (through unobserved heterogeneity). Chapter 2, in particular, studies a widely adopted mixed proportional hazards model, the Timing-of-Events (ToE) approach to dynamic treatment evaluation (Abbring and van den Berg, 2003). The main goal of the chapter is to assess the performance of the ToE approach under different specification choices faced by researchers in empirical applications. Gaure et al. (2007) did this in a Monte Carlo exercise, by estimating the model by using fully artificial data. In this type of simulation exercise, the researcher has full control over the data generating process. As a consequence, the true size of the treatment effect is known in advance, so that one can estimate the ToE model under alternative situations (e.g. different degrees of unobserved heterogeneity, sample size, or exogenous variation) and assess how likely it is to return the true treatment effect. One possible criticism, however, is that it is not clear to what extent the results from this exercise apply to the real-world situations where the model is used. Moreover, in setting up the simulations there is an inherent degree of arbitrariness that, by definition, cannot be avoided.

In Chapter 2, these criticisms are overcome by using a modified version of the Empirical Monte Carlo (EMC) simulation design originally proposed by

[4] See Ham and LaLonde (1996), van den Berg (2001), and Abbring and Heckman (2007).

Huber et al. (2013). The idea is to use real data to estimate the probability of being treated as a function of a rich set of covariates. The estimated selection model is then used to simulate fictitious (placebo) treatments for the non-treated units. Hence, we know that the true treatment effect is zero and we know the selection model (since we estimated it), so that conditional on the covariates used to estimate the selection model, the treatment is as good as random (unconfoundedness). The simulations can then be used in a similar way as described above. However, in this case we use realistic simulations. Clearly, the extent to which we are able to generate realistic simulations relies on the quality of the data used. This is an appealing feature of the EMC design, since it allows to leverage high-quality administrative data that today are available in different countries. Moreover, the use of realistic simulations allows us to characterize the confounders that need to be adjusted for when estimating selection models. This type of exercise, which is in the same spirit as in a prominent literature started by LaLonde (1986), cannot be meaningfully implemented in a standard Monte Carlo design where the researcher is the one determining in advance the relevant covariates.

There are several reasons to adopt event history analysis methods in general and when doing policy evaluation in particular. A first one is that many phenomena of interest are inherently dynamic, such as transitions from cohabiting to marriage, from unemployment to work, from illness to death, and so on. In this case, least squares methods are not well-suited to deal with right-censoring or other types of missing data problems common when considering duration variables. Moreover, standard reduced-form duration models easily allow for controlling for duration dependence non-parametrically.[5] Naturally, duration models can also be applied to situations where we are not necessarily interested in a causal interpretation of the model estimates. For instance, in the fourth chapter of this thesis, event history analysis models are used to describe and predict family formation and childbearing transitions over time. Suppose that, as in the chapter, there exist two competing models that are not nested into each other. In the analyses, I again take advantage of a simulation design based on real data to support the researcher decision in choosing among the two alternative duration models. Once the two models are estimated, they are treated as the true data generating process and used to simulate event histories. The two set of simulated event histories are then compared to the ones in the real sample by using different measures of how “distant” the observed and simulated sequences are from each other. This allows the researcher to take an informed decision on which model to select on the basis of its predictive performance.

Another important reason to use event history analysis methods in economics is that duration distributions are predicted by different classes of rele-

[5] I take advantage of this in the first chapter where I use a difference-in-differences duration model in order to estimate the total effect of two monitoring and sanctions policies.

vant structural models characterized by accumulation of information and optimal decision-taking over time. These include job search models (see e.g., van den Berg, 2001); learning models (Jovanovic, 1979; 1984); dynamic discrete-time discrete choice models based on latent processes (Heckman and Navarro, 2007; Abbring and Heckman, 2007; Abbring, 2010); and related continuous-time models (Abbring, 2012). For example, a job search model with monitoring and benefit sanctions such as in Abbring et al. (2005) and Lalive et al. (2005) can be used to frame the empirical estimation of sanction imposition effects in the first chapter of this thesis. Overall, these theoretical models can be estimated by using data on duration variables, and they can be used to interpret reduced-form duration models.

Finally, evaluating the effect of policies in an event-history setting also allows one to focus on specific aspects that would be completely lost or highly problematic to study in a static framework. These are, for instance, the role of information shocks and of the dynamic accumulation of information on the timing of the exit from unemployment to job (Crépon et al., 2018),[6] and the explicit distinction between ex ante and ex post policy effects (Abbring and Heckman, 2007). This latter distinction is key in the first chapter of this thesis. In fact, from a policy perspective, focusing solely on ex post sanction effects might have two limitations. First, a lack of alignment with what policy makers can do in practice, which is to control the degree of harshness of the UI system by modifying the monitoring and sanctions rules. A second limitation has to do with a lack of alignment with the policy goal, which is not to punish per se, but rather to deter lack of job search in the entire population of UI recipients. By definition, the notion of deterrence is not captured by sanction imposition effects, which in fact identify the effect of receiving a sanction on top of any existing ex ante effects.

The distinction between ex ante or ex post effects can be very informative for the design of policies. For instance, large ex ante effects in the context of monitoring and sanctions might suggest focusing on making the threat of sanction imposition more credible (i.e. improving the monitoring technology). In the context of training programs, if the ex ante effect of program eligibility is negative and the effect of training participation is positive, then it might be preferable to offer the program earlier in the unemployment spell. Without distinguishing an ex ante and ex post component, it might also be the case that intention-to-treat analyses return a net effect of zero, which of course is not really informative on the mechanism behind the policy ineffectiveness.

In the remainder of this introduction I summarize the chapters of this thesis.

[6] In general, embedding information deriving from time-varying variables is only possible in a dynamic setting. Moreover, time-varying covariates can help to identify the model parameters of interest (Heckman and Taber, 1994; Gaure et al., 2007; and the second chapter of this thesis).

Chapter 1: Threat Effects of Monitoring and Unemployment Insurance Sanctions: Evidence from Two Reforms

The first chapter of this thesis studies threat effects of unemployment insurance (UI) monitoring and benefit sanctions on job exit rates. In this context, "monitoring" consists of all procedures used for detecting lack of job search activity or related non-compliance with the job search requirements, and a "sanction" is a temporary benefit suspension (i.e. a monetary fine) imposed on those who are caught not complying with the job search rules.

The key policy objective of monitoring and sanctions is to deter lack of compliance with job search rules in the entire population of jobseekers, so as to alleviate well-known moral hazard problems introduced when UI benefits are in place (see e.g., Fredriksson and Holmlund, 2006). In line with this, in this chapter I precisely focus on threat effects, which are ex ante policy effects defined as the change in job-search behavior arising from the fear of being sanctioned. By contrast, a large and established literature on UI benefit sanctions has almost exclusively analyzed how people respond to actually receiving a sanction (the so-called sanction imposition effects). However, sanction imposition effects do not directly inform us about the fundamental policy goal of monitoring and sanctions: if monitoring and sanctions policies are successful in deterring moral hazard, then they will especially be so among those who are never sanctioned.

In this work I provide quasi-experimental estimates of monitoring and sanctions on re-employment rates. For identification, I exploit two reforms of the Swedish UI rules that made the system considerably stricter at different points in time for the UI recipients and for the longer-term unemployed that receive alternative activity support benefits. Since each of the two reforms only affected one of the two groups leaving the other one unaffected, I set up a difference-in-differences exercise where in correspondence of each reform date I compare the re-employment rates of the two jobseeker groups, before and after the reform date.

The results show that men and long-term unemployed individuals respond to the tighter monitoring and the threat of sanctions by finding jobs faster. These estimates should be interpreted as total reform effects, potentially driven by changes in both sanction imposition effects and threat effects. For this reason, I propose a simple decomposition exercise that allows me to compare sanction imposition effects and threat effects in the same policy setting. This is the second main contribution of this work. I estimate sanction imposition effects on the re-employment rates using the ToE model, the standard approach adopted in the literature. I find large significant sanction imposition effects. However, the decomposition exercise shows that these effects explain very little of the overall reform effects, so that most of the total effects arise through threat of monitoring and sanctions. A direct policy implication is that the impact of monitoring and sanctions may be severely underestimated when focusing only on the sanction imposition effects as done in the literature.

Chapter 2: Empirical Monte Carlo Evidence on Estimation of Timing-of-Events Models

In this chapter, written together with Gerard van den Berg and Johan Vikström, we use a so-called Empirical Monte Carlo simulation design to study the estimation and performance of the Timing-of-Events (ToE) model. The ToE approach was proposed by Abbring and van den Berg (2003), who specify a bivariate Mixed Proportional Hazard (MPH) model and establish conditions for semi-parametric identification of all model components.[7] The ToE approach has been used in a variety of settings where the outcome of interest is the time until the realization of a given event of interest, and the focus is on quantifying the effect of a treatment taken at any elapsed time. For instance, I use a ToE model in the first chapter of this thesis to estimate the effect of benefit sanctions on the time spent in unemployment.

In this project, we modify the Empirical Monte Carlo approach to explore key specification choices that researchers routinely encounter when estimating the ToE model. We use rich administrative data on Swedish jobseekers eligible to participate in a training program (the treatment). In a first step, we estimate a univariate duration model for the selection into treatment using information on all treated and control units. Then we drop the treated units since they do not play any further role in the simulation design. The estimated selection model is used to generate (placebo) time to treatment durations for all non-treated units. These jobseekers were never really treated, and we do not simulate their re-employment durations (i.e. we simply use the observed durations for all units). Then, by construction, the effect of these placebo treatments is known to be zero. The fact that real data is used instead of a data generating process chosen by the researcher makes the simulation exercise arguably more relevant for real applications.

With these realistically-simulated data we perform two types of analyses. First, we inspect which covariates are important confounders that need to be controlled for when estimating selection models. This is similar to what Lechner and Wunsch (2013) do in th German setting. We show that short-term employment history variables (e.g. capturing the share of time spent in employment), together with baseline socio-economic characteristics, regional and inflow timing information, are able to remove a large share of the selection bias. Overall, adjusting for employment history appears to be relatively more important than adjusting for unemployment, earnings and welfare variables. We also find that adding information about long-term labor market history on top of controlling for short-term history is unimportant.

[7] A notable result in their paper is that identification does not require exclusion restrictions, nor the sequential conditional independence assumptions invoked in the dynamic matching literature. To achieve identification, the authors exploit variation provided by exogenous regressors. They also show that both the random effects and the MPH assumptions can be relaxed, while keeping identification non-parametric, when information on multiple independent spells for each cross-sectional unit is available.

Next, we omit some of the covariates of the selection model and estimate ToE models with a discrete support point distribution for the unobserved heterogeneity. The model is able to adjust for a large share of the unobserved heterogeneity, in particular when exploiting calendar-time variation for identification. However, we also find that using too many or too few mass points to specify the discrete support distribution generally leads to large bias. In this respect, information criteria, in particular those penalizing parameter abundance, can be a useful way to select the appropriate number of support points.

Chapter 3: Targeted Wage Subsidies and Firm Performance

Employment subsidies are an important policy tool to help the disadvantaged to get back to employment (Card et al. 2010, 2017; Kluve, 2010). In this chapter, written with Oskar Nordström Skans and Johan Vikström, we study the impact of wage subsidies targeted to the long-term unemployed on firm-level performance measures (firm size, wage sum, investments, profits, value added, per-worker productivity). The rationale for taking the perspective of the firm is that wage subsidies can in principle lead to crowding out of workers currently in the firm and may have distortionary effects if they are allocated to firms that would otherwise struggle to stay in the market. We also study how the impact of the subsidies changes with the degree of caseworker discretion, and we contribute to the still scarce literature that focuses on how active labor market policies affect the employer-employee match.

For identification of the effect of subsidies on firm performance, we compare firms hiring through subsidies to other observably identical firms that also hire unemployed individuals, but without a subsidy. We only match on pre-treatment levels, but the treated and comparison firms show remarkably similar pre-treatment trends, including in measures that we do not match upon. When using the matched sample to study post-treatment outcome trajectories, we find very different results in two policy systems. Between 1998 and 2006, all targeted wage subsidies needed caseworker approval. This staff-selection scheme is compared to a rules-selection system in place since 2007, which removed the caseworker approval from the subsidies allocation process. Under the staff-selection regime, treated firms substantially and persistently outperform the comparison firms after the treatment in a range of production measures. On the other hand, under the new rules, we find no corresponding changes, but only a positive effect on firms' survival rates. This suggests larger crowding-out effects and more windfall gains.

In order to interpret these results, we show that the difference between systems is not due to differences in the hired workers' characteristics, business cycle conditions, or the increasing share of immigrant workers over time. An alternative hypothesis is that caseworkers act as gatekeepers guarding against displacement of non-subsidized jobs and screen out firms on the margin of exit. In accordance with this, we find that investments (interpreted as a measure of the firm's own expectations about future performance) are lower for treated firms only in the rules-selection scheme. This is consistent with the hypothesis that caseworkers guard against an over-allocation of subsidies to firms with poor internal expectations about future performance. Overall, from a policy perspective, our results suggest that targeted employment subsidies can have large positive effects on post-match outcomes of the hiring firms, at least if the policy environment allows for pre-screening by caseworkers.

Chapter 4: Comparing Discrete Time Multi-state Models Using Dissimilarities

In the last chapter of this thesis, together with Raffaella Piccarreta and Marco Bonetti we consider the situation in which the researcher has access to data on activities (or states) experienced by cross-sectional units over time, and the goal is to describe such state trajectories with alternative non-nested event history models. Such situations occur in a variety of settings, both in biomedical studies and in the social sciences. For example, in epidemiology the health condition of treated individuals is typically observed over time, and in each period the patient can experience remission, occurrence of diseases, or death; in demography, one may be interested in studying family formation sequences; in economics, typical event histories are characterized by transitions between employment, unemployment and out-of-labor force.

Different classes of multi-state models can be used to describe the occurrence over time of events of different kinds (see e.g., Putter et al., 2007). We focus on parametric semi-Markov models for the probability of transitioning towards a given set of states. Such models can prove useful from a descriptive point of view in assessing the relationship of covariates with the evolution of state-trajectories. Specifically, we might be interested in comparing the predictive performance of competing models, that is, their ability to generate trajectories that are "similar" to those observed in the sample at hand. The "similarity" criterion that we adopt in this project is based on a *distance* measure commonly used in sequence analysis. However, any properly defined distance measure can be used within the comparison framework that we propose. In particular, our main goal is to introduce criteria to suitably compare collections of pairwise dissimilarities computed between observed and model-generated sequences. We apply such distance-based criteria to data on family formation and childbearing decisions collected as part of the Dutch Fertility and Family Surveys (FFS) study.

We first use the FFS data to estimate two discrete-time multi-state models, the Multi-State Life Table (MSLT) approach and the State Change model (SCM), respectively proposed by Cai et al. (2006, 2010) and Bonetti et al. (2013). We then use the two estimated models to simulate event history sequences, which we compare to those observed in the FFS sample. We do so by using three alternative distance-based approaches that we propose, and we find that the MSLT model performs better than the SCM. We conclude by discussing possible extensions of the overall strategy presented in this chapter.

References

Abbring, J. H. (2010). Identification of dynamic discrete choice models. *Annual Review of Economics*, 2(1):367–394.

Abbring, J. H. (2012). Mixed hitting-time models. *Econometrica*, 80(2):783–819.

Abbring, J. H. and Heckman, J. J. (2007). Chapter 72 Econometric evaluation of social programs, Part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In *Handbook of Econometrics*, volume 6, pages 5145–5303. Elsevier.

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

Abbring, J. H. and van den Berg, G. J. (2005). Social experiments and instrumental variables with duration outcomes. IFAU working paper, 2005:11.

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

Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2):3–30.

Banerjee, A. V. and Duflo, E. (2009). The experimental approach to development economics. *Annual Review of Economics*, 1(1):151–178.

Bonetti, M., Piccarreta, R., and Salford, G. (2013). Parametric and nonparametric analysis of life courses: An application to family formation patterns. *Demography*, 50(3):881–902.

Boone, J., Fredriksson, P., Holmlund, B., and van Ours, J. C. (2007). Optimal unemployment insurance with monitoring and sanctions. *The Economic Journal*, 117(518):399–421.

Cai, L., Hayward, M., Saito, Y., Lubitz, J., Hagedorn, A., and Crimmins, E. (2010). Estimation of multi-state life table functions and their variability from complex survey data using the SPACE program. *Demographic Research*, 22:129–158.

Cai, L., Schenker, N., and Lubitz, J. (2006). Analysis of functional status transitions by using a semi-Markov process model in the presence of left-censored spells. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(4):477–491.

Card, D., Kluve, J., and Weber, A. (2010). Active labour market policy evaluations: A meta-analysis. *The Economic Journal*, 120(548):F452–F477.

Card, D., Kluve, J., and Weber, A. (2017). What works? A meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association*, 16(3):894–931.

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

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48:424–455.

Fredriksson, P. and Holmlund, B. (2006). Improving incentives in unemployment insurance: A review of recent research. *Journal of Economic Surveys*, 20(3):357–386.

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

Ham, J. C. and LaLonde, R. J. (1996). The effect of sample selection and initial conditions in duration models: Evidence from experimental data on training. *Econometrica*, 64(1):175.

Hausman, J. A. and Woutersen, T. (2014). Estimating a semi-parametric duration model without specifying heterogeneity. *Journal of Econometrics*, 178:114–131.

Heckman, J. J. and Navarro, S. (2007). Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics*, 136(2):341–396.

Heckman, J. J. and Taber, C. R. (1994). Econometric mixture models and more general models for unobservables in duration analysis. *Statistical Methods in Medical Research*, 3:279–299.

Heckman, J. J. and Vytlacil, E. J. (2007). Chapter 70 Econometric evaluation of social programs, Part I: Causal models, structural models and econometric policy evaluation. In *Handbook of Econometrics*, volume 6, pages 4779–4874. Elsevier.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

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

Jovanovic, B. (1979). Job matching and the theory of turnover. *Journal of Political Economy*, 87(5, Part 1):972–990.

Jovanovic, B. (1984). Matching, turnover, and unemployment. *Journal of Political Economy*, 92(1):108–122.

Keane, M. P. (2010). Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics*, 156(1):3–20.

Kluve, J. (2010). The effectiveness of European active labor market programs. *Labour Economics*, 17(6):904–918.

Lalive, R., van Ours, J. C., and Zweimüller, J. (2005). The effect of benefit sanctions on the duration of unemployment. *Journal of the European Economic Association*, 3(6):1386–1417.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–620.

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

Neyman, J. (1923). Statistical problems in agricultural experiments. *Journal of the Royal Statistical Society II*, Suppl. (2):107–180.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6(1):34–58.

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

1. Threat Effects of Monitoring and Unemployment Insurance Sanctions: Evidence from Two Reforms

Acknowledgments: For their guidance I am indebted to Oskar Nordström Skans and Johan Vikstöm. For valuable comments and discussions I also thank Steffen Altmann, Chiara Binelli, Cristina Bratu, Rebecca Diamond, Thomas Dohmen, Peter Fredriksson, Hans Grönqvist, Erik Haking, Caroline Hall, Mounir Karadja, Paul Muller, Björn Öckert, Luca Repetto, Amelie Schiprowski, Alex Solis, Isaac Sorkin, Gerard van den Berg, and seminar audiences at Uppsala University, UCLS Uppsala, IZA Bonn, Stockholm University, University of Milan-Bicocca, IRVAPP Trento, JRC Ispra. This work was supported by FORTE.

1.1 Introduction

Unemployment Insurance (UI) systems provide an important safety net by replacing forgone labor earnings for workers who involuntarily lose their jobs. However, just as for any insurance scheme, UI systems may induce moral hazard.[1] In the case of UI systems, moral hazard may arise in the form of reduced job search. In order to reduce moral hazard, and thereby be able to provide more insurance without adverse labor market consequences, many countries have resorted to the use of monitoring and sanction schemes. In this paper, I provide a comprehensive set of estimates on the effectiveness of such policies.

A useful starting point when thinking about monitoring and sanctions is to note that attempts to deter the misuse of the UI systems through such policies closely resemble attempts to prevent crime through punishment within criminal prosecution systems. In the crime literature, the terms *deterrence* or *threat effect* refer to the change of behavior due to the fear that a given conduct will be sanctioned; *sanction imposition effect*, instead, denotes the change of behavior deriving from the actual experience of punishment.[2] This literature has found that policies based on deterrence can be effective in reducing crime, especially in the case of swift-and-certain punishment regimes that provide salient incentives (see e.g., Weisburd et al., 2008; Hawken and Kleiman, 2009).[3] Most attention has been paid to deterrence for two main reasons: first, deterrence can directly modify the behavior of *all* individuals eligible for sanctions (not just of those actually sanctioned); second, since it can be effective also for individuals that are not directly caught misbehaving, deterrence has substantial cost-saving potential compared to the actual sanction imposition.

Deterrence is, however, absent from almost all studies in the context of UI systems. These have instead analyzed the effect of imposing monetary fines (benefit sanctions) on the individuals actually sanctioned. This paper brings together the insights of the crime and UI literatures. Starting from the idea that deterrence is central also in UI systems, I study both threat and sanction imposition effects in the same policy setting. In doing so, I provide estimates of threat effects of stricter monitoring and UI benefit sanctions through a quasi-experimental design.

In the context of UI systems, benefit sanctions are used to correct moral hazard problems arising when unemployed individuals are granted UI benefits. While jobseekers can insure themselves against unexpected income losses due to job separations, the UI benefits receipt is made conditional on exerting

[1] UI systems may also be associated with adverse selection problems, although this has been emphasized less in the literature (for an example of this, see Landais et al., 2017).

[2] Conceptually, crime deterrence can be seen in an expected utility framework where higher probability of apprehension and higher sanction size reduce the value of misbehaving (Becker, 1968). A similar setting applies to UI systems where jobseekers are monitored and a lack of job search is sanctioned with monetary fines.

[3] For reviews on crime deterrence, see Chalfin and McCrary (2017), Nagin (2013a, 2013b).

sufficient job search effort, which is monitored by caseworkers at the public employment service (PES). Inactivity and lack of cooperation may lead to UI benefit sanctions, corresponding to temporary suspensions of benefits.

Monitoring and UI benefit sanctions can be theoretically justified as being welfare enhancing (Boone et al., 2007). In practice, however, efficiency gains can be reached either by modifying the behavior of the UI recipients actually sanctioned (sanction imposition effect) or by modifying the UI recipients' search effort through the threat of sanction imposition (threat effect). From the policy-maker's perspective, if monitoring is costly and imperfect, it is the threat effect that really matters. This is because the main objective is to diminish moral hazard in the entire population of jobseekers exerting low search effort, not just among those actually sanctioned. Despite their relevance, however, empirical evidence on threat effects is extremely scarce.

The main contribution of this paper is to fill this gap by providing the first credible evidence of threat effects in UI monitoring and sanctions systems. I exploit variation induced by two reforms of the Swedish monitoring and sanctions system that substantially increased the strictness of the system. For each of the two reforms, I compare the job exit rates of two groups of jobseekers before and after the policy change in a difference-in-differences (DID) setting. The jobseekers that I compare are the unemployed individuals receiving UI benefits (UI group) and the longer-term unemployed that have exhausted their UI benefits and receive activity support benefits (AS group).[4] Individuals in these two groups compete for jobs in the same labor market, are exposed to the same business cycle conditions, and all start their unemployment spell by receiving UI benefits. The main difference between the two groups is that AS recipients, by definition, have been unemployed longer.

In September 2013, following a pre-reform period where sanctions were almost non-existent and the monitoring intensity was moderate, the stock of UI recipients started being subjected to a considerably stricter policy regime. The reform resulted in a substantial increase in the number of UI sanctions issued. Moreover, monitoring became stricter due to the mandatory requirement of submitting monthly reports of the job search activity. In January 2014, a second reform introduced the same monthly activity reporting tool for the AS recipients. Therefore, the first reform allows me to estimate the effect of stricter monitoring and sanctions on the UI group job exit rate (using the AS jobseekers as controls), while the second reform allows me to estimate the effect of stricter monitoring on the AS group (using the UI jobseekers as controls). Importantly, the two reforms allow me to study the two relevant policy margins in this context: the joint introduction of stricter monitoring and sanctions, and the introduction of stricter monitoring only.

[4] Specifically, AS benefits are given conditional on participating in labor market programs, whereas UI benefits are given to individuals who are openly unemployed.

Identification of the policy parameters of interest is facilitated by the fact that all jobseekers that I sample are characterized by relatively long unemployment durations. In order to take into account the fact that AS recipients are comparatively longer-term unemployed, and hence likely more negatively selected as compared to the UI recipients, I estimate DID-duration models where I control for duration dependence non-parametrically. I additionally adjust for a rich set of time and seasonality fixed effects in order to control for differential trends that would otherwise invalidate identification. For estimation, I use rich administrative data providing information on individual-level unemployment histories at the daily level, daily benefit payments and sanctions information, and background characteristics for the entire population of jobseekers.

I find large and significant reform effects for male jobseekers, and especially for the long-term unemployed individuals affected by the second reform (21 percent increase in the job exit rate). The fact that jobseekers tend to respond later during their unemployment spell is in line with existing evidence on active labor market policies (ALMPs), see e.g. Card et al. (2017). Conversely, I do not find significant reform effects for women, which is consistent with some existing evidence on ALMPs (Card et al., 2017; Bergemann and Van Den Berg, 2008).[5] I run several checks to corroborate these findings. First, I rule out the existence of differential trends by performing placebo exercises where I shift the reform dates back in time and, separately, move forward the duration threshold for the UI individuals' eligibility to transition to the AS group. Moreover, I check for and find no support for group compositional differences across the reform dates (which would confound the reform estimates). Several robustness checks also support the findings in the main analyses.

The second main contribution of this paper is a decomposition of the estimated total effect of the first reform into its threat and sanction imposition effects. In order to estimate sanction imposition effects, I follow the convention in the existing literature and use a flexible bivariate duration model where I jointly model the exit to job rate and the sanction process (Abbring and van den Berg, 2003). I find a 29 percent increase in the job exit rate as a consequence of sanction imposition. This result is consistent with previous evidence on sanction effects. Moreover, results are similar in size when splitting the sample based on gender. This shows that the heterogeneous total reform effects do not arise because of different sanction imposition effects. Instead, they must be driven by differences in threat effects.

To quantify the size of threat effects, I perform a decomposition exercise where I subtract the sanction imposition component from the estimated total reform effect. To make these two quantities comparable, I adjust for the probability of being sanctioned and for the proportion of the spells duration that

[5] With respect to monitoring and sanctions policies, evidence from other countries does not point to a clear direction in terms of heterogeneous effects by gender (see e.g. McVicar, 2014).

on average is covered by a sanction. I find that for male UI jobseekers, a large part of the total reform effect is attributable to the threat component, which accounts for a 10.3 percent increase in the job exit rate out of the total 11 percent increase due to the reform. For women, the weighted sanction effect is even smaller in size, and accounts for a negligible part of the (insignificant) total reform effects. This is consistent with the fact that for this group the total reform effect was not found to be significantly different from zero. All in all, the results from the decomposition exercise suggest that the sanction imposition effects emphasized in the literature explain very little of the overall effects of sanctions.

Despite the fact that the objective of monitoring and sanctions is to deter moral hazard in the form of violations of job search requirements, almost all studies of UI sanctions (see below for details) have focused on estimating the effect of sanction imposition on the individuals actually sanctioned. A likely reason for the lack of evidence on deterrence effects is that identification is challenging. It requires for the researcher to compare counterfactual outcomes under different policy settings characterized by different sanctions schedules and/or probabilities of apprehension. Moreover, in order for the policies to change the job search behavior of UI claimants, it is crucial that the policy differences are substantial and salient. These are core aspects of the two reforms considered in this paper.

One exception is Boone et al. (2009), who provide direct evidence of the threat effects of benefit sanctions. Through a small-scale laboratory experiment, the authors compare two systems characterized by identical expected benefits, one with constant benefits and the other with higher baseline benefits and a positive probability of being sanctioned. They find that the threat of introducing the sanctions system on the job acceptance probability is equal to 14.1 percentage points, while the sanction effect equals 10 percentage points. However, it is unclear to what extent these results translate into a real-world setting.

The only two other papers studying threat effects of sanctions are Lalive et al. (2005) and Arni et al. (2013), which exploit within-regional differences in the rate at which warnings are issued. They show a positive correlation between the cross-PES offices variation in the job finding rate and the variation in the propensity of issuing warnings. Lalive et al. (2005), in particular, find an elasticity of the job exit rate with respect to the warning rate of 0.13. In a simulation exercise, both studies show relevant sanction effects (with unemployment duration reduced by almost 3 weeks for the sanctioned) and substantial threat effects (with a reduction of the unemployment rate of about 7 days for all jobseekers).

This paper also relates to a broad empirical literature on the effect of sanction imposition mentioned above. Taken together, papers in this field (almost) unambiguously find that sanction imposition increases job exit rates through increased search effort and/or reduced reservation wage,[6] whereas the quality of the jobs found is persistently worsened.[7] Moreover, since sanctions are coupled with monitoring, and often with elements of job-search assistance, the literature on benefit sanctions partly overlaps with that on ALMPs.[8]

The remainder of the paper is structured as follows. Section 1.2 outlines the institutional background. Section 1.3 describes the identification of the causal parameters of interest, the sampling criteria and the data used. Sections 1.4 and 1.5 present the main analyses results and the comparison between threat effects and sanction imposition effects, respectively. Finally, Section 1.6 summarizes and concludes.

1.2 Institutional background

1.2.1 Unemployment Insurance and activity support entitlement

In Sweden, UI benefit sanctions rules apply to all UI recipients. Openly unemployed jobseekers older than 20 years can be eligible for either basic UI compensation or income-related UI compensation (IAF, 2014c). The entitlement conditions for basic UI benefits are registering at a PES office, actively seeking work, being able and willing to work at least three hours each working day and 17 hours per week, and having fulfilled a *work condition* (have worked for at least 6 out of the 12 months prior to unemployment, at least 80 hours per month). Individuals eligible for basic UI benefits gain the right to income-related UI benefits if they additionally have been a voluntary member of a UI fund for at least 12 months (*membership condition*). Full-time unemployed UI recipients are entitled to a full 300-day period of daily cash transfers paid at most 5 times per week, which corresponds to 420 calendar time days. In the time frame considered in this paper, the size of UI payments is 320-680 Swedish Crowns (SEK) per day ($\approx$ 35-75 €). The lower bound corresponds to the basic UI. Jobseekers eligible for income-related benefits are entitled to 80 percent of their former salary for the first 200 days of unemployment and 70 percent for the remaining 100 days, capped at SEK 680 per day.[9]

[6] van der Klaauw et al. (2004) and Abbring et al. (2005) find large re-employment effects after sanction impositions for UI and welfare recipients in the Netherlands, respectively. Similar results have been found in many other settings, such as Switzerland (Lalive et al., 2005), Denmark (Svarer, 2011), Germany (Hofmann, 2008; van den Berg et al. (2013); Müller and Steiner, 2008), and Norway (Røed and Westlie, 2012).

[7] See e.g., Arni et al. (2013) and van den Berg and Vikström (2014). Other studies have also found differential effects of sanctions and financial bonuses (van der Klaauw and van Ours, 2013), and for different types of unemployment benefits (Busk, 2016).

[8] For exhaustive reviews on ALMPs see Card et al. (2017), Card et al. (2010), Kluve (2010), Crépon and van den Berg (2016), and Caliendo and Schmidl (2016).

[9] By international comparison, the Swedish system is relatively generous. See Immervoll and Knotz (2018) and Grubb (2000) for cross-countries job search requirements and UI eligibility criteria.

In the main analyses, the sample is restricted to full-time unemployed individuals that start their unemployment spell with a full 300-day UI period. This allows me to know at which duration time the individuals exhaust their UI benefits. I refer to this first group of jobseekers as the *UI group*.

After exhausting their UI benefits, jobseekers become eligible to receive activity support (AS) upon enrolling in the so-called *Job and Development Program*.[10,11] The daily transfers are equal to 65 percent of the previous earnings, with the same minimum and maximum levels as for the UI benefits. Since I restrict my attention to jobseekers with full UI replacement period at the inflow, in my sample the people that reach 420 unemployment duration days are eligible to transition to the Job and Development Program. I refer to this second group of jobseekers as the *AS group*.[12]

1.2.2 Monitoring and sanctions before the reforms

A central feature of the Swedish UI system is that benefit recipients need to actively search for a new job. Newly unemployed individuals that register at a PES office are required to agree on a personalized plan of action decided together with a caseworker, with the goal of exiting from unemployment. This makes the right to receive UI compensation *conditional* on exerting a sufficient level of search effort.

The jobseeker's activities are monitored by caseworkers at the PES. Caseworkers inform jobseekers about the conditions for UI entitlement, the requirement of looking for a suitable job, the importance of meetings at the PES, and the underlying reasons for being sanctioned (that is, mishandling the job search process and prolonging or causing unemployment). After the initial creation of the action plan, which in most cases takes place within one month since the PES registration (IAF, 2014c), caseworkers have meetings with the unemployed individuals. During these meetings, caseworkers pro-

[10] The Job and Development program provides long-term unemployed with targeted activities corresponding to 75 percent of the individual's potential labor supply. After 450 days in the program, participants enter into a workfare scheme and are assigned to full-time work in low-qualified occupations. Eligibility conditions are to be unemployed and registered at the PES and (i) to have exhausted a full set of UI benefits, or (ii) to have been unemployed or in an ALMP for 14 months. Special rules apply to former participants in the youth guarantee and to parents of minor children (Arbetsförmedlingen, 2017).

[11] Note that the AS benefits are also given to jobseekers that participate in other labor market programs, possibly before they exhaust their UI benefits. As it is discussed more in detail in Section 1.3, this has implications on who is actually treated by the two reforms.

[12] As mentioned before, jobseekers ineligible for UI benefits become eligible to enroll in the Job and Development Program if they have been registered as unemployed or enrolled in a labor market policy program for 14 months. This group of unemployed people is excluded from my analyses since everyone in the sample starts with 300 days of UI benefits. Moreover, since special eligibility rules apply to young unemployed individuals, I focus on jobseekers older than 24 years.

pose ALMPs, refer appropriate vacancies, and provide counseling. Meetings are also used to monitor the jobseekers' compliance with the UI rules. Before the reforms that I study, the meetings were the only form of monitoring.

Benefit sanctions are monetary fines corresponding to a suspension of the UI benefits. Inactivity, refusal of job offers, and job quits are valid reasons for a sanction. In case the rules are violated, the caseworker sends a notification to the UI fund, which decides whether to impose a sanction.[13]

The Swedish sanctions system is characterized by a staircase model, with increasing sanction size for each violation of the rules (IAF, 2014c). Overall, sanctions are grouped into three categories: job offer rejections, lack of compliance with the general UI eligibility rules, and job quits with no valid cause. In the pre-reform period, the refusal of suitable job offers without an acceptable reason is punished with a 25 percent benefits reduction for 40 days at the first offense, with a 50 percent reduction for 40 days the second time, and with benefit suspension until a new work condition is fulfilled the third time. UI recipients can also be sanctioned for infringements related to violations of the UI entitlement conditions. These include unreported employment, failure to actively search for a job, not showing up at meetings, not signing the action plan, and failing to apply for assigned jobs. In these cases UI benefits are suspended until a new work condition is fulfilled.[14]

Two main aspects characterize the monitoring and sanctions system before September 2013. First, the per-jobseeker number of sanctions imposed was close to zero (see Figure 1 below). In this period, Sweden was among the EU countries with the lowest sanction rate (Gray, 2003). As discussed by van den Berg and Vikström (2014), one main reason for such a low sanction rate is that the system was perceived as too harsh by caseworkers, who therefore were reluctant to use this policy instrument. The second feature of the pre-reform UI system is that monitoring intensity was rather low. Monitoring occurred only through meetings with the caseworker, which on average took place less than once a month (0.8 jobseeker meetings per month; Liljeberg and Söderström, 2017). Thus, the pre-reforms period is characterized by moderate monitoring and extremely low sanction rate.

[13] The proportion of notifications leading to a sanction for the 2013-2014 period is close to 80 percent (IAF, 2014b). Individuals can in principle appeal to a sanction, but this rarely happens. The decision is taken quickly, in most cases within 2 or 3 weeks since the notification.

[14] During the time frame of the analyses, AS recipients might lose the right to receive activity support benefit in case of expulsion from the Job and Development Program (due to unreported employment or other gross violations of entitlement conditions; IAF, 2014c), but this happens very rarely.

1.2.3 Two reforms of the monitoring and sanctions system

The September 2013 reform for the UI recipients

In September 2013 a reform of the system was implemented for the UI recipients. Its objective was to improve the job search incentives of the unemployed through enhanced monitoring technology and increased sanction rate (IAF, 2014a; Arbetsförmedlingen, 2014).

A first main policy change was the introduction of a new monitoring system based on *monthly activity reports*. Latest the 14th of each month, UI recipients now have to hand in a summary of all job search activities in the last month. Typically, the reports are submitted electronically, and caseworkers should use them to monitor the UI recipients' job search effort and to provide job search assistance.[15] Recall that in the pre-reform period the monitoring activity of the caseworkers was exclusively carried out during meetings with the jobseekers. Importantly, the stated policy purpose of the activity reports was *not* to replace meetings (IAF, 2014a). This is confirmed by the observed meetings intensity, that did not change after the reform (Liljeberg and Söderström, 2017). Thus, the new activity reports provided caseworkers with a new and improved way of detecting violations of the rules, and led to tighter monitoring of the jobseekers.

A second major policy change was a quick and substantial increase in the number of sanctions imposed. Different factors contributed to the sharp increase in the sanction rate. First, the sanctions schedule was made less punitive with the purpose of encouraging caseworkers to use this policy instrument.[16] Second, failing to submit an activity report was included among the reasons for being sanctioned. Third, some notifications started being sent automatically to the UI funds (failure to show up at meetings or to submit an activity report). Overall, the aim was to make the sanction process more efficient and less arbitrary.

The changes of the sanctions system had a tremendous impact on the number of sanctions. Figure1 shows the total number of sanctions per jobseeker. Before the reform, the sanction rate was very close to zero; after the reform, the number of sanctions increased dramatically.

As mentioned above, the reform also introduced a less harsh sanction schedule. Figure 2 shows that before the reform, the average sanction size was

[15] According to survey evidence, in about 80 percent of the cases the activity reports are inspected by the caseworker within 14 days (Arbetsförmedlingen, 2014).

[16] Under the new rules, job offer refusal sanctions correspond to 5, 10, 45 days of suspension for the first 3 times, and to the loss of entitlement until new work requirement (capped at 112 days) for the fourth time. UI eligibility sanctions (including the failure to submit an activity report) correspond to a first time warning, 1 day, 5 days, 10 days, and loss of entitlement for the subsequent infringements.

Figure 1. Number of sanctions per unemployment spell

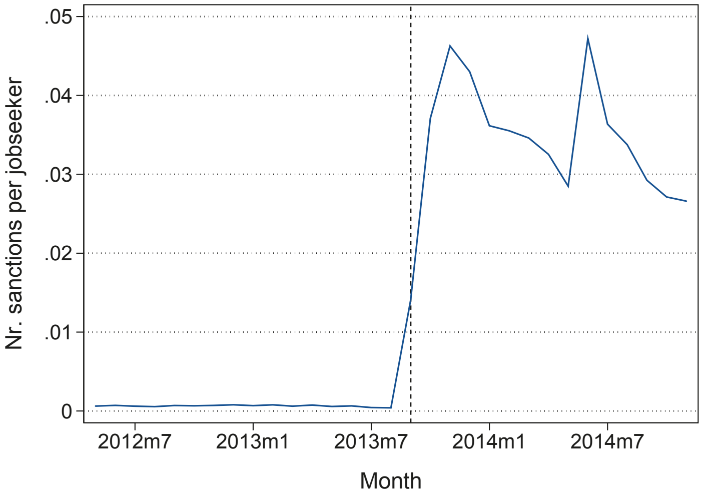
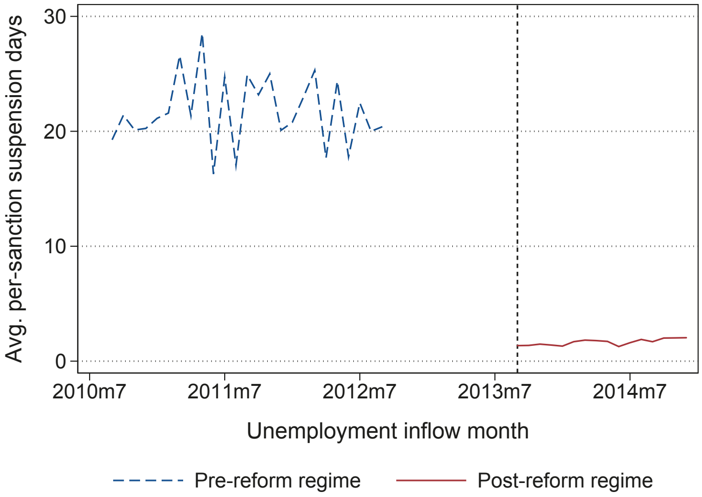


Figure 2. Average length of the sanctions, in days



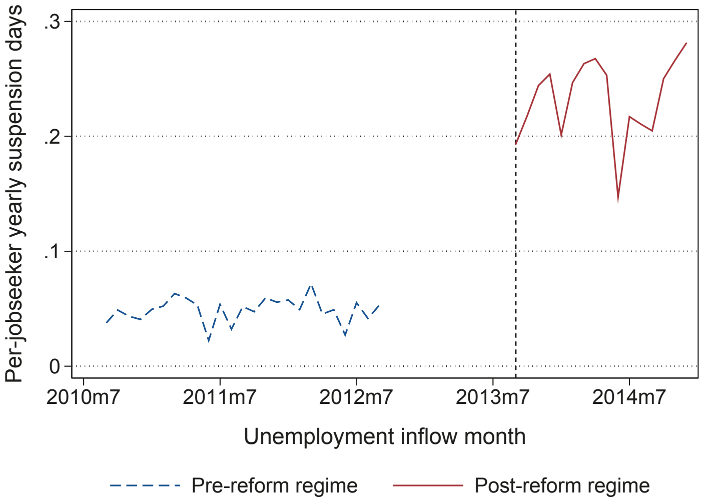
around 20 days of suspension. After the September 2013 reform, the average sanction size decreased to roughly 2.5 days.[17]

To compare the relative importance of the drastically increased sanction rate and the reduced sanctions size, Figure 3 shows the average number of UI

[17] I follow for one year the group of jobseekers inflowing into full-time unemployment in a given month, and compute for this group of jobseekers the share of individuals sanctioned over the year. I repeat this for each inflow into unemployment month, and stop exactly one year before the first reform in order to not mix up the two reform regimes.

suspension days within the first 12 months of unemployment. This provides a measure of the expected sanction cost, which reflects changes in both the rate and the size of the sanctions.[18] Figure 3 shows that the expected sanction cost increased dramatically as a result of the new rules. This is because before the reform the sanction rate is virtually zero, and after the reform the increase in the number of sanctions more than outweighs the decrease in their size. Thus, unless jobseekers are extremely risk averse, the new stricter system provides greatly enhanced job search incentives compared to the old one. Moreover, van den Berg and Vikström (2014) find that the size of the sanction imposed is secondary compared to the shock of being sanctioned, which suggests that the increased sanction rate is relatively more important than its decreased size. This has also been confirmed outside UI systems, e.g. for the enforcement of court-ordered financial obligations (Weisburd et al., 2008) and of probation and parole (Hawken and Kleiman, 2009).

Figure 3. Expected sanction cost per-newly unemployed



In sum, the new monthly activity reports and the sharp increase in the number of sanctions implied a substantially stricter monitoring and sanctions system. No other changes to the UI system were made.

The January 2014 reform for the AS recipients

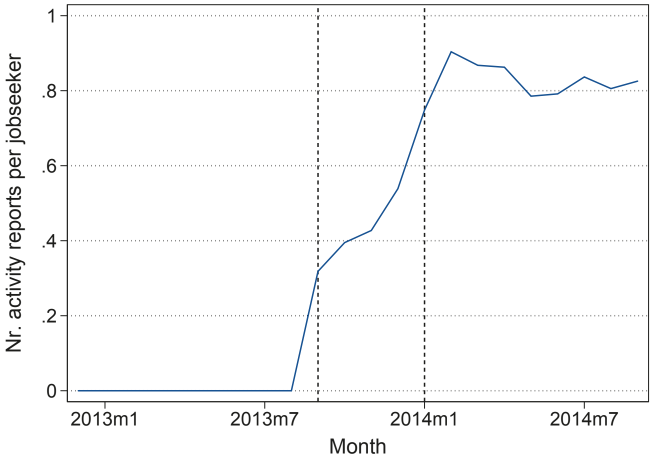
In January 2014, a second reform of the monitoring system was implemented for the AS group, that is the jobseekers who have exhausted their UI benefits

[18] Similarly as done for Figure 2, in order to keep the two periods separated, I stop summing up the sanctions one year before September 2013. Sanctions of an "indefinite" length – in practice capped at a higher bound number of days – are assumed to last their maximum possible duration.

and start receiving activity support benefits. The goal of the reform was to make the overall monitoring system similar for both UI and AS jobseekers.

Before January 2014, the AS group was only subject to monitoring through caseworker meetings.[19] The reform enhanced monitoring by extending the system of monthly activity reports already in place for the UI jobseekers to all AS recipients. Figure 4 illustrates this by showing the aggregate number of monthly activity reports per jobseeker. The figure shows a first increase in September 2013, due to the introduction of the activity reports for the UI recipients, and a second increase in January 2014, relative to the second reform affecting the AS group. The second reform did not change the sanction rules, so that the sanction rate remained practically zero for the AS group also after the rule changes. Thus, this second reform implied substantially tighter monitoring for the AS recipients, with no changes in the sanctions regime.[20]

Figure 4. Number of per-jobseeker activity reports



[19] The AS group was subject to benefit sanctions only in the extremely rare cases of expulsion from the Job and Development Program.

[20] Despite the probability of being sanctioned is low for the AS recipients, losing the benefits is still a possibility in case of gross violations of the eligibility rules. Other reasons why the AS recipients may increase their job search due to the reform include: (i) if they value the regular submission of the activity reports (e.g. because this may enhance the quality of the job search assistance provided by the caseworker), and (ii) if they dislike being caught breaking the rules when not submitting the activity report.

1.3 Empirical strategy and data

1.3.1 Difference-in-differences (DID) design

In order to estimate the effects of the two policy changes, I use the rollout of the two reforms for the UI and AS groups in a DID setting. Recall that *all* sampled individuals start in the UI group (the full-time unemployed with a complete number of UI benefit days at the inflow). Out of these, the UI jobseekers that remain unemployed after 420 days are eligible to transition to the Job and Development Program and to receive AS benefits.[21] Jobseekers in the UI and AS groups are exposed to similar business cycle conditions and compete on the same labor market. The main difference between the two groups is that the AS jobseekers are longer-term unemployed. In order to make these two groups more similar, I sample spells with durations relatively close to and centered around 420 days, the threshold in correspondence of which jobseekers are eligible to transition to the AS group (see 1.3.2 for more details).

The outcome of interest is the re-employment rate, that is the hazard of exiting from full-time unemployment to job.[22] For each reform, I compare the outcomes of the two groups before and after the date of the policy change (the treatment). The estimated parameter is the *total effect* of the policy shift, averaged across the treated individuals. Since the final goal is to quantify threat effects, these total reform estimates still need to be decomposed into threat and sanction imposition effects. The model used for estimating sanction effects is described in Section 1.5.1, whereas the decomposition exercise is presented in Section 1.5.2.

Consider the September 2013 reform that made the monitoring and sanctions regime stricter for all UI recipients without changing the existing rules for the AS group. In this case, I compare the re-employment rate of the UI recipients (the treated group) to that of the AS recipients (the comparison group), before and after September 2013. This returns the average effect of the stricter monitoring and sanctions reform on the UI recipients. I use a similar DID approach for the second reform, where the AS recipients (the treated group) are compared to the UI recipients (the controls) before and after January 2014. In this case, I estimate the average effect of the reform on the AS recipients.

Throughout the DID analyses, individuals are classified as transitioning to the AS group at 420 unemployment duration days, i.e. when they exhaust their UI benefits and are *eligible* to enroll in the Job and Development Program and collect activity support benefits. All estimates should accordingly be interpreted as Intention-to-treat estimates (ITT). The ITT strategy is motivated by the fact that the actual AS-transition is not a deterministic function of the time

[21] I refer to these longer-term unemployed as the *AS group*, although, as mentioned earlier, jobseekers can collect activity support benefits also if they participate in other types of programs, possibly before they exhaust their UI benefits.

[22] I consider both full-time, part-time and subsidized jobs to define the event of exiting to job.

spent in unemployment: it usually, but not always, occurs at 420 days after the first UI payment.[23] This is because the unemployed may not use benefits at the full speed, for instance because they receive sickness benefits or are on parental leave during the unemployment spell. Hence, using the transition eligibility allows me to avoid using the actual timing of the transition, which in general is not random.[24] The identification strategy exploits two sources of variation: calendar time and unemployment duration. Note that spells crossing the 420-day threshold and the reform dates contribute to the identification of the parameters of interest.

The identification of the reform effects in the DID setting requires absence of differential time trends in the two groups and no anticipatory effects of the reform. If this is the case, the observed average pre- and post-reform outcomes of the comparison group can be used to retrieve the counterfactual average outcome for the treated group (e.g., for the first reform, what would have happened to UI recipients in the absence of the new monitoring and sanctions rules). By design, all time-fixed differences in the two groups are netted out.

Formally, the model is the following. Let d be unemployment duration (in days), m and y calendar month and year, and $g = UI, AS$ the jobseeker group. Define $D^{(1)} \equiv D_d^{UI} \cdot D^{\text{Sept2013}}$ to be the first reform indicator, i.e. the time-varying treatment variable equal to one for the UI group after September 1$^{\text{st}}$ 2013, and equal to zero otherwise. Here, D_d^{UI} is a time-varying indicator for being in the UI group, and D^{Sept2013} is a time-varying indicator for being in the post-September 2013 period. Moreover, $D^{(2)} \equiv (1 - D_d^{UI}) \cdot D^{\text{Jan2014}}$ is the second reform indicator equal to one for the AS group after January 1$^{\text{st}}$ 2014. I estimate the following Cox model for the hazard of exiting unemployment:

$$\ln\theta(g,d,m,y) = \ln\lambda_d + \beta_1 D^{(1)} + \beta_2 D^{(2)} + \lambda_{my}, \qquad (1.1)$$

where the two parameters of interest are β_1, the effect of being in the new monitoring and sanctions regime for the UI group, and β_2, the effect of being

[23] Graph A.1 in the Appendix reports the distance (in weeks) between the UI benefits exhaustion and the date the Job and Development Program starts, in the raw sample of jobseekers inflowing into full-time unemployment within the time window considered in the main analyses.

[24] Jobseekers that have been unemployed less than 420 days are always classified as UI recipients, including those that participate in programs before 420 duration days and hence receive AS benefits. Hence, some jobseekers in the UI group are not treated by the first reform, while others are actually treated by the second reform. The number of misclassified UI recipients due to this type of pre-420 duration exits to programs is likely low since all sampled individuals are long-term unemployed relatively close to exhaust their UI benefits (see Section 1.3.2). These jobseekers likely have little incentives enrolling in programs that confiscate leisure time, especially the closer they get to UI exhaustion. Overall, this misclassification would imply an attenuation bias in the positive estimates from the two reforms in case the jobseekers wrongly classified as UI recipients are negatively selected as compared to their openly unemployed counterparts (for instance due to lock-in effects from program participation).

subject to the activity reports monitoring regime for the AS group.[25] Note also that over a given spell $D^{(1)}$ and $D^{(2)}$ can switch on and off depending on the group the unemployed person belongs to and if the reform has been implemented or not.

The job exit hazard $\theta(\cdot)$ on the left-hand side of (1.1) is the instantaneous (daily) probability of exiting to job conditional on being unemployed up to duration time d. It is modeled as a function of the baseline hazard $\ln \lambda_d$, that captures non-parametrically the unemployment duration dependence; λ_{my}, a set of year-specific monthly fixed effects capturing calendar time-specific effects common to the two groups. In robustness specifications I further add λ_{mg}, a set of monthly fixed effects that controls for group-specific seasonality.[26] As explained in the next section, I sample spells close to and centered around 420 duration days. This makes the common trends assumption more likely to hold. In order to further balance trends in the two groups, I control for monthly time fixed effects, and in robustness specifications I additionally adjust for the group-specific seasonality fixed effects. Moreover, since a main difference between the two groups is that the AS recipients are by definition longer-term unemployed (and hence potentially more negatively selected over unemployment time than the UI recipients), I also non-parametrically control for duration dependence. Finally, after setting up the estimation model, I formally test for the absence of differential group trends by estimating placebo reform effects where I anticipate the reform dates to test whether they are statistically different from zero. Moreover, in a different placebo exercise I move forward the duration threshold for the UI individuals' eligibility to transition to the AS group.

One concern is that the effect of the first reform may change the composition of the controls in the second reform (since UI jobseekers are treated during the first reform and are used as comparison group later on). To formally assess this possibility (dynamic selection), I test for changes in a range of observable characteristics between the UI and AS groups before and after the two reforms.

1.3.2 Data description

I use information from several Swedish administrative registers. Data from the Swedish Public Employment Service provides information on all unemployment spells (at daily level) and rich background characteristics. Population

[25] The residual variation exploited for identification of β_1 and β_2 comes from within month differences in the two groups, after netting out monthly seasonality fluctuations specific for the UI and AS recipients and the duration dependence component.

[26] The main effects D^{Sept2013} and D^{Jan2014} are implicitly controlled for through the λ_{my} terms. The main effect D_d^{UI} is omitted in (1.1), as in the DID specifications that I estimate I assign the transition to the AS group at $d = 420$ (ITT framework). Hence, D_d^{UI} cannot be separately identified from the baseline hazard.

registers from Statistics Sweden (LOUISE) provide additional background characteristics. I use the register called ASTAT from the Swedish Unemployment Insurance Board (IAF) to link information on the number of UI benefit days. The same register includes daily information on all benefit sanctions.

Sampling and descriptive statistics

I construct the analyses sample in the following way. First, I select all unemployment spells starting with full-time unemployment. Age at inflow is restricted to be between 25 and 50. This is done because young people are subject to special eligibility rules for participation in the Job and Development Program, and older workers may be eligible for early retirement schemes and other targeted policies. The analyses also exclude all individuals with disabilities. Next, I retain only spells with a full 300 UI days at the start of the spell (equivalent to 420 calendar time days). Moreover, I focus on spells of a duration of between 280 an 560 days, i.e. relatively close to and centered around the 420-day threshold. Shorter spells are ignored, and all ongoing spells are right-censored after 560 days.[27] I start to sample unemployment spells two years before the first reform of September 2013, and I include spells up until March 2015, where any ongoing spells are right-censored. This ensures that enough pre-reform observations are available to capture the pre-treatment trends through the rich set of time and seasonality fixed effects.

Table 1 shows descriptive statistics. The columns report group averages in the three periods (before September 2013; between September 2013 and January 2014; and after January 2014). All characteristics are measured at the time of inflow into unemployment. The table shows that the AS group is composed of jobseekers that are less educated and more likely to be immigrants and married. Compared to the UI group, they also have a weaker attachment to the labor market and a lower income in the three years preceding the start of the spell. All this shows that the longer-term unemployed (the AS group) have less favorable characteristics than their shorter-term unemployed counterparts (the UI group). Note that this is not a problem for the identification of the reform effects, since the DID model adjusts for all time-fixed differences between the two groups. What would be problematic are *changes* in group differences over time. However, this does not appear to be the case, since Table 1 shows that the group differences are stable over time. Later, I formally test for such dynamic selection patterns.

[27] The advantage of sampling jobseekers relatively close to the transition threshold is that the two groups compared are more similar than they would otherwise be when sampling short-term unemployed as well. One potential disadvantage is that restricting the sample to spells lasting at least 280 days might introduce sample selection if the first reform has a strong impact on the short-term unemployed, hence affecting the probability of sampling jobseekers thereafter. In the robustness analyses section, I implicitly check for this possibility by testing for compositional changes of the two groups over time.

Table 1. *Group averages in the three reform periods*

	Before Sept. 2013		Sept. 2013– Jan. 2014		After Jan. 2014	
	UI	AS	UI	AS	UI	AS
Age	37.41	37.81	37.52	37.73	37.52	37.89
Education: compulsory	0.19	0.21	0.19	0.22	0.19	0.21
Education: secondary	0.47	0.47	0.47	0.45	0.46	0.46
Education: upper	0.34	0.32	0.34	0.33	0.35	0.33
Any child below 18	0.42	0.42	0.41	0.42	0.41	0.42
Immigrant	0.48	0.53	0.52	0.56	0.52	0.56
Married	0.40	0.42	0.41	0.43	0.40	0.42
Male	0.57	0.58	0.60	0.60	0.59	0.60
Unemployed 24 months before	0.27	0.30	0.32	0.36	0.31	0.36
Any program in last 24 months	0.03	0.03	0.03	0.04	0.03	0.03
Duration of last unempl. spell	200	221	232	253	239	260
Any program in last 4 years	0.04	0.05	0.05	0.05	0.05	0.05
Previous income (past 3 years)	1658	1543	1699	1503	1763	1624
Inflow year: 2010	0.31	0.35	0.00	0.00	0.00	0.00
Inflow year: 2011	0.33	0.40	0.00	0.00	0.00	0.00
Inflow year: 2012	0.36	0.25	0.60	1.00	0.09	0.46
Inflow year: 2013	0.00	0.00	0.40	0.00	0.91	0.54
No. observations	32,185	16,565	8,468	4,972	12,132	7,999

Notes: Average observables in the UI and AS groups, by reform period as defined by the reform dates (Sept. 2013 and Jan. 2014). All socio-economic characteristics and previous labor market history measured at the inflow into unemployment. Previous income in 100s SEK.

1.4 The total effects of the two reforms

1.4.1 Main results

I start by estimating the effects of the two reforms by gender. This has been shown to be a relevant dimension according to which ALMPs effects vary (see e.g., Card et al., 2017; Bergemann and Van Den Berg, 2008) Table 2 presents the estimates using the DID model presented in Section 3 for the exit rate to a job (re-employment rate).

Panel A shows the results for men. To start with, Column 1 presents placebo estimates where I shift the entire observation window and anticipate the reform dates by two years. Apart from this, the overall data structure, sampling criteria and estimated model are kept exactly as in the main analyses. Any significant placebo estimates would raise doubts on the validity of the identification strategy and the parallel-trends assumption. This is not the case, since placebo estimates in Column 1 are insignificant and very close to zero.

Next, Column 2 of Panel A reports the estimates for the actual reform period. The table shows that the re-employment rate for male jobseekers is significantly affected by the first reform (11 percent increase).[28] The effect of the second reform is even larger, with a 21 percent increase of the re-employment rate. The results are robust to the additional inclusion of socio-economic characteristics (Column 3).

Table 2. *Total effects of the monitoring and sanction reforms, by gender*

	Placebo period	Reform period	
	(1)	(2)	(3)
Panel A: Men			
Reform 1: Monitoring and sanctions, UI recipients	-0.02 (0.08)	0.11^{*} (0.07)	0.11^{*} (0.07)
Reform 2: Monitoring, AS recipients	0.04 (0.08)	0.21^{***} (0.07)	0.21^{***} (0.07)
No. individuals	18,301	25,682	25,682
Spell duration	✓	✓	✓
Calendar Time FE	✓	✓	✓
Covariates			✓
Panel B: Women			
Reform 1: Monitoring and sanctions, UI recipients	0.006 (0.09)	-0.05 (0.08)	-0.03 (0.08)
Reform 2: Monitoring, AS recipients	0.05 (0.10)	-0.06 (0.08)	-0.03 (0.08)
No. individuals	13,884	19,066	19,066
Spell duration	✓	✓	✓
Calendar Time FE	✓	✓	✓
Covariates			✓

Notes: DID-Cox model estimates for the re-employment rate using the data described in Section 1.3.2. The covariates include: dummy for any children, age, migrant status, married, education. Robust standard errors in parentheses. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

These results may appear puzzling since the second reform provides individuals with stricter monitoring, while the first reform introduces both stricter sanctions and stricter monitoring. However, remember that the two reforms affect different groups of jobseekers: the second reform affects the long-term

[28] Since the model coefficients measure changes in log re-employment rates, estimates are interpreted as percentage changes in the re-employment rate when the corresponding covariates are increased by one unit. In the pre-September 2013 period, the re-employment rate the month preceding the 420 duration days threshold is equal to 0.113 (and similar for men and women).

unemployed (AS group), while the first one affects more the shorter-term unemployed (UI group). If the long-term unemployed react differently to monitoring incentives, this explains the different effects of the two reforms. In fact, a common finding in the literature is that the long-term unemployed tend to benefit more from ALMPs (Card et al., 2017). Another difference between the two reforms is that both the pre- and post-reforms strictness of the system for the two groups is different.[29] Hence, we should not necessarily expect the first reform to have a larger effect than the second one.

Interestingly, Panel B shows no significant effects of any of the two reforms for women. One way of interpreting the heterogeneous effects for men and women is that gender differences may reflect differential attitudes towards risk-taking behavior. Experimental evidence have robustly shown that women tend to be more risk averse than men (Croson and Gneezy, 2009; Charness and Gneezy, 2012). If women are more likely to comply with the rules from start,[30] whereas men tend do so only when there is a sizable threat of sanction imposition, this can explain the heterogeneity of the total reform effects. To better understand this, the next section reports estimates of the sanction imposition effects separately by gender. This may reveal whether the gender differences are due to differential threat effects or differential sanction imposition effects.

Table 3. *Total effects of the monitoring and sanction reforms*

	Placebo	Reform period	
	-1 Year	Main	With covariates
	(1)	(2)	(3)
Reform 1: Monitoring and sanctions, UI recipients	-0.006 (0.06)	0.05 (0.05)	0.05 (0.05)
Reform 2: Monitoring, AS recipients	0.05 (0.06)	0.10* (0.06)	0.12** (0.06)
No. individuals	32,185	44,748	44,748
Spell duration	✓	✓	✓
Calendar Time FE	✓	✓	✓
Covariates			✓

Notes: DID-Cox model estimates for the re-employment rate using the data described in Section 1.3.2. The covariates include: dummy for any children, age, migrant status, gender, married, education. Robust standard errors in parentheses. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

[29] The UI group was subject to sanctions already before the rules changed (although the sanction rate was very low). Hence, UI jobseekers pass from a moderate monitoring and sanctions system to a stricter one. Instead, AS jobseekers pass from an even milder pre-reform period with even lower probability of being sanctioned to a stricter monitoring-only one.

[30] Consistently, before September 2013, women are 30% less likely to be sanctioned than men.

Table 3 shows the effects of the two reforms when pooling men and women. As before, the placebo estimates in Column 1 are insignificant. Column 2 shows a 10 percent increase in the exit to job rate of AS jobseekers due the monitoring reform. The point estimate for the first reform is also positive, but not significantly different from zero.

1.4.2 Robustness Analyses

This section presents three sets of robustness analyses. First, I test for the lack of changes in compositional differences between the UI and AS groups before and after the reforms (dynamic selection). Second, I present the results from alternative model specifications to test the robustness of the main analyses estimates. Finally, I implement additional placebo checks to test the parallel trends assumption.

Dynamic selection

Identification in the DID model relies on a comparison of the re-employment rates around the AS threshold after 420 days of unemployment. Thus, different spell segments are compared to each other (early parts being UI, later parts being AS). A potential concern, is that any treatment effects during the first part of the spells (i.e. for the first reform, the effect of stricter monitoring and sanctions for the UI recipients) may change the composition of jobseekers that remain in the second part of the spells. This creates the so-called dynamic selection problem, which may confound the estimated effects due to the changes in the composition of the groups.

To address this, I replace the outcome (re-employment rate) with observed characteristics (such as socio-economic variables) measured at the unemployment inflow. Otherwise, I estimate the DID model as in the main analyses. This offers one way of studying the assumptions underlying the DID model, since significant estimates for these observed variables would indicate problems with dynamic selection. Specifically, I regress each observed characteristic on the two reform indicators, the AS group indicator and the interaction between the two. This DID exercise allows me to compute the outcome averages for the UI and AS groups in the three calendar time periods defined by the two reforms (see Columns 1-3 of Table 4). For each reform, the regression coefficients on the interaction term return the difference in the two groups averages across the given reform date (Columns 4 and 6).

Reassuringly, Table 4 reveals no significant estimates and all point estimates are very close to zero (see the *p*-values in Column 5 and 7). This is true also when considering the entire set of covariates in a joint test. I also construct a measure of predicted unemployment duration using all the observed covariates, and use this as an outcome in the same DID regression framework. The group differences in predicted unemployment across the reforms are also

Table 4. *Dynamic selection and compositional differences*

	AS and UI mean differences			Period 2 vs 1		Period 3 vs 2	
	Period 1	Period 2	Period 3	Est.	p-value	Est.	p-value
Outcome	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Age	0.392	0.205	0.363	-0.187	0.214	0.158	0.354
Education: compulsory	0.021	0.022	0.017	0.002	0.783	-0.005	0.552
Education: secondary	-0.001	-0.016	-0.001	-0.015	0.133	0.015	0.189
Education: upper	-0.020	-0.007	-0.017	0.013	0.175	-0.010	0.374
Any child below 18	0.002	0.005	0.003	0.002	0.857	-0.002	0.846
Immigrant	0.048	0.044	0.039	-0.005	0.599	-0.004	0.695
Married	0.014	0.023	0.020	0.008	0.409	-0.002	0.836
Male	0.013	-0.001	0.017	-0.014	0.150	0.019	0.102
Unemployed 24 months before	0.029	0.039	0.044	0.010	0.271	0.005	0.636
Any program in last 24 months	0.003	0.007	0.005	0.004	0.246	-0.002	0.667
Duration of last unemployment spell	20.93	21.02	21.37	0.084	0.989	0.348	0.959
Any program in last 4 years	0.004	0.004	0.001	0.001	0.864	-0.003	0.591
Past avg. income (in last 3 years)	-0.146	-0.188	-0.133	-0.043	0.315	0.056	0.246
Joint significance p-value				0.421		0.892	
Predicted unemployment duration	1.557	1.384	1.449	-0.172	0.431	0.065	0.793

Notes: DID regressions for AS and UI compositional differences across regimes. Columns (1)–(3): average outcome differences between the two groups in the policy regimes defined by the two reforms dates. Columns (4) and (6): DID estimates when respectively comparing (2) to (1) and (3) to (2). Previous income averaged in the three years preceding the unemployment inflow (in log-scale). Predicted unemployment duration computed by: (i) regressing unemployment duration on all observables using period 1 spell parts; and (ii) using the estimated model to predict for all periods. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

insignificant. All this suggests that dynamic selection is not an issue, hence supporting my main results and identification strategy.

Robustness of the reform effects

Table 5 presents results from several robustness analyses with the baseline results in the first column for comparison.

In the main analyses, I adjust for general seasonal variation through time-varying calendar time indicators. Here, Column 2 of Table 5 reports model estimates where I add group-specific monthly dummy variables, which additionally adjust for different seasonal dynamics in the UI and AS groups. Despite the second reform effect is not significant anymore, the point estimates are very robust to the inclusion of these seasonality controls, hence ruling out that the observed effects are due to group-specific seasonality effects.

As explained above, in the main analyses the transition of the UI jobseekers to the AS group is assigned at 420 days, without using the actual transition date (which is potentially endogenous). This comes at the cost of increasing noise, since some unemployed individuals transition to the AS group already before this threshold, and others do so after the threshold. This is not problematic for identification, since the exogenous 420-day threshold is used for *all* jobseekers, but it may reduce the precision of the estimates. Therefore, in Column 3 of Table 5 I explore whether it is possible to obtain a stronger first stage for identification. To this end, I “dummy out” the first month after the 420-day threshold, so that the spell parts immediately following the AS transition do not contribute to the estimation of the reform effects.[31] This procedure returns very similar estimates to the ones in the main analyses.

The UI recipients that remain unemployed and exhaust their UI benefits eventually transition to the AS group. Here, one concern is that workers may increase their search effort just before exhausting their UI benefits (see e.g., Card et al. 2007). However, note that the DID model flexibly adjusts for duration dependence (through the baseline hazard), and this also controls for increased exit rates just before benefit exhaustion. However, one may worry that these anticipatory effects change in correspondence with the two reforms. To check for this, Column 4 of Table 5 reports estimates where the period before the AS transition is “dummied out" in a similar way as above (with the pre-420 month indicator interacted with the two reforms variables). The estimates are robust to this exercise.

Next, I report robustness analyses with respect to the sampling window. In the main analyses, all spells range between 280 and 560 days. The last two columns of Table 5 show results when varying the size of the duration window. Specifically, I extend this window (Column 5) and tighten the window

[31] Specifically, I add a time-varying indicator switching to one during the 30 days following the 420-day threshold and I interact it with the two reforms indicators.

Table 5. *Robustness analyses for the total reform effects*

	Baseline (1)	Group-specific seasonality (2)	Control for month post UI exhaustion (3)	Control for month before UI exhaustion (4)	Duration 250-590 (5)	Duration 310-530 (6)
Reform 1: Monitoring and sanctions, UI recipients	0.05 (0.05)	0.05 (0.06)	0.04 (0.06)	0.05 (0.05)	0.03 (0.05)	0.04 (0.06)
Reform 2: Monitoring, AS recipients	0.10* (0.06)	0.10 (0.07)	0.13* (0.07)	0.11* (0.06)	0.08 (0.05)	0.09 (0.06)
No. individuals	44,748	44,748	44,748	44,748	51,607	39,253
Spell duration	✓	✓	✓	✓	✓	✓
Calendar Time FE	✓	✓	✓	✓	✓	✓

Notes: Robustness estimates of the main results when using the full sample. Column 1: baseline results (spells range between 280 and 560 days); Column 2: additional inclusion of group-specific seasonal dummies; Columns 3 and 4: partition out the month following and preceding the 420-day threshold, respectively; Columns 5 and 6: sampling spells ranging in 250-590 and 310-530 duration days, respectively. Robust standard errors in parentheses. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

(Column 6) around the 420-day threshold. In both cases the results are very similar to those in the main analyses.

Finally, Table A.1 reports the same robustness checks separately for men and women. The estimates show that also the main results for these two groups are robust.

Extended placebo analyses

Table 6. *Placebo analyses for the total reform effects*

	Placebo calendar time		Placebo spell duration	
	-2 Years (1)	-1 Year (2)	480-760 (3)	580-860 (4)
Reform 1: Monitoring and sanction, UI recipients	-0.001 (0.06)	-0.006 (0.06)	-0.09 (0.08)	-0.08 (0.10)
Reform 2: Monitoring, AS recipients	-0.07 (0.07)	0.05 (0.06)	-0.08 (0.08)	-0.06 (0.11)
No. individuals	40,184	32,185	19,319	13,358

Notes: Placebo estimates when anticipating the reform dates (Columns 1 and 2), and when delaying the transition to the AS group (Columns 3 and 4). Spells used in Columns 1 and 2 range between 280 and 560 duration days. Robust standard errors in parentheses. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

Identification of the reform effects is based on variation across calendar time and spell duration. Table 6 shows the results from extended placebo analyses where I misplace the reform dates and the duration time thresholds.

First, I study placebo effects for different placebo reform dates. To this aim, I show results when moving the entire sampling window back in time one and two years, respectively in Columns 1 and 2. The dates are moved by exactly one or two years to preserve the same seasonal structure that characterizes the sampling window of the main analyses. The resulting placebo estimates are always insignificant.

Second, in the main analyses, the duration of all sampled spells ranges between 280 and 560 days, with the UI to AS threshold at 420 days. In Columns 3 and 4 of Table 6, this sampling window is shifted to 480–760 days and 580–860 days, with placebo thresholds at 620 and 720 days. Otherwise, the model structure is the same with reform dates at September 2013 and January 2014. Since at these thresholds there are no reform changes, I expect the corresponding placebo estimates to be zero. From the table we see that the point estimates are negative but insignificant, supporting the main results. The only potential issue is the size of the placebo estimates. However, their negative sign indicates that, if anything, the positive estimates from the real period should be biased towards zero.

1.5 Relationship between threat and sanction effects

1.5.1 Sanction imposition effects

To obtain the threat effect of the monitoring and sanctions regime, it is necessary to decompose the reform effect into a threat effect component and a sanction imposition component. To estimate sanction imposition effects, I focus on the sanctions imposed during the new monitoring and sanctions regime, when the large increase in the sanctions rate took place. To this aim, I sample unemployment spells starting after September 2013 and merge UI benefit sanctions to the spells. Durations are right-censored at the end of 2015. I proceed as in the main analyses, and select only spells of full-time and non-disabled unemployed, aged between 25 and 50 at the inflow. I sample only spells of UI recipients.[32] I do not distinguish between the different types of UI benefit sanctions, and I focus on the first sanction during the unemployment spell.[33]

Identification of sanction effects

To estimate the effect of a sanction I use a bivariate duration model commonly referred to as the *Timing-of-Events* (ToE) model (Abbring and van den Berg, 2003). This model is the standard approach for the estimation of sanction effects (see e.g., Arni et al., 2013; van den Berg and Vikström, 2014).

In this framework, the goal is to identify the causal effect of a sanction on the re-employment rate (θ_e, the outcome of interest). The challenge is that sanctions are not random events. Many observable and unobservable factors may influence the sanction rate, and these factors are likely to also affect the re-employment rate. Hence, I jointly model the re-employment rate and the sanction rate, θ_s. Let d be time in unemployment, λ_{ed} and λ_{sd} are baseline hazard functions capturing duration dependence, x is a set of determinants observable to the researcher, and D_d is a time-varying treatment indicator taking the value one after a sanction has been imposed. The model includes the unobserved heterogeneity terms $v = (v_e, v_s)'$, that are allowed to be correlated; each captures the effect of unobserved determinants respectively on the re-employment rate and the sanction rate. The model is:

$$\ln\theta_e(d,x,D,v_e) = \ln\lambda_{ed} + x'\beta_e + \delta D_d + v_e \qquad (1.2)$$

$$\ln\theta_s(d,x,v_s) = \ln\lambda_{sd} + x'\beta_s + v_s, \qquad (1.3)$$

where δ represents the treatment effect of interest (here assumed to be constant, but it can be allowed to vary with duration d, time since treatment, or observed characteristics x).

Identification relies on the following assumptions (Abbring and van den Berg, 2003). First, individuals must not be able to anticipate the *exact* timing

[32] Being more restrictive by selecting only those with the full amount of UI benefits at the inflow does not qualitatively change the results.

[33] To avoid misclassification, I restrict the spells so that they are least 15 days long.

of the sanction (*no anticipation*). In this setting, several aspects of the sanction assignment process are unknown to the jobseeker, for instance because the actual decision is taken by the UI fund. Moreover, even if some jobseekers might anticipate the timing of a notification, UI funds typically decide upon imposing a sanction soon after they are notified. This leaves jobseekers with little time to adjust their job search behavior in anticipation of the sanction imposition. A second assumption is the Mixed Proportional Hazard (MPH) structure in (1.2) and (1.3) (*MPH assumption*). Third, x and v should be independently distributed (*random effects assumption*). The last two assumptions can be relaxed if multiple-spell data is used (Abbring and van den Berg, 2003).

If these and some additional regularity conditions hold, the model is nonparametrically identified. Note that identification does not require exclusion restrictions (the x vector is the same in the two hazard rates). This makes the model particularly appealing in this setting, since quasi-experimental variation in the assignment of sanctions is not available and exclusion restrictions would be hard to justify. Intuitively, identification is achieved by a quick succession of events. If a sanction is rapidly followed by a transition from unemployment to employment, this is evidence of a causal effect, whereas any selection effects do not give rise to the same type of quick succession of events.

Estimation of sanction effects

In order to estimate the ToE model, it is necessary to specify the baseline hazards, the distribution of the unobserved heterogeneity and select the covariates. I follow the common practice in the literature and use a discrete support point distribution for the unobserved heterogeneity (Lindsay, 1983; Heckman and Singer, 1984). To select the number of support points, I rely on the evidence in Gaure et al. (2007) and Lombardi et al. (2019).

In the simulation study by Gaure et al. (2007), the authors find that the general approach of approximating the unobserved heterogeneity through a discrete distribution performs well. However, they also highlight that unjustified restrictions, such as pre-defining a small number of support points for the discrete distribution, may result in large bias. Lombardi et al. (2019) also study ToE specification issues, but use a different simulation approach based on actual data (the so-called Empirical Monte Carlo design; see Huber et al., 2013). The use of data on real outcomes and covariates to simulate placebo treatment spells has the advantage of providing evidence more closely linked to real applications and based on less arbitrarily chosen data generating processes. One central conclusion is that it is important to use information criteria to select the number of support points.

Here, I use three information criteria: the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the Hannan-Quinn information criterion (HQIC). The number of support points is selected based on the number that maximizes the given information criterion. To search for the support points values, I use the same search algorithm as in Gaure et al. (2007) and Lombardi et al. (2019).

For the baseline hazard functions, I use a piecewise constant distribution (8 duration pieces). The observed covariates include a rich set of baseline socio-economic characteristics (gender, age and education dummies), regional dummies, quarterly inflow indicators, regional unemployment rate at the time of inflow, and a set of variables capturing previous labor market history.[34]

Sanction effects estimates

In accordance with the analyses of the reforms, I estimate sanction effects both when using the full sample and separately for men and women. All information criteria previously defined return 4 mass points as the preferred specification, for both the full sample and the split sample estimations.

Table 7, Column 1, reports the sanction effect for the full sample. Here, the point estimate of 0.291 indicates that jobseekers exit to job roughly 29 percent faster after being sanctioned. This is consistent with the fact that sanctions decrease the value of staying unemployed, leading to increased job-search intensity and/or decreased reservation wages. Interestingly, the estimated effect is very similar in size to the baseline results in van den Berg and Vikström (2014), who study sanction effects in the Swedish setting before September 2013. Overall, the size of the estimated sanction effect is large, but smaller than the effect of sanctions in other countries. For instance, for the Netherlands Abbring et al. (2005) find that a sanction doubles the job exit rate. For Switzerland, the total effect of a warning and a sanction increases the re-employment rate by around 50 percent (Lalive et al., 2005). Note that these cross-country comparisons do not take into account differences in sanction size, which can vary across countries (see e.g., Grubb, 2000; McVicar, 2014).

From Columns 2 and 3 of Table 7, we see that the effects are very similar for men and women: men exit to job 24 percent faster after a sanction, while the same increase is 22 percent for women. The fact that the sanction effect is very similar for men and women is contrary to what was found in the analysis of the total reform effects, where we saw large effects for men but small and insignificant effects for women (for both reforms). Thus, it is clear that sanction effects do not drive the heterogeneity in the total reform effects previously found. Instead, such heterogeneous patterns must be due to differences in threat effects.

[34] In additional analyses, I specify a more detailed set of inflow dummies (monthly) and extend the previous labor market history characteristics to include short-term history variables (up to 2 years before the inflow). Results are qualitatively similar and available upon request.

Table 7. *Sanction effects in the new monitoring and sanctions regime*

	All (1)	Men (2)	Women (3)
Sanction effect	0.291*** (0.047)	0.221*** (0.072)	0.241*** (0.055)
No. individuals	178,843	96,824	82,019

Notes: Timing-of-Events estimates. Unobserved heterogeneity approximated with 4 mass points. Controls include: timing of inflow; socio-economic characteristics; local labor market (region, regional unemployment rate); unemployment history (up to 10 years before the unemployment inflow). Standard errors in parentheses. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

1.5.2 The relationship between threat and sanction effects

In this section, I decompose the total effect of the first reform into the threat effect and sanction effect components. This allows me to compare the relative importance of the two elements, both for the full sample and when splitting it according to gender. Note that different aspects make the decomposition not straightforward. First, threat effects may have an impact on the sanction rate. Second, UI jobseekers were subject to sanctions already before September 2013 (although as mentioned, the sanction rate was very close to zero). Lastly, the magnitude of the threat effects may in principle change over the time spent in unemployment. In the decomposition exercise, I simplify the analysis by assuming constant sanction rate, by not considering pre-reform sanction effects, and by assuming constant threat effects over duration time. The decomposition is performed according to the following formula:

$$\text{Threat effect} = \text{Total effect} - \text{Sanction effect } \times (p \cdot coverage), \tag{1.4}$$

where the threat effect on the left-hand side is computed as the difference between the total effect of the September 2013 reform and the weighed sanction imposition effect. The size of the sanction imposition effect is rescaled to make it comparable to the total reform effect. In particular, the weighting term $p \cdot coverage$ is a function of (i) p, the share of the sanctioned individuals among those used in the sanction effect estimation; and (ii) $coverage$, the fraction of the spell length that on average is covered by the imposed sanctions for the subset of sanctioned individuals.

Table 8 shows that a large part of the total reform effects estimated with the DID model is due to threat effects, not to the actual imposition of sanctions. In fact, after rescaling the sanction effects to make them comparable to the total reform effects, their size becomes relatively small. In particular, when looking at the full sample and comparing weighted sanction effect and threat effect (Columns 5 and 6), the threat of being in a stricter system leads to a

4.2 percent increase in the exit to job rate, which is more than five times the weighted sanction effect.

An even more extreme pattern is found for male UI recipients. For them the threat effect (10.3 percent job exit increase out of the total 11 percent increase) is larger than for the full sample. For women, sanction imposition effects are similar in size to those of men and become extremely small after weighting them. For this group, there is no the threat effect since both reform effects were not found to be significantly different from zero.

Table 8. *Threat and sanction imposition effects comparison*

Group	Total reform effect	Proportion sanctioned	Spell part covered by sanction	Sanction effect	Weighted sanction effect	Threat effect
	(1)	(2)	(3)	(4)	(5)	(6)
All	0.05	0.060	44.23%	0.291	0.008	0.042
Men	0.11	0.073	44.87%	0.221	0.007	0.103
Women	-0.05	0.044	42.98%	0.241	0.005	-0.054

Notes: Threat effects computed as the difference between the total effect of the September 2013 reform (Column 1) and the weighted sanction imposition effect (Column 5). The weighting factor is equal to the share of jobseekers sanctioned during the post-reform period (Column 2) multiplied by the average spell part covered by the sanction (Column 3).

1.6 Conclusions

This paper explores threat effects in the context of UI systems, where the job search behavior of jobseekers is monitored and lack of search activity is sanctioned with UI benefits suspension. Despite the goal of monitoring and sanctions is to deter lack of job search of all the unemployed, threat effects have received very limited attention in the UI literature.

One result is that male jobseekers significantly and robustly increase their job finding rates in response to a shift to a stricter monitoring and sanctions system. In line with existing evidence, the effects are larger for the long-term unemployed and no effects are found for women. These overall reform effects can be the result of changes in threat effects, changes in sanction imposition effects, or a combination of the two. However the decomposition exercise shows that the threat effects largely dominate the sanction imposition effects.

Overall, this study shows that the threat of sanction imposition can enhance the job search effort of the eligible jobseekers, above and beyond the effect of actual sanction imposition. It also shows that sanction imposition effects emphasized in the literature only account for a minor part of the reform effects since they are small compared to the threat effects.

Appendix

Figure A.1. Time between UI exhaustion and Job and Development Program start (in weeks)

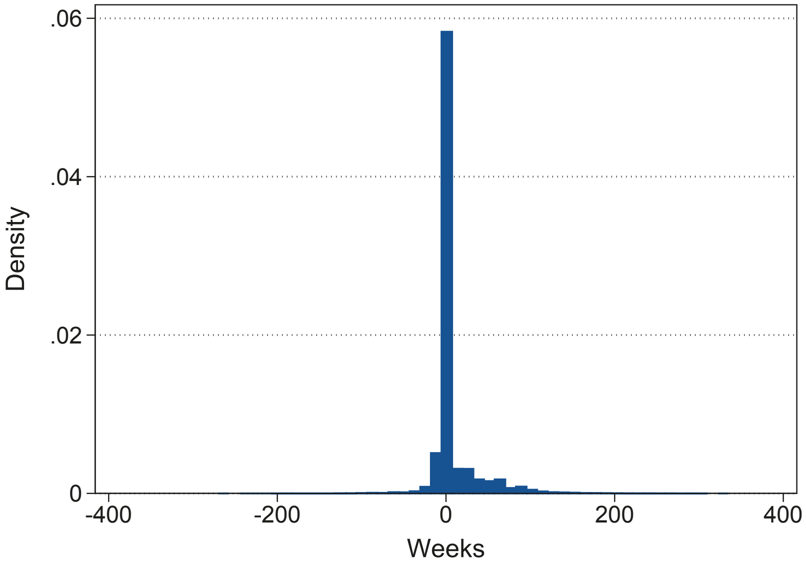


Table A.1. *Robustness analyses for the total reform effects, by gender*

	Baseline (1)	Group-specific seasonality (2)	Control for month post UI exhaustion (3)	Control for month before UI exhaustion (4)	Duration 250-590 (5)	Duration 310-530 (6)
Panel A: Men						
Reform 1: Monitoring and sanctions, UI recipients	0.11* (0.07)	0.09 (0.08)	0.12 (0.08)	0.12* (0.07)	0.10 (0.06)	0.08 (0.08)
Reform 2: Monitoring, AS recipients	0.21*** (0.07)	0.17* (0.10)	0.24*** (0.09)	0.20** (0.08)	0.20*** (0.07)	0.19** (0.08)
No. individuals	25,682	25,682	25,682	25,682	29,475	22,687
Spell duration	✓	✓	✓	✓	✓	✓
Calendar Time FE	✓	✓	✓	✓	✓	✓
Panel B: Women						
Reform 1: Monitoring and sanctions, UI recipients	-0.05 (0.08)	-0.009 (0.10)	-0.08 (0.09)	-0.04 (0.09)	-0.07 (0.08)	-0.03 (0.09)
Reform 2: Monitoring, AS recipients	-0.06 (0.09)	-0.02 (0.12)	-0.06 (0.11)	-0.03 (0.09)	-0.12 (0.09)	-0.05 (0.10)
No. individuals	19,066	19,066	19,066	19,066	22,132	16,566
Spell duration	✓	✓	✓	✓	✓	✓
Calendar Time FE	✓	✓	✓	✓	✓	✓

Notes: Robustness estimates of the main results when splitting the sample by gender. Column 1: baseline results (spells range between 280 and 560 days); Column 2: additional inclusion of group-specific seasonal dummies; Columns 3 and 4: partition out the month following and preceding the 420-day threshold, respectively; Columns 5 and 6: sampling spells ranging in 250-590 and 310-530 duration days, respectively. Robust standard errors in parentheses. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

References

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

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

Arbetsförmedlingen (2014). Ökad sökaktivitet genom tydligare krav och ökad uppföljning (Increased search activity through clearer requirements and increased monitoring). Swedish PES reports 2014.

Arbetsförmedlingen (2017). The Job and development guarantee programme. Swedish PES reports 2017.

Arni, P., Lalive, R., and van Ours, J. C. (2013). How effective are unemployment benefit sanctions? Looking beyond unemployment exit. *Journal of Applied Econometrics*, 28(7):1153–1178.

Becker, G. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, 76(2):169–217.

Bergemann, A. and van den Berg, G. J. (2008). Active labor market policy effects for women in Europe — A survey. *Annales d'Économie et de Statistique*, (91/92):385–408.

Boone, J., Fredriksson, P., Holmlund, B., and van Ours, J. C. (2007). Optimal unemployment insurance with monitoring and sanctions. *The Economic Journal*, 117(518):399–421.

Boone, J., Sadrieh, A., and van Ours, J. C. (2009). Experiments on unemployment benefit sanctions and job search behavior. *European Economic Review*, 53(8):937–951.

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

Caliendo, M. and Schmidl, R. (2016). Youth unemployment and active labor market policies in Europe. *IZA Journal of Labor Policy*, 5(1).

Card, D., Chetty, R., and Weber, A. (2007). The spike at benefit exhaustion: Leaving the unemployment system or starting a new job? *American Economic Review*, 97(2):113–118.

Card, D., Kluve, J., and Weber, A. (2010). Active labour market policy evaluations: A meta-analysis. *The Economic Journal*, 120(548):F452–F477.

Card, D., Kluve, J., and Weber, A. (2017). What works? A meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association*, 16(3):894–931.

Chalfin, A. and McCrary, J. (2017). Criminal deterrence: A review of the literature. *Journal of Economic Literature*, 55(1):5–48.

Charness, G. and Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior and Organization*, 83(1):50–58.

Crépon, B. and van den Berg, G. J. (2016). Active labor market policies. *Annual Review of Economics*, 8:521–546.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–474.

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

Gray, D. (2003). National versus regional financing and management of unemployment and related benefits: The case of Canada. OECD Social, Employment and Migration Working Papers, No. 14.

Grubb, D. (2000). Eligibility criteria for unemployment benefits. OECD Economic Studies, No. 31, 2000/II.

Hawken, A. and Kleiman, M. (2009). Managing drug involved probationers with swift and certain sanctions: Evaluating Hawaii's HOPE. American Psychological Association report.

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

Hofmann, B. (2008). Work incentives? Ex-post effects of unemployment insurance sanctions - Evidence from West Germany. IAB Discussion Paper.

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

IAF (2014a). Arbetsförmedlingens underrättelser om ifrågasatt rätt till arbetslöshetsersättning, lämnade under 2013 och första kvartalet 2014 (Employment service's notifications of disputed right to unemployment-benefit made in 2013 and the first quarter of 2014). Swedish Unemployment Insurance Board report, 2014:21.

IAF (2014b). Arbetslöshetskassornas sanktioner efter underrättelser om ifrågasatt ersättningsrätt (Unemployment insurance funds sanctions following notifications of disputed right to benefit). Swedish Unemployment Insurance Board report, 2014:23.

IAF (2014c). The Swedish unemployment insurance act (amended September 1, 2013). Swedish Unemployment Insurance Board.

Immervoll, H. and Knotz, C. (2018). How demanding are activation requirements for jobseekers. OECD Social, Employment and Migration Working Papers, No. 215.

Kluve, J. (2010). The effectiveness of European active labor market programs. *Labour Economics*, 17(6):904–918.

Lalive, R., van Ours, J. C., and Zweimüller, J. (2005). The effect of benefit sanctions on the duration of unemployment. *Journal of the European Economic Association*, 3(6):1386–1417.

Landais, C., Nekoei, A., Nilsson, P., Seim, D., and Spinnewijn, J. (2017). Risk-based selection in unemployment insurance: Evidence and implications. Working paper.

Liljeberg, L. and Söderström, M. (2017). Hur ofta träffas arbetssökande och arbetsförmedlare? (How often do jobseekers and casworkers meet?). IFAU working paper, 2017:16.

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

Lombardi, S., van den Berg, G. J., and Vikström, J. (2019). Empirical Monte Carlo evidence on estimation of timing-of-events models. Working paper.

McVicar, D. (2014). The impact of monitoring and sanctioning on unemployment exit and job-finding rates. *IZA World of Labor, 2014: 49*.

Müller, K.-U. and Steiner, V. (2008). Imposed benefit sanctions and the unemployment-to-employment transition: The German experience. SSRN Electronic Journal.

Nagin, D. S. (2013a). Deterrence: A review of the evidence by a criminologist for economists. *Annual Review of Economics*, 5(1):83–105.

Nagin, D. S. (2013b). Deterrence in the twenty-first century. *Crime and Justice*, 42(1):199–263.

Røed, K. and Westlie, L. (2012). Unemployment insurance in welfare states: The impacts of soft duration constraints. *Journal of the European Economic Association*, 10(3):518–554.

Svarer, M. (2011). The effect of sanctions on exit from unemployment: Evidence from Denmark. *Economica*, 78(312):751–778.

van den Berg, G. J., Uhlendorff, A., and Wolff, J. (2013). Sanctions for young welfare recipients. IZA Discussion Papers, No. 7630.

van den Berg, G. J. and Vikström, J. (2014). Monitoring job offer decisions, punishments, exit to work, and job quality. *The Scandinavian Journal of Economics*, 116(2):284–334.

van der Klaauw, B., van den Berg, G. J., and van Ours, J. C. (2004). Punitive sanctions and the transition rate from welfare to work. *Journal of Labor Economics*, 22(1):211–241.

van der Klaauw, B. and van Ours, J. C. (2013). Carrot and stick: how re-employment bonuses and benefit sanctions affect exit rates from welfare. *Journal of Applied Econometrics*, 28(2):275–296.

Weisburd, D., Einat, T., and Kowalski, M. (2008). The miracle of the cells: An experimental study of interventions to increase payment of court-ordered financial obligations. *Criminology and Public Policy*, 7(1):9–36.

2. Empirical Monte Carlo Evidence on Estimation of Timing-of-Events Models

with Gerard J. van den Berg and Johan Vikström

Acknowledgments: We thank Paul Muller, Oskar Nordström Skans, seminar participants at IFAU, and conference participants at EEA-ESEM 2018 and EALE 2018 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). Vikström and Lombardi acknowledge support from FORTE.

2.1 Introduction

The Timing-of-Events (ToE) approach focuses on the effect of a treatment given during a spell in some state on the rate of leaving that state, if systematic unobserved confounders cannot be ruled out. To this purpose, Abbring and van den Berg (2003) specify a bivariate Mixed Proportional Hazard (MPH) model and establish conditions under which all parts of the model, including the treatment effect, are non-parametrically identified. The fact that this approach allows for unobserved confounders is one reason for why it has been applied in many settings. An early example is Abbring et al. (2005) on the effect of benefit sanctions on the transition rate out of unemployment, with unobserved factors such as personal motivation potentially affecting both the time to a benefit sanction (treatment) and time in unemployment (outcome). Recent examples include Crépon et al. (2018), Richardson and van den Berg (2013), Caliendo et al. (2016), Busk (2016), Lindeboom et al. (2016), Holm et al. (2017), Bergemann et al. (2017) on labor market policies; Van Ours and Williams (2009, 2012), and McVicar et al. (2018) on cannabis use; van Ours et al. (2013), van den Berg and Gupta (2015), Palali and van Ours (2017) in health settings; Bijwaard et al. (2014) on migration; Jahn and Rosholm (2013) on temporary work; and Baert et al. (2013) on overeducation.

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

In this paper, we use a new simulation design based on actual data to evaluate these and related specification issues for the implementation of the ToE model in practice. To this end, we modify the novel Empirical Monte Carlo design (EMC) proposed by Huber et al. (2013). In their study, they compare different methods to estimate treatment effects under unconfoundedness.[1] The key idea is to use actual data on treated units to simulate placebo treatments for non-treated units and then base the simulations on these placebo treatments. This ensures that the true effect is zero, that the selection model is known, and

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

that the unconfoundedness assumption holds by construction. The fact that real data is used instead of a data generating process chosen by the researcher makes the simulation exercise arguably more relevant for real applications.

Previous EMC implementations study estimators based on conditional independence assumptions. Here, we propose and implement a variant of the basic EMC approach, which allows us to study the ToE model. In our simulation design, we use rich administrative data on Swedish jobseekers, with information on participation in a training program (the treatment). For each jobseeker, we create detailed background information. This is used to estimate a duration model for the time to treatment using data on both treated and non-treated units. We then use the estimated model to simulate placebo treatment durations for each non-treated unit. By construction, the effect of these placebo treatments is zero and the treatment assignment process is known. With the simulated data we estimate various ToE models. Here, the key aspect is that we leave out some of the variables that were used to simulate the placebo treatments. Since the excluded variables were used to generate the placebo treatments, and since they also affect the outcome duration (re-employment rate), we obtain a bivariate duration model with correlated unobserved determinants, i.e. the ToE setting. This new simulation design allows us to use real data to examine a number of ToE-specification issues.[2]

An important question that has been studied for a long time is how to best specify the distribution of unobserved heterogeneity. Initial simulation evidence was provided by Heckman and Singer (1984), Ridder (1987), and Huh and Sickles (1994). More recently, Baker and Melino (2000) study a univariate duration model with unobserved heterogeneity and duration dependence. One conclusion is that model specifications with too many support points over-correct for unobserved heterogeneity (through an overly-dispersed unobserved heterogeneity distribution), which leads to bias in all model components. Gaure et al. (2007) also use simulated data and examine a similar bivariate duration model as the one analyzed in this paper. One finding is that a discrete support points approach is generally reliable if the sample is large and there is some exogenous variation, such as variation due to time-varying covariates. On the other hand, unjustified restrictions – such as pre-specifying an extremely low number of support points for the unobserved heterogeneity – or deviations from the model assumptions, may cause substantial bias.

Our study adds to this evidence by using a simulation design based on actual data. This leads to several conclusions. In the main analyses, we leave out a large number of variables from the model, so that the estimated effect of the placebo treatment is far from the true zero effect, i.e. there is substantial bias.

[2] Recently, Advani et al. (2018) use the LaLonde (1986) data to provide a critical assessment of the internal validity of the EMC simulation design. This critique is rebutted by Huber et al. (2016), who, among other things, stress that the LaLonde data is small in size (hence, the Monte Carlo samples are not drawn from an infinite population) and also contains only a few covariates (hence, the selection process is not well-captured).

However, two support points are already able to eliminate a large share of the bias. We also find a substantial risk of over-correcting for unobserved heterogeneity. With too many support points, the average bias is more than twice as large as with a few support points, and the variance increases in the number of support points. The over-correction problem occurs because too many support points lead to an overly-dispersed distribution of unobserved heterogeneity, and to fit the data this is compensated in the model by bias in the treatment effect and the duration dependence.

Another result is that information criteria are useful for selecting the number of support points. In particular, the Akaike information criterion (AIC), Bayesian information criterion (BIC), and Hannan-Quinn information criterion (HQIC) all perform well. They protect against over-correction by penalizing parameter abundance. They also guard against under-correction by rejecting models without or with only weak correction for unobserved heterogeneity. However, information criteria with little penalty for parameter abundance, such as those solely based on the maximum likelihood (ML criterion), should be avoided altogether. This is because they tend to favor models with too many support points, and this leads to over-correction problems.

We mainly focus on the above-mentioned specification choices, but the simulation results also indicate that the ToE model is generally able to adjust for a significant share of the bias due to unobserved heterogeneity. This holds in our baseline model, where the only source of variation is across cross-sectional units through time-fixed covariates. When we introduce more exogenous variation in the form of time-varying covariates through the unemployment rate in the local labor market (measured at monthly intervals), the bias is further reduced. This holds even when the setting is characterized by substantial heterogeneity induced by omitting a large set of covariates, including a wide range of short- and long-term labor market history variables. The importance of time-varying covariates echoes the results in Gaure et al. (2007).

The results on how to specify the distribution of unobserved heterogeneity are not only relevant for ToE models, but also for all other selection models with random effects, including univariate duration models, general competing risks models,[3] non-parametric maximum likelihood estimators for non-duration outcomes and structural models with unobserved heterogeneity. Univariate duration models with unobserved heterogeneity are for instance used in studies of factors behind duration dependence in aggregate re-employment rates. The latter may be explained by individual-level duration dependence or dynamic sorting of unemployed with low exit probabilities into long-term unemployment (e.g., Abbring et al., 2001). In labor economics, competing risks models are used in studies of unemployment durations with competing

[3] The ToE model is a type of competing risks model where one duration (treatment duration) is assumed to have a causal impact on the other duration (outcome duration). More generally, there are many other competing risks models with related unobserved heterogeneity.

exits to employment and non-employment (e.g., Narendranathan and Stewart, 1993) as well as exits to different types of jobs (Baert et al., 2013; Jahn and Rosholm, 2013). In health economics and epidemiology, two often studied competing risks are disease relapse and death (e.g., Gooley et al., 1999). Non-parametric maximum likelihood estimators have also been extensively used when modelling non-duration outcomes. One example is consumer choice analysis (Briesch et al., 2010) and univariate or multinomial choice models with unobserved determinants (Ichimura and Thompson, 1998; Fox et al., 2012; Gautier and Kitamura, 2013).

Another important contribution of our paper is that we evaluate the relevance of different sets of covariates when measuring causal effects of active labor market programs. This is relevant for evaluations based on conditional independence (CIA) assumptions, but it is also relevant for identification strategies that allow for unobserved heterogeneity, as we help to characterize the unobserved heterogeneity that needs to be taken into account. The relevance of different covariates has been an important topic ever since discussions based on the seminal work by LaLonde (1986). Even though LaLonde mainly focused on the performance of different non-experimental methods, the ensuing discussions using the LaLonde data also focus on the availability of variables in the data and its implications for the performance of non-experimental methods (Dehejia and Wahba, 1999, 2002; Smith and Todd, 2005). Several other studies also use survey data from experimental designs to assess non-experimental methods, which has led to many important conclusions on the relevance of different set of covariates.[4] Other studies have used novel survey data to assess the importance of usually unobserved characteristics such as personality traits.[5]

In a related study, Lechner and Wunsch (2013) use EMC methods and real data to examine the relevance of different covariates. They use detailed data on unemployed jobseekers in Germany and analyze job search assistance and training programs. Their idea is to use essentially all variables that are important for the selection process and used in various CIA-based evaluations of active labor market programs. With these data they perform simulations in a similar way as Huber et al. (2013), i.e. they simulate placebo treatments for the non-treated using the full data. Then, to assess the relative importance of different variables, they leave out alternative blocks of covariates, re-estimate

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

[5] Caliendo et al. (2017) study the relevance of measures of personality traits, attitudes, expectations, social networks and intergenerational information. They find that such usually unobserved 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.

the model, and compare the size of the bias (non-zero treatment effect) across specifications.

We use our EMC-simulated data in a similar way. We use the Swedish data to construct analogous variables as for the German setting in Lechner and Wunsch (2013). This allows us to examine to what extent the results in Lechner and Wunsch (2013) carry-over to other countries and programs. However, we also include additional covariates not used by Lechner and Wunsch (2013). First, since we model treatment durations and not binary treatment indicators, we also include previous employment and unemployment durations in the set of covariates. This is because previous durations may capture aspects related to how long one stays unemployed in a better way than non-duration history variables such as the employment rate over a certain time period. 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 this end, we use parental income, which is a commonly used proxy for such general unobserved skills. Third, since we model the treatment duration, time-varying covariates, such as local business cycle conditions, may play a role, especially for longer unemployment spells. Another difference compared to Lechner and Wunsch (2013) is that we consider a duration outcome framework.

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

This paper is also related to Muller et al. (2017), who compare the ToE approach, CIA-based matching methods and a quasi-experimental benchmark using a policy discontinuity. For all three methods, they use the same data to to evaluate a Dutch job-search assistance program, and find that the alternative approaches yield similar results. Our evaluation of the ToE approach is very different. We focus on the specification of the distribution of unobserved heterogeneity, other specification choices and the importance of different covariates.

The paper proceeds as follows. Section 2 presents the Timing-of-Events model proposed by Abbring and van den Berg (2003). Section 3 describes the simulation design and the data used in the simulations. 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. In Section 5, we present the EMC simulation results, and Section 6 concludes.

2.2 The Timing-of-Events model

This section presents the ToE approach as introduced by Abbring and van den Berg (2003). They specify a bivariate duration model for the duration in an initial state and the duration until the treatment of interest: T_e and T_p, with t_e and t_p being their realizations. The model includes individual characteristics, X, and unobserved individual characteristics V_e and V_p, with realizations (x, v_e, v_p). Abbring and van den Berg (2003) assume that the exit rate from the initial state, $\theta_e(t|D(t),x,V_e)$, and the treatment rate, $\theta_p(t|x,V_p)$, follow the Mixed Proportional Hazard (MPH) form:[6]

$$\begin{aligned} \ln\theta_e(t|x,D,V_e,t_p) &= \ln\lambda_e(t)+x'\beta_e+\delta D(t)+V_e, \\ \ln\theta_p(t|x,V_p) &= \ln\lambda_p(t)+x'\beta_p+V_p, \end{aligned} \tag{2.1}$$

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

Abbring and Van den Berg (2003) show that all components of this model, including the treatment effect, δ, and the unobserved heterogeneity distribution, G, are identified under the following assumptions. The first assumption is no-anticipation, which means that future treatments are not allowed to affect current outcomes. This holds if the units do not know the exact time of the treatment or if they do not react on such information.[7] A second assumption is that X and V should be independently distributed, implying that the observed characteristics are uncorrelated with the unobserved characteristics. A third assumption is the proportional hazard structure (MPH model). We discuss these assumptions in more detail when we describe our simulation design. Abbring and van den Berg (2003) also impose several regularity conditions.

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

[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.

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

nous variation is generated through the time-invariant characteristics, x, which create variation in the hazard rates across units. Strictly speaking, this is the only variation that is needed for identification.

Previous studies suggest that time-varying variation, i.e. variables that change with the elapsed duration, for instance due to business cycle variation or seasonal variation, is a useful and more robust source of additional exogenous variation (Eberwein et al., 1997; Gaure et al., 2007). The intuition is that such time-varying covariates shift the hazard rates, and this helps to identify the influences of the unobserved heterogeneity. More specifically, current factors have an immediate impact on the exit rate, whereas past factors affect the current transition probabilities only through the selection process (for a more detailed discussion, see van den Berg and van Ours, 1994, 1996). We therefore examine both ToE models with only time-invariant covariates and models with time-varying covariates.

2.3 Simulation approach

2.3.1 The basic idea

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

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

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

One important reason to use the Swedish unemployment spell data is that there are many examples of evaluations that estimate ToE models using this type of data.[8] The use of unemployment spells also affects how we design our simulation study. Unemployment durations and labor market program entries are typically measured at the daily level, which is also the case in our setting. 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.[9]

Next, let us relate our simulated data to the assumptions made in the ToE approach. By construction, the no-anticipation assumption holds, because the units cannot anticipate and react to placebo treatments. However, there are

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

[9] Continuous-time models are often estimated in the literature, even when using discrete data (daily, weekly, monthly or yearly). For instance, this is the case for Palali and van Ours (2017), Tatsiramos (2010), Jahn and Rosholm (2013), Kyyrä et al. (2013), McVicar et al. (2018), Muller et al. (2017), van Ours and Williams (2009), van Ours and Williams (2012), and van Ours et al. (2013).

other ToE assumptions that may not hold in this simulation design. First, the assumption requiring independence between X and V (random effects assumption) may not hold in our simulations, since the excluded variables representing unobserved heterogeneity may be correlated with the variables that were actually used in the ToE estimation. To explore this, in extended simulations we estimate ToE models when leaving out blocks of variables that are alternatively highly or mildly correlated with the observables. It turns out that the degree of correlation between the observed and unobserved factors is relatively unimportant. Second, since the outcome duration is not modeled, the outcome hazard (re-employment rate) may not follow the MPH structure. Third, a duration model without embedded unobserved heterogeneity is used to model the treatment selection process. This means that although we use an extremely rich set of variables to estimate the selection process, if there are some omitted characteristics, the model will be misspecified.

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

2.3.2 The relevance of different covariates

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

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

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

Note that this procedure holds under the assumption of CIA with the full set of covariates. Lechner and Wunsch (2013) provide good arguments as to why CIA should be valid in their German setting when they use their full set of covariates, and Vikström (2017) provides similar arguments for Sweden. This can of course always be questioned, for instance because treatment selection is based on unobserved motivation and skills. Thus, we study the relevance of the different observed covariates, keeping in mind that there may also be important information that is not included in our data.

2.3.3 The training program

One often-studied treatment for jobseekers is labor market training. This motivates our use of data on a Swedish vocational training program called AMU (Arbetsmarknadsutbildning). The program and the type of administrative data that we use resemble those of other countries. The main purpose of the program, which typically lasts for around 6 months, is to improve the skills of the jobseekers so as to enhance their chances of finding a job. Training courses include manufacturing, machine operator, office/warehouse work, health care, and computer skills. The basic eligibility criterion is to be at least 25 years old. During the training, participants receive a grant. Those who are entitled to unemployment insurance (UI) receive a grant equal to their UI benefits level, while for those not entitled to UI the grant is smaller. In all cases, training is free of charge.

Previous evaluations of the AMU training program include Harkman and Johansson (1999), de Luna et al. (2008), Richardson and van den Berg (2013), and Vikström and van den Berg (2017). These papers also describe the training program in more detail.

2.3.4 Data sources and sampling

We combine data from several administrative registers and surveys. The Swedish Public Employment Service provides daily unemployment and labor market program records of all unemployed in Sweden. We use this information to construct spell data on the treatment duration (time to the training program) and the outcome duration (time to employment), both measured in days. We sample all unemployment spells starting during the period of 2002–2011.[10] 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 jobseeker enters into training multiple times, only the first instance is considered.

For each spell, we construct detailed information on individual-level characteristics. We start by constructing similar covariates as in the German data in Lechner and Wunsch (2013).[11] The population register LOUISE provides basic socio-economic information, such as country of origin, civil status, regional indicators and level of education. Matched employer-employee data (RAMS) and wage statistics from Statistics Sweden are used to construct information on the characteristics of the last job (wages, type of occupation, skill-level), and to retrieve information on the characteristics of the last firm (firm size, industry and average worker characteristics). From Unemployment Insurance (UI) records we obtain information on UI eligibility.

Data from the Public Employment Service is used to construct unemployment history variables. It is also used to construct information on the regional unemployment rate. Earnings records and information on welfare participation are used to construct employment, out-of-labor force and earnings histories. For the history variables, we construct both short-run history (last two years) and more long-run history (last ten years). Altogether, this captures many aspects of the workers employment and earnings history in the last two or ten years.

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

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

[11] 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.

covariates, and include the local unemployment rate in the region during each month as a time-varying covariate (Sweden has 21 regions).

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

All covariates that are used in the analyses are summarized in Table A.1. The statistics in the table show that immigrants from outside Europe, males, married and the less educated jobseekers are overrepresented among the training participants. Training participants also also more likely to be employed in firms with lower wages, and there are fewer previous managers and more mechanical workers among the treated workers. All labor market history measures point in the same direction: training participants have worse unemployment and welfare characteristics in the last two and ten years.

2.3.5 Simulation details

Selection model

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

$$\theta_p(t|x) = \lambda_p(t) \cdot \exp(x\beta_p). \tag{2.2}$$

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 A.1. The model estimates, also reported in Table A.1, show that the daily treatment rate peaks after roughly 300 days. They also confirm the same patterns found for the sample statistics: immigrants, younger workers, males, high-school graduates, and UI recipients are more likely to be treated. Short- and long-term unemployment and employment history variables are also important determinants of treatment assignment.

After estimating the selection model by using the full population of actual treated and controls (i.e. the never treated), the treated units are discarded and play no further role in the simulations. Next, we use (2.2) to simulate the placebo times to treatment for each non-treated, T_s, which is generated according to (dropping x to simplify the notation):

$$\exp\left(-\int_0^{T_p} \theta_p(\tau)d\tau\right) = U, \tag{2.3}$$

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 (2.3), we can retrieve T_p for each observation.[12]

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

Simulations

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

2.3.6 Implementation of the bivariate duration model

We estimate a continuous-time ToE model for the treatment and outcome hazards as defined in Equation (2.1). The unknown distribution of the unobserved heterogeneity is approximated by a discrete support points distribution (Lindsay, 1983; Heckman and Singer, 1984; Gaure et al., 2007).

Likelihood function

For each unit $i = 1, \ldots, N$ we formulate the conditional likelihood contribution, $L_i(v)$, conditional on the vector of unobserved variables $v = (v_e, v_p)$. Then, the individual likelihood contribution, L_i, is obtained by integrating $L_i(v)$ over the distribution of the unobserved heterogeneity, $G(V)$. For the

[12] The actual distribution for the integrated hazard will depend on the specification of the selection model (2.2). In the simple case where all covariates are time-fixed and the placebo treatments are generated by using a proportional hazard model that has two piecewise constant parts, with θ_s^0 for $t \in [0, t_1)$ and θ_s^1 for $t > t_1$:

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

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

duration dependence ($\lambda_e(t)$, $\lambda_p(t)$), we use a piecewise constant specification with $\lambda_s(t) = \exp(\alpha_{sm})$ where the spell-duration indicators are $\alpha_{sm} = \mathbb{1}\left[t \in [t_{m-1}, t_m)\right]$, for $m = 1, \ldots, M$ cut-offs. We fix the cut-offs to 31, 61, 122, 183, 244, 365, 548, 2160. The actual observed variables used in the model are explained in the next section.

To set up $L_i(v)$, we split the spells into parts where all right-hand side variables in (2.1) are constant. Splits occur at each new spell-duration indicator and when the treatment status changes. In all baseline ToE specifications, the covariates specified are calendar-time constant. In additional specifications where the time-varying local unemployment rate is included, calendar-time variation leads to additional (monthly) splits. Spell part j for unit i is denoted by c_{ij}, and has length l_{ij}. Let C_i be the set of spell parts for unit i. Each part, c_{ij}, is fully described in terms of l_{ij}, α_{sm}, x_i and the outcome indicator, y_{sij}, which equals one if the spell part ends with a transition to state s and zero otherwise. There are two such possible states (job exit and treatment start). Then, with approximately continuous durations, $L_i(v)$ is:

$$L_i(v) = \prod_{c_{ij} \in C_i} \left[\exp\left(-l_{ij} \sum_{s \in S_{it}} \theta_s(t, x_i, D_{it}, v_s|\cdot) \right) \times \prod_{s \in S_{it}} \theta_s(t|\cdot)^{y_{sij}} \right], \tag{2.4}$$

with

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

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

$$\mathcal{L} = \sum_{i=1}^{N} \left(\sum_{w=1}^{W} p_w \ln L_i(v_w) \right) \equiv \sum_{i=1}^{N} L_i. \tag{2.5}$$

Search algorithm

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

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

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

Step 3: Perform ML maximization with respect to the full parameters vector $\vartheta_W = (\beta, v, p)$ by using $\hat{\vartheta}_{W-1}$ and $(\tilde{v}_W, \tilde{p}_W)$ as initial values. Return $\hat{\vartheta}_W$.

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

Step 1 corresponds to a model without unobserved heterogeneity, since $\hat{v}$ cannot be distinguished from the intercept in X. In *Step 2* the algorithm searches for a new support point in the $[-3, 3]$ interval.[13] In this step, all other parameters of the model are fixed. This explains why in *Step 3* we perform a ML maximization over all parameters, including the new support point. At the end of the procedure we obtain $\overline{W}$ maximum likelihood estimates: $\{\hat{\vartheta}_W,\, \mathcal{L}(\hat{\vartheta}_W)\}_{W=1}^{\overline{W}}$.

Information criteria

We use different approaches to choose between the $\overline{W}$ estimates. First, we report results where we pre-specify the number of support points (up to six points). An alternative approach is to increase the number of support points until there is no further improvement in the likelihood (ML criterion).

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.[14] The ML criterion is defined as $ML = \mathcal{L}(\hat{\vartheta}_W)$, where only likelihood increases greater than 0.01 are considered. The criteria are calculated for each replication, so that the selected number of support points may vary both across replications and criteria. This allows us to compute the average bias and the mean square error for all information criteria.

2.4 Available covariates and evaluations of ALMPs

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

[13] 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.

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

Table 1. *Estimated treatment effect bias when controlling for different covariates*

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

Notes: Estimated biases using the full sample of placebo treated and non-treated with control for for different blocks of covariates. The number of observations is 2,564,561. Hazard rate estimates for time in unemployment using a parametric proportional hazard model with piecewise constant baseline hazard (8 splits). *, ** and *** denote significance at the 10, 5 and 1 percent levels.

sample of placebo treated and placebo non-treated units is used to estimate a parametric proportional hazard (PH) model. Here, the baseline hazard is specified in the same way as for the model used to simulate the placebo treatments.[15] Table A.1 lists all covariates in each block.

The main results are given in Table 1. In each panel of the table, we start with the covariates from the proceeding panels and add additional information to the covariates already in the model, so that the model is extended sequentially by adding blocks of covariates one by one. We add the covariates in a similar order as Lechner and Wunsch (2013), who argue that the order resembles the ease, likelihood and cost of obtaining the respective information. This will, for instance, reveal the relevance of adding information on long-term labor market history on top of the more basic covariates such as short-term history and baseline socio-economic characteristics.

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

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

Table 2 examines individual short-term employment and unemployment variables. They are added in addition to the baseline covariates. The aim

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

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

Table 2. *Estimated bias of the treatment effect when controlling for different short-term labor market history variables*

	Est.	Std. Err.
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 unempl. if in last 24 months	0.0616***	(0.00243)
No. 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.

is to understand what specific aspects of employment and unemployment that are the most important to adjust for. In the comparisons, we control for either past employment duration, different measures of the share of time spent in employment (employment rate), employment status at a given point in time, or other history variables. A reason for this exercise is that we model treatment durations and not a binary treatment status. Accordingly, it may be the case that previous durations capture aspects of the ongoing unemployment spell in a better way than previous employment rates and employment status at a given point in time. We also compare the relevance of similarly constructed short-term unemployment history variables.

The results show that information on previous employment duration reduces the bias considerably: from 6.2% in the baseline specification to 3.9% (Panel A). However, adding information on past employment rates or other short-term employment history variables reduces the bias even more, leading to biases of -0.04% and 0.2%, respectively (Panel B and C). In particular, Panel B shows that all covariates measuring past employment rate single-handedly capture a large part of the bias. For instance, by only adjusting for months employed in the last six months before the unemployment spell, the bias reduces from 6.2% to 1.7%. Panel C also shows that employment status one year before the unemployment spell has a similar impact on the bias. On the other hand, employment status two years before the spell and other short-term employment variables appear to be less important. Interestingly, the bias is positive or close to zero in all cases, so that the reversal of the bias sign that was observed in Panel B of Table 1 occurs only once all short-term employment variables are included together. That is, even if some short-term history variables are more relevant, they all capture different aspects of the selection process, so that adjusting for both previous employment durations and rates is important.

Panels D to F of Table 2 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. For instance, while adjusting for previous employment duration reduced the bias from 6.2% to 3.9%, now including previous unemployment duration only reduces the bias from 6.2% to 5.5%. All in all, this suggests that for training programs with emphasis on human capital accumulation, the most important characteristics to control for are those related to employment history.

Next, we 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 minor impact on the bias of the estimated treatment effect. The same holds when in Panel D we adjust for various characteristics of the last job (e.g., previous wage and occupation) as well as for detailed information about the last firm (e.g., industry and composition of worker). Lechner and Wunsch (2013) also find that, after controlling for calendar time, regional conditions and short-term labor market history, adding additional covariates such as long-term labor market history is relatively unimportant. This is also consistent with the results in Heckman et al. (1998), Heckman and Smith (1999), Mueser et al. (2007) and Dolton and Smith (2010), who find that it is important to control for regional information, labor market history and pre-treatment outcomes. However, one difference compared to Lechner and Wunsch (2013) is that in this setting adjusting for short-term employment history is enough to obtain small bias, whereas Lechner and Wunsch (2013) find that it is important to also adjust for all aspects of the short-term history (employment, unemployment, out-of-labor-force status, earnings) to obtain a low bias.

Finally, Panel D examines the relevance of parental income, the idea being that father's and mother's income proxy for more general unobserved skills. This may be important if unobserved skills are not captured by the covariates discussed so far, which are mainly related to labor market attachment. However, parents' income turns out to have limited impact on the bias, at least once we control for both short- and long-term labor market history variables. This indicates that labor market histories are also able to capture more general unobserved skills.[17]

2.5 Specification of ToE models

This section presents the main simulation results. The focus here 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. Results are presented in the form of average bias, variance of the placebo estimates, and mean squared error (MSE).

2.5.1 Baseline results

Table 3 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–B, Table A.1).[18] 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.

[17] This confirms the results in Caliendo et al. (2017), who find that once one controls for rich observables of the type that we include here, additional (usually unobserved) characteristics measuring personality traits and preferences become redundant.

[18] Here, we control for time-fixed regional unemployment rate (measured as the month of inflow into unemployment). In Table 7, we estimate ToE models where this covariate varies on a monthly basis.

The first row shows that the baseline model 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. In fact, with six support points the average bias (6.4%) is more than twice as large as the average bias with two support points (2.7%). Moreover, both the variance and the MSE increase in the number of support points (Columns 2–3).

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

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

Interestingly, the problem with over-correcting for unobserved heterogeneity does not occur to the same extent in the simulated data used by Gaure et al. (2007). They highlight that the main problem is under-correction with too few

[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.

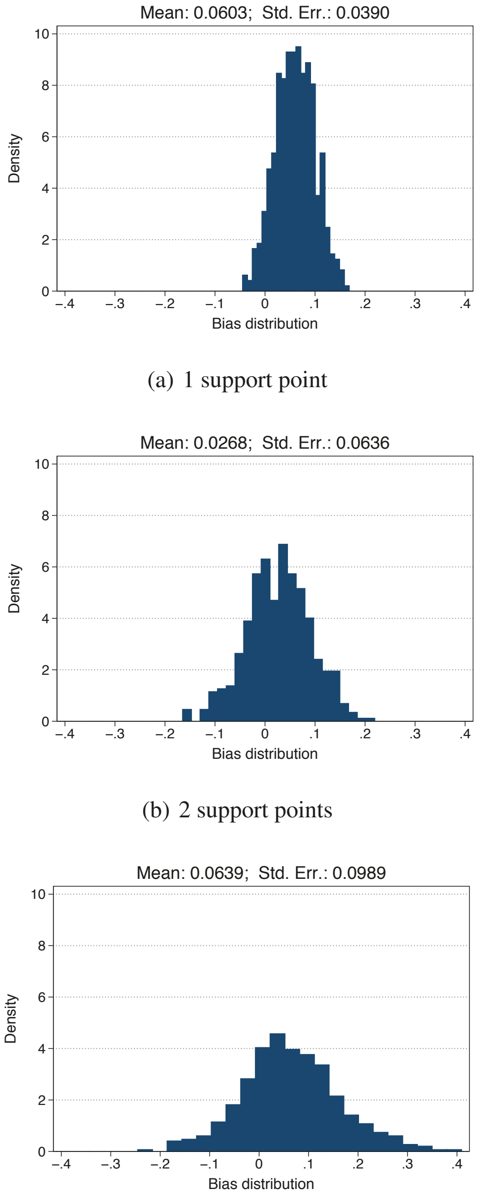
[20] Here, we focus on the bias of the treatment effect, but previous simulation studies using simulated data show that failing to account for unobserved heterogeneity also leads to biased spell-duration and covariate effects (Gaure et al., 2007).

Table 3. *Bias and variance of the estimated treatment effect for a pre-specified number of support points and support points according to model selection criteria*

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

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

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



(a) 1 support point

(b) 2 support points

(c) 6 support points

Note: Distribution of the estimated bias of the treatment effect from a ToE model with different specifications of the discrete support point distribution. Simulations using 500 replications with 10,000 random drawings from the full sample of placebo treated and placebo non-treated. Hazard rate estimates for time in unemployment. Each model uses a piecewise constant baseline hazard (8 splits) and includes socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate.

support points.[21] Our simulation results that are based on real data, instead, suggest that both under- and over-correction are important problems when estimating ToE models. Thus, finding a way to select the appropriate number of support points appears to be important.

2.5.2 Information criteria

Panel B of Table 3 provides simulation results when the distribution of the unobserved heterogeneity (number of support points) is specified by using alternative information criteria. Panel C reports the average number of support points that are selected according to each criterion. The ML criterion, where the number of support points is increased as long as the likelihood is improved, leads to 4.11 support points on average. The bias and variance are large compared to simply pre-specifying two or three support points. Hence, the ML criterion tends to select too many support points, leading to an over-correction problem (too many spurious support points are included). As a result, criteria with little penalty for parameter abundance, such as the ML criterion, should be avoided altogether.

The results for AIC, BIC and HQIC are much more encouraging. All three criteria produce models with rather few unobserved heterogeneity support points (often two support points). In this setting, this corresponds to the specifications with the lowest bias achieved when pre-specifying a low number of support points. We conclude that these more restrictive information criteria protect against over-correction problems due to too many support points. They do so by penalizing the number of parameters in the discrete heterogeneity distribution. They also guard against under-correction problems (too few support points) by favoring models with unobserved heterogeneity over models without unobserved heterogeneity (one support point).

A comparison between the AIC, BIC and HQIC criteria reveals rather small differences. As expected, the two more restrictive information criteria (BIC and HQIC) lead to models with fewer support points, and the average bias is slightly lower than for the less restrictive AIC criterion. The variance is also slightly lower for BIC and HQIC than for AIC. This is because these more restrictive criteria tend to select fewer support points and the variance of the estimated treatment effects is increasing in the number of support points. However, later we will see that none of the three criteria is superior in all settings. All three penalize parameter abundance, and this protects against problems of over-correction due to spurious support points. In some cases, the risk of under-correcting is relatively more important, and this favors the less restrictive AIC criterion. In other cases, the opposite holds, and this favors

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

the more restrictive BIC and HQIC criteria. Thus, using all three criteria and reporting several estimates as robustness check appears to be a reasonable approach.

The main interest here is in providing background information on the alternative specification choices. However, Table 3 also provides some insights on the overall idea of using ToE models to adjust for unobserved heterogeneity. In general, the table shows that the ToE approach corrects for a large share of the bias, which is reduced from 6.0% for the model without unobserved heterogeneity to around 2.7% when information criteria are used to select the number of support points (see Column 1 of Table 3). This holds even though the only source of exogenous variation derives from time-fixed observed covariates. In subsequent analyses, we explore if additional sources of exogenous variation in the form of time-varying covariates can reduce the bias even more.

2.5.3 Sample size

In Columns 4–6 and 7–9 of Table 3, the sample size is increased to 40,000 and 160,000 observations, respectively. For both these sample sizes we see that two support points are associated with the lowest bias, but here the increase in the bias after three support points is smaller than for 10,000 observations. For instance, with 10,000 observations, going from two to six support points increases the bias from 2.7% to 6.4%, and with 40,000 observations, it increases from 2.2% to 3.7%. For the largest sample with 160,000 observations, the increase in the bias when going from twp to six support points is even smaller. This suggests that over-correction due to too many support points tends to be a problem with small samples. However, note that what constitutes a small sample size most likely differs across applications. For instance, it might be related to the number of parameters in the model, the fraction of treated units, the number of exit states, and the variation in the observed variables.

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

2.5.4 Excluded covariates

We next vary the unobserved heterogeneity by excluding different sets of covariates when estimating the ToE models. In the baseline simulations, the ToE model includes baseline socio-economic characteristics, inflow time dummies and regional information. Here, we generate more unobserved heterogeneity by excluding additional covariates (all the socio-economic characteristics in Panel A of Table 1) and less heterogeneity by excluding fewer covariates (earnings history in Panel F of Table 1). Table 2 shows that these models generate a bias of 9.5% and 4.0%, respectively, in the full sample of placebo treated and controls (Panels A and B). These values can be compared to the bias of 6.2% in the baseline setting.

Columns 1–3 of Table 4 report the results for the model with more extensive unobserved heterogeneity. Again, the ToE model adjusts for a large share of the bias due to unobserved heterogeneity. For instance, with a sample size of 10,000, the bias for the specification without unobserved heterogeneity is 9.4%, but it drops to 2–3% when we adjust for unobserved heterogeneity using the AIC, BIC or HQIC criteria (Panel A). As before, these more restrictive criteria return the lowest bias, whereas the ML criterion leads to a model with too many support points.[22] Again, this is consistent with previous results. It confirms that it is important to use an appropriate information criterion to select the number of support points, because this avoids problems with over-correction due to spurious support points.

Overall, the specification with less substantial unobserved heterogeneity, obtained by excluding fewer covariates, produces similar patterns (Columns 4–6 of Table 4). The main difference concerns the relative performance of the AIC, BIC and HQIC criteria. Consider the results for a sample size of 40,000. With more extensive unobserved heterogeneity (Columns 1–3), the bias for the AIC criterion is 0.9%, whereas it is 1.8% and 1.9% for the BIC and HQIC criteria, respectively. This suggests that the more restrictive information criteria (BIC and HQIC) may under-correct for unobserved heterogeneity by favoring models with too few support points, and this leads to larger bias. This pattern is reversed when we create less substantial unobserved heterogeneity by excluding fewer covariates (Columns 4–6). Here, the average bias is lower for the more restrictive BIC and HQIC criteria than for AIC. This is because for this specification, there likely is a larger risk of over-correcting for unobserved heterogeneity, which leads to better bias performance for the criteria with a larger penalty for parameter abundance. From this, we conclude that neither one of the information criteria is superior in all settings.

2.5.5 Degree of correlation between X and V

Since we use single-spell data, identification of the ToE model requires independence between the included covariates and the unobserved heterogeneity (random effects assumption). This may not hold in our setting, because we create unobserved heterogeneity by leaving out certain blocks of covariates, and these excluded covariates may be correlated with those that we include when we estimate the ToE model. We therefore perform additional simulation exercises leaving out different blocks covariates from the model. We consider

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

Table 4. *Bias and variance of the estimated treatment effect when* excluding different sets of covariates*, by model selection criteria and sample size*

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

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

Table 5. *Bias and variance of the estimated treatment effect when adding to the baseline model* covariates more or less correlated *with those left in the error term*

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

Notes: The three model specifications correspond to the baseline model of Table 3 augmented with Welfare benefit history (last 2 years), Previous firm most common occupation dummies and Last occupation dummies, for the positive correlation, small positive correlation and negative correlation specifications, respectively. Correlation coefficients computed from the outcome model using all actual treated and control units, by correlating the linear predictor of the covariates included in the model with the linear predictor of all covariates left in the error term. Estimated bias, variance and mean squared error of the treatment effect from a ToE model with different specifications of the discrete support point distribution. Simulations set as for Table 3.

three settings with strongly positive, mildly positive and negative correlation between the covariates used in the ToE model and the excluded covariates.[23] We select covariates to include in the model so 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 5 shows the simulation results with samples of size 10,000. It shows that the information criteria perform similarly as before. The ML criterion selects a larger number of support points which leads to larger bias, and the AIC, BIC and HQIC criteria select more parsimonious models characterized by lower bias than for the ML criterion. Importantly, this holds regardless of the degree of correlation between the observed and the unobserved variables. It holds with a strong positive correlation (Columns 1–3), mildly positive correlation (Columns 4–6) and negative correlation (Columns 7–9). This is reassuring: even when the variables left out from the model are largely related with those left in the ToE model, the relative performance of the information criteria does not appear to be affected. We obtain similar results when drawing samples of size 40,000 (Panel B of Table 5).

2.5.6 Estimation of the unobserved heterogeneity

So far we have focused on the treatment effect. The overall performance of the ToE model can be also checked by inspecting to what extent the estimated discrete distributions for the unobserved heterogeneity approximates the true one. To 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 create it by leaving out certain blocks of covariates. However, since we do not simulate the outcome durations, the exact composition of V_e is unknown.

Specifically, for each actual treated and control unit, we use the coefficients of the estimated selection model reported in Table A.1 to compute the linear predictor of the variables left out from the model. This linear predictor corresponds to V_p in the model. We compare the first two moments with the corresponding moments for the estimated unobserved heterogeneity from the ToE models (with samples of size 10,000). We include the estimated constant in the linear predictor, which leads to relatively small values of both true and approximated $\exp(V_p)$.

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

Table 6. *Comparison between the actual and the estimated distribution of the unobserved heterogeneity for the treatment duration*

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

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

The results from this exercise are shown in Table 6. The table reports results for the true unobserved heterogeneity (Panel A) and the estimated unobserved heterogeneity (Panels B–C). Panel B shows that larger numbers of support points tend to overestimate the dispersion of the unobserved heterogeneity. On the other hand, the mean of the unobserved heterogeneity distribution tends to be slightly underestimated, regardless of the number of support points chosen. Panel C indicates that the ML criterion returns an unobserved heterogeneity with too large variance when compared to the true variance, whereas for the more restrictive information criteria (AIC, BIC and HQIC) the variance is too small. However, overall, the ToE model appears to approximate well the true underlying unobserved heterogeneity distribution of the selection model.[24]

2.5.7 Exogenous variation with time-varying covariates

Identification of the ToE model requires variation in the observed exogenous covariates, which is needed to produce exogenous changes in the hazard rates.

[24] 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.

This was the only source of exogenous variation exploited in the baseline simulations above. It resulted in several insights on how to specify the unobserved heterogeneity distribution when estimating ToE models. Overall, we found that the ToE model was able to adjust for a large part of the selection due to unobserved heterogeneity, but it did not eliminate the bias entirely. For this reason, we now consider an additional source of identification in the form of time-varying covariates (local unemployment rate). The idea is that time-varying covariates should be useful for identification since they generate shifts in the hazard rates that help to recover the distribution of the unobserved heterogeneity. Specifically, the time-varying covariate used is time-varying unemployment rate measured at the monthly level for each county (län). We refer to it as local unemployment rate. This time-varying covariate was included in the selection model to simulate the placebo treatments.[25] Here, the samples are of size 10,000.

The results from this exercise are presented in Table 7. The first row of Panel A shows that the bias without adjusting for unobserved heterogeneity (one support point) is 5.6%. As before, additional support points are then stepwise included (Panel A). The results confirm what was found in the baseline simulations. First, if we under-correct for unobserved heterogeneity (no unobserved heterogeneity) this leads to sizable bias; if we over-correct for unobserved heterogeneity the bias is also large. Second, the ML criterion tends to select models with an overly-dispersed unobserved heterogeneity distribution, which is associated with large bias. Third, the three criteria that penalize parameter abundance (AIC, BIC and HQIC) all perform well, since they lead to models characterized by low bias.

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

[25] On the other hand, the (time-fixed) local unemployment rate measured at the inflow month was included among the covariates throughout the main analyses with only time-invariant covariates.

Table 7. *Bias and variance of the estimated treatment effect with* time-varying local unemployment rate*, by model selection criteria and sample size*

Specification	Time-varying unemployment rate		
	Bias (1)	SE (2)	MSE (3)
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	

Notes: Simulations with 10,000 observations. Estimated bias, variance and mean squared error of the treatment effect from a ToE model with different specifications of the discrete support point distribution. Simulations using 500 replications with random drawings from the full sample with placebo treated and placebo non-treated. Hazard rate estimates for time in unemployment. Each model uses a piecewise constant baseline hazard (8 splits). The ToE model also includes baseline socio-economic characteristics, inflow year dummies, regional indicators and local unemployment rate.

2.6 Conclusions

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

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

The model is also, in general, able to approximate well the true underlying unobserved heterogeneity distribution of the treatment equation. Another key conclusion is that exogenous variation in the form of time-varying covariates (local unemployment rate) is a useful source of identification. This result holds even though ToE models that only rely on variation in the observed covariates also tend to produce good results, as long as an appropriate information criterion is used.

The fact that the unobserved heterogeneity is generated based on realistic simulations has allowed us to inspect which covariates are important confounders that needs to be controlled for when estimating models for the selection into treatment. In this case, the main conclusion is that it is important to adjust for short-term labor market histories when evaluating labor market programs for jobseekers, whereas adding long-term labor market histories appears to be less important. This is consistent with the results in Lechner and Wunsch (2013) in a German setting. We also find that, in general, controlling for short-term employment histories appears to be more effective than controlling for short-term unemployment histories. In particular, we show that adjusting for variables measuring the share of time spent in employment in the near past is able to reduce the bias to a large extent. Other types of short-term employment history variables, such as previous employment durations, also turn out to be important, but relatively less so.

Appendix

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

	Treated	Control	Selection model	
			Est.	Std. Err.
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.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: College or higher	0.22	0.31	-0.0490**	(0.0250)
Panel B: Inflow time and regional information				
Beginning unempl.: 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 unempl. 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 empl. 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 empl. in last 24 mos.	2.31	2.42	-0.0069***	(0.0015)
No. 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)

Continue to next page

Table A.1 – continued from previous page

	Treated	Control	Selection model	
			Est.	Std. Err.
Panel D: Short–term unemployment history (2 years) and unemployment duration				
Time unempl. in last spell	107.11	89.43	0.0000	(0.0000)
Missing time unempl. 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. in last 24m	15.12	14.76	0.0001	(0.0001)
No. unempl. spells in last 24 mos.	0.82	0.88	0.0033	(0.0060)
Unemployed 6m before	0.20	0.16	0.0171	(0.0151)
Unemployed 24m before	0.24	0.22	-0.0327***	(0.0121)
Any program in last 24 months	0.03	0.02	0.0579**	(0.0291)
Panel E: Short–term welfare history (2 years)				
Welfare benefits -1 year	4928.00	3742.27	0.0318***	(0.0078)
Welfare benefits -2 years	4258.73	3542.66	0.0075	(0.0095)
On welfare benefits -1 year	0.19	0.14	0.0028	(0.0166)
On welfare benefits -2 years	0.17	0.14	-0.0720***	(0.0163)
Panel F: Earnings history (2 years)				
Earnings 1 year before	111684.78	110247.91	0.0095*	(0.0055)
Earnings 2 years before	111858.48	110612.95	-0.0157*	(0.0094)
Panel G: Long-term employment history (10 years)				
Months employed in last 10 years	58.19	62.91	-0.0022***	(0.0002)
No. 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 unempl. in last 10y	256.77	290.49	-0.0000***	(0.0000)
No. unempl. spells in last 10 years	3.63	3.83	0.0074***	(0.0018)
Average unempl. duration	95.31	90.15	-0.0001***	(0.0000)
Duration of last unempl. 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)
No. 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 avg. welfare benefits last 4y	4239.77	3533.38	-0.0213	(0.0142)
Yearly av.g welfare benefits last 10y	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)

Continue to next page

Table A.1 – continued from previous page

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

Continue to next page

Table A.1 – continued from previous page

	Treated	Control	Selection model	
			Est.	Std. Err.
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. Earnings and benefits are in SEK, parents' income in 100s SEK; all monetary values are inflation-adjusted.

References

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

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

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

Advani, A., Kitagawa, T., and Słoczyński, T. (2018). Mostly harmless simulations? On the internal validity of empirical Monte Carlo studies. IZA Discussion Paper, No. 11862.

Arni, P., Lalive, R., and van Ours, J. C. (2013). How effective are unemployment benefit sanctions? Looking beyond unemployment exit. *Journal of Applied Econometrics*, 28(7):1153–1178.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Eberwein, C., Ham, J. C., and LaLonde, R. J. (1997). The impact of being offered and receiving classroom training on the employment histories of disadvantaged women: Evidence from experimental data. *The Review of Economic Studies*, 64(4):655–682.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Kyyrä, T. (2010). Partial unemployment insurance benefits and the transition rate to regular work. *European Economic Review*, 54(7):911–930.

Kyyrä, T., Parrotta, P., and Rosholm, M. (2013). The effect of receiving supplementary UI benefits on unemployment duration. *Labour Economics*, 21:122–133.

Lalive, R., van Ours, J. C., and Zweimüller, J. (2005). The effect of benefit sanctions on the duration of unemployment. *Journal of the European Economic Association*, 3(6):1386–1417.

Lalive, R., van Ours, J. C., and Zweimüller, J. (2008). The impact of active labour market programmes on the duration of unemployment in Switzerland. *The Economic Journal*, 118(525):235–257.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–620.

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

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

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

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

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

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

Muller, P., van der Klaauw, B., and Heyma, A. (2017). Comparing econometric methods to empirically evaluate job-search assistance. Unpublished manuscript.

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

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

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

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

Røed, K. and Raaum, O. (2006). Do labour market programmes speed up the return to work? *Oxford Bulletin of Economics and Statistics*, 68(5):541–568.

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

Tatsiramos, K. (2010). Job displacement and the transitions to re-employment and early retirement for non-employed older workers. *European Economic Review*, 54(4):517–535.

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

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

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

van den Berg, G. J. and Vikström, J. (2014). Monitoring job offer decisions, punishments, exit to work, and job quality. *The Scandinavian Journal of Economics*, 116(2):284–334.

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

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

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

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

Vikström, J. and van den Berg, G. J. (2017). Långsiktiga effekter av arbetsmarknadsutbildning (Long-term effects of labor market training). IFAU working paper, 2017:17.

3. Targeted Wage Subsidies and Firm Performance[*]

with Oskar Nordström Skans and Johan Vikström

Acknowledgments: We are grateful for helpful suggestions from Anders Forslund, Francis Kramarz, Gerard van den Berg, Marco Caliendo, Martin Olsson, Olof Rosenqvist and from participants to EEA-ESEM 2017, EALE 2017, IZA Labor Market Policy Evaluation conference 2017. We also benefited from useful comments from seminar participants at Uppsala University and IFAU. This work was supported by FORTE.

[*] This chapter was published as: Lombardi, S., Skans, O. N., and Vikström, J. (2018). Targeted wage subsidies and firm performance. *Labour Economics*, 53: pp. 33–45·

3.1 Introduction

Targeted wage subsidies that reduce part of the wage costs for private firms hiring unemployed workers are an integral part of active labor market policies (ALMP) in most Western countries. The main objective is to help disadvantaged workers find jobs, and most studies tend to find that the policy tool is very efficient in this dimension (for surveys see, e.g., Card et al. 2010, 2017 and Kluve 2010). Despite these positive estimates, policy prescriptions tend to be cautious because of concerns regarding demand side responses (see e.g. Neumark, 2013). These concerns include crowding out of unsubsidized hires and fears that wage subsidies allocate workers to unproductive firms that are able to hire and compete on the market only due to the subsidies. Yet, there exists very little systematic evidence on the characteristics of the firms that hire with targeted subsidies, and on the impact the subsidies have on these firms.

In this paper, we make three distinct additions to the literature: we document the extent to which the characteristics of subsidized firms differ from those of other recruiting firms, we describe the extent to which key firm-level outcomes change due to the subsidies, and we analyze whether these patterns depend on the degree of caseworker discretion when subsidies are allocated. Together, this provides new empirical evidence on key concerns regarding wage-subsidy distortions. The results also provide some novel (and rare) evidence on how ALMPs affect the allocation of workers across firms, an issue that has received much recent attention within the wider labor-economic literature (see e.g. Card et al. 2013, Song et al. 2018 and Card et al. 2018).

Our analysis uses detailed Swedish administrative data on workers and firms in order to study the impact of targeted wage subsidies. We start from spell data on unemployed workers and the subsidies they receive and link this information to a matched employer-employee database which allows us to follow the employing firms over time. Data from business registers provides information on profits, sales, wage sums, value added and investments for the same firms.

Our analysis compares firms recruiting with subsidies (defined as treated) to other observably identical firms. We focus on small- and medium-sized firms throughout in order for the subsidies to be of a non-trivial magnitude relative to firm-performance measures. For the causal analysis, we compare treated firms to firms that hire unemployed workers without using subsidies. We adjust for pre-existing differences in firm size and separations, sum of wages paid and average workers' characteristics by matching on observable pre-treatment levels in these dimensions. We show that, after matching, the treated and matched controls have identical pre-treatment trends (which we do not match on). Furthermore, both pre-treatment trends and levels are remarkably similar in key dimensions that we do *not* match on, most notably productivity and profits. We find no evidence that the subsidies are allocated to low-performing firms. The pre-hire performance of the subsidized firms is remarkably similar to that of other recruiting firms, despite the fact that the subsidized hires (by design) have much longer pre-match unemployment spells. The main difference between the two groups of firms is that subsidized firms are smaller. But in terms of productivity, profits and staff composition, similarities in both levels and trends are striking.

We analyze two very different policy systems. Between 1998 and 2006 all targeted wage subsidies in Sweden needed to be approved by a caseworker at the public employment office. The caseworkers could also propose suitable employer-employee matches (see e.g. Lundin, 2000). This staff-selection scheme is contrasted to a new rules-selection system introduced in 2007, which granted all employers that hired an eligible long-term unemployed worker the right to receive a wage subsidy, thus substantially reducing the role of caseworkers in the allocation of the subsidies.

In the regime where caseworkers pre-approved subsidized matches, treated firms substantially outperform the comparison firms *after* the treatment, both in terms of the number of employees and in terms of various production measures, despite having identical pre-match trajectories. This pattern is persistent and it does not come at the cost of decreased productivity per worker. That is, in this system, the subsidies are clearly associated with positive changes in firm performance. In the second system, when long-term unemployed are entitled to subsidies without caseworker approval, the results are less clear. We find no corresponding change in firm size and productivity measures among surviving firms. This would suggest larger crowding-out effects and more windfall gains. On the other hand, the subsidies have a clear positive effect on firms' survival rates in the rules selection regime.

We show that the difference between systems is not due to differences in the hired workers' characteristics. If anything, caseworkers target more vulnerable workers and detailed controls for worker characteristics does not change the conclusion. Further evidence suggests that business cycle conditions and/or the increasing share of immigrant workers are unlikely explanations for the differences between systems. A possible hypothesis for the different findings is instead that caseworkers act as gatekeepers guarding against both displacement of non-subsidized jobs and windfall gains, and screening against firms on the margin of exit. As a corroborate of this hypothesis, we show results indicating that caseworkers guard against an overallocation of subsidies to firms with poor internal expectations about future performance. This exercise uses data on investments which (in line with standard investment theory) we interpret as a forward-looking variable capturing the firm's own expectations about future performance and we find that investments are lower for treated firms in the rules-selection scheme but not in the staff-selection scheme.

Our paper is related to several strands of the existing literature. In a recent paper, Cahuc et al. (2018) use a French reform in 2008 to study the effectiveness of hiring credits. Firms with fewer than 10 employees that hire a worker with a wage less than 1.6 times the minimum wage were eligible for the credit. The main result is of a strong and immediate employment effects of the credits. Using experimental variation, Crépon et al. (2013) find that a job placement assistance program in France displaces employment of non-treated unemployed individuals searching for jobs in the same area as the treated workers. In our paper, we find evidence of a different type of displacement, namely that of non-subsidized workers already employed in the firms hiring with the subsidies. Kangasharju (2007) uses Finnish data that links firms and workers, and finds that employment subsidies in Finland increased the firms' payroll by more than the size of the subsidy. Other studies on displacement effects include those that have used surveys of employers. For instance, Bishop and Montgomery (1993) survey more than 3500 private employers in the US and conclude that at least 70% of the tax credits granted to employers are payments for workers who would have been hired in the absence of any subsidy. In a similar vein, Calmfors et al. (2002) discuss Swedish survey-based evidence. Andersson et al. (2016) evaluate a *training* program in the U.S. and consider various measures of firm quality as outcomes. These measures include firm size, turnover, as well as firm-effects defined in Abowd et al. (1999). Overall, they find modest effects on the quality of the firms where the formerly unemployed workers find jobs.[1]

Finally, two recent studies examine how active labor market programs affect firm behavior and firm-level outcomes. Blasco and Pertold-Gebicka (2013) study a large scale randomized experiment on the effects of counseling and monitoring, and examine if this affected the firms in areas exposed to the experiment. Lechner et al. (2013) exploit that German local employment offices determine the mix of ALMPs to study firm level effects. In this paper, we use data that links firms and workers to study firms that are actually targeted by the subsidies, whereas these two studies focus on effects on all firms in a certain area.[2]

The paper is structured as follows. Section 2 provides the institutional background and discusses the potential role of caseworkers. Section 3 explains the data and outlines the empirical strategy. Section 4 presents the results. Finally, Section 5 concludes.

[1] For survey evidence how wage subsidies affect the unemployed workers covered by the subsidies see Card et al. 2010, 2017; Kluve 2010). For recent evidence on Swedish data, see Sjögren and Vikström (2015) on targeted employment subsidies and Egebark and Kaunitz (2018) and Saez et al. (2017) on non-targeted payroll tax reductions for youths. The latter of these papers also study spillover (wage) effects within the firms through rent sharing.

[2] Other papers studying spillover effects at the market level include, for instance, Blundell et al. (2004), Lise et al. (2004), Ferracci et al. (2018), Pallais (2014), Gautier et al. (2018) and Lalive et al. (2015). These studies use geographical variation and/or theoretical models to study spillover effects at a more general level, including market equilibrium effects. In contrast, we focus on the allocation of workers across firms and on how targeted wage subsidies affect firm performance.

3.2 Background

3.2.1 The targeted wage subsidies

In Sweden, targeted wage subsidies and all other aspects of Active Labor Market Policies are administrated by the Swedish Public Employment Service (PES). The overall aim of the agency is to promote a well-functioning labor market for both unemployed individuals and firms. The PES provides different policy measures targeted to unemployed individuals, including job search counseling, labor market training, practice programs and targeted wage subsidies. Another aim is to support firms in the recruitment process, in particular by maintaining a free and publicly available vacancy database. The PES is divided into 280 local public employment offices. Each unemployed individual is assigned to a caseworker at the local office, and caseworkers are responsible for enrolling the people assigned to them into policy programs and to provide job-search assistance.

In this paper we focus on targeted wage subsidies. These subsidies target different sets of unemployed individuals and reimburse part of the firms' labor costs by crediting their tax accounts when an eligible person is hired. The aim is to provide firms with incentives to hire those that otherwise would struggle to find non-subsidized jobs. From the perspective of the long-term unemployed, the subsidized job can be a stepping-stone towards a non-subsidized job. Workers hired through these subsidies are subject to exactly the same regulations (including employment protection laws) as non-subsidized workers.

We analyze two different subsidy systems. The first, the Employment Subsidy Program (Anställningsstöd) was in place between 1998 and 2006. The program was targeted *and* selective. It was mainly *targeted* to individuals unemployed for at least 12 months and at least 20 years old.[3] The program replaced 50 percent of the labor cost (including payroll taxes) for a maximum duration of 6 months. The program was *selective* in the sense that each subsidized job had to be approved by a caseworker at the local PES office. The importance of caseworkers is confirmed by implementation surveys. Lundin (2000) shows that caseworkers sometimes initiate the subsidized match, even though firms always have the opportunity to decline suggestions from the caseworker. In addition, Harkman (2002) shows that caseworkers have fairly strong and varying views on the appropriateness of these (and other) programs. Taken together, this means that caseworkers influence how the subsidies are allocated to different firms and workers. We therefore refer to this subsidy system as the *staff-selection* system.

The second scheme we study is the "New Start Jobs program," introduced in January 2007. This program is targeted but not selective.[4] Similar to the

[3] Workers with special needs or workers with extensive unemployment histories may obtain a subsidized job before 12 months of unemployment.

[4] Note that the subsidy can be paid on top of the youth reduction in payroll taxes introduced in 2018 which was studied by Egebark and Kaunitz (2018) and Saez et al. (2017).

staff-selection system, the new subsidies *target* individuals who have been unemployed for at least 12 months. However, the system is not selective since any worker who has been unemployed for at least 12 months during the last 15 months has the *right* to receive the subsidy if they find a job.[5] The overall size of the subsidy is similar to the previous system. The New Start Jobs program has a slightly lower replacement rate but a longer duration. It replaces 31.42 percent of the wage cost for a time equal to the duration of unemployment (i.e. at least 12 months). Overall, if anything the New Start Jobs subsidies are more generous than those in the staff-selection system.

Thus, the main difference between the two policy systems is that the Employment Subsidy Program involves caseworker approval, whereas the New Start Job system does not. Under the new system, firms employing an eligible individual have the right to use the subsidy.[6] That is, caseworkers do not have to approve each subsidy, and in most cases they are not even involved in the allocation of the subsidy. Under the new system, caseworkers can still act as facilitators in forming new employer-employee matches, but their counseling activity is neither required for starting new subsidized jobs nor binding. Instead, firms are solely responsible for initiating the procedures to apply for the targeted wage subsidy. Since the allocation of the subsidies is determined by the rules for the subsidy and not by caseworkers, we refer to this second program as the *rules-selection scheme*.

In both cases, firms hiring through subsidies are subject to the regulations as other hires in most other dimensions. As a consequence, the same employment protection laws apply to both the subsidized and the non-subsidized workers.

3.2.2 Conceptual differences between the two policy regimes

We will examine if the subsidies are targeted to low-performing firms and if they are associated with large windfall gains for employers, and if the empirical patterns related to these concerns differ between two different policy regimes. The first regime is a system with staff-selection, where subsidies have to be approved by a caseworker, and the second is the rules-selection regime where all unemployed job seekers are eligible for the subsidies.

The caseworkers' involvement can affect the allocation of workers across firms, either by not approving firms that merely use the subsidies to replace

[5] Differently from the Employment subsidy program, the New Start Jobs subsidy does not require the individual to be registered as unemployed. Poor health, incarceration or other reasons for non-employment could suffice. This also implies that some subsidized jobs may start before 12 months of unemployment if these workers qualify through other types of non-employment, so that the 12 months eligibility threshold is not strictly binding.

[6] The only requirement is that the the prospective worker provides sufficient documentation of eligibility. The firms also have to fulfill some basic requirements, such as not having significant amounts of unpaid taxes. From January 2017 a new requirement is that the participating firms need to have a collective agreement with a labor union.

non-subsidized jobs and/or by allocating the subsidies such that the quality of the match between workers and firms is higher.[7] Caseworkers can thus affect sorting and selection which may lead to improved firm outcomes. This also implies that the setting may differ from the traditional evaluation one, in the sense that the role of sorting is interesting in itself. This also implies that positive outcomes arising from an allocation towards firms with a more positive forward trajectory is a legitimate successful outcome of the allocation process, at least from the perspective of the caseworker. However, we retain the terms treated and comparison/control to refer to firms hiring with and without the subsidies, respectively.

3.3 Empirical strategy and data

3.3.1 Data

We use data from several Swedish administrative registers. Data from the Swedish Public Employment Service provides information about all registered unemployed individuals. It contains detailed information about all individuals receiving targeted wage subsidies through our two systems (Employment subsidies and the New Start Jobs), including the start and the end date of each subsidy. By using unique personal and firm identifiers, this data is merged to a matched employer-employee database from Statistics Sweden (RAMS register).[8] This database contains information on all employment episodes for all employees in Sweden. Each employment episode is linked to the corresponding firm and provides us with information on yearly labor income and basic information about the firm. Using the matched employer-employee data we can follow firms and workers over time, which allows us to construct a firm level panel data set with information on the number of employees and the hiring and separation rates in each year.[9] We focus on both the total number of workers and the number of workers who were hired using the employment subsidies. The latter includes both workers currently covered by the employment subsidies and workers remaining in the firm after the subsidy has ended.

We also use information on firms' operating costs and profits, assets value, revenues, yearly turnover, investments, value added and other firms' produc-

[7] Caseworkers' gatekeeper role within public employment offices has rarely been studied before, despite evidence of the importance of gatekeeper roles having been found in other public sector areas. For instance, Engström and Johansson (2012) and Markussen et al. (2013) show that medical doctors can act as gatekeepers in disability and sickness insurance systems.

[8] The PES data does not include information on the hiring firm, and the matched employer-employee data does not include information on the exact subsidy start date. Since a worker can start multiple jobs, we need another way to link each wage subsidy to a particular firm. We do this by only keeping the job with the highest salary.

[9] The number of hires is the number of workers employed in the firm during the current year who were not employed in the same firm the previous year. The number of separations corresponds to the number of workers employed in the firm the previous year but not the current one.

tion measures. This data is obtained from Statistics Sweden's business register of firm-level accounts. Operating profits are the difference between operating revenues (generated from the firm's core business activities) and operating expenses (such as costs of goods and production), minus depreciation and amortization. Value added is the total value that is added at each stage of production excluding costs for intermediate goods and services, and is equivalent to total revenues minus intermediate consumption of goods and services. Worker productivity is defined as the total firm's valued added divided by the number of workers. Investments per worker are the total yearly amount spent on land and machinery, net of the disinvestments in the same categories and divided by firm size.

Finally, population registers from Statistics Sweden are used to construct information on the characteristics of the employees at the firm-year level. These include age, level of education, civil status, immigrant status and gender.

3.3.2 Sampling and comparison group

We compare firms recruiting through subsidies (defined as treated) to other observably identical firms. Let us illustrate the sampling procedure for treatments in year t. We first sample all firms with fewer than 30 workers in year $t-1$. The reason for this is one subsidized job constitutes a small treatment for large firms. We therefore focus on small- and medium-sized firms for which we expect to see effects. We also exclude firms with only one worker, and select the firms that survive until year t.[10] This implies that we observe at least one year of firm history.[11] Next, we use the PES information on the employment subsidies to identify firms with subsidized hires during the first quarter of year t. We focus on jobs starting during the first quarter both because our firm-level outcomes are measured on a yearly basis and in order to diminish the influence of short term-vacancies that are used across the summer.

We use the matched employer-employee data to sample firms observed during the 1998-2008 period. The justification for the 2008 restriction is that the subsidy rate was doubled for all new New Start subsidies starting in January 2009 and onwards. Moreover, by focusing on this time period we also avoid sampling firms during the great recession (the unemployment rate in Sweden started to rise during the first quarter of 2009, but the impact was much smaller

[10] In most cases firms with only one worker are firms where the owner is the only worker (self-employed). Most of these firms never intend to grow, therefore they are not at the risk of using the subsidies, which explains why we exclude them from our analyses.

[11] We drop firms that grow to more than 60 workers within five years. The reason for this is that disproportionately fast-growing firms are likely to be driven by mergers. As robustness checks, we have used different firm size cutoffs and we have studied whether the treatment affects the probability that the firms grow to more than 60 workers but, reassuringly, we found no significant effects and tiny point estimates (-0.004 (se 0.003) and -0.003 (se 0.004) for the two regimes, respectively).

than in Europe as a whole). For each firm we only study the first wage subsidy within our observation period. This sampling procedure gives us 8,679 treated firms in the staff-selection system and 3,411 treated firms in the rules-selection system.[12]

As comparison group, we select firms that hire from the pool of long-term unemployed the same years and quarters, but without using the subsidy (not during the entire calendar year). We ensure that they have not hired with the subsidy in the past, but allow the comparison firms to use the wage subsidies in the future (5.3% of the comparison firms do this within 5 years). As for the treated firms, we focus on firms hiring from the pool of long-term unemployed in the first quarter of the year. A long-term unemployed is defined as an individual who finds a job after at least *six months* of unemployment according to the PES data. Since these comparison firms also hire at least one formerly unemployed worker in the same quarter as the treated firms, they are arguably in a somewhat similar situation as the treated firms.[13] We repeat the sampling procedure each year, which means that a firm can be selected as comparison firm in multiple years.

For both types of subsidies the general rule is that the workers become eligible after 12 months of unemployment. However, we use a 6 months threshold for the comparison group to ensure that we use ineligible, but otherwise similar, workers. Since workers hired after 6 months should have more favorable unobserved characteristics than workers hired after more than 12 months of unemployment, any positive estimates for the subsidies should be considered as "conservative" (i.e. biased towards zero). Note however that, as discussed in Section 2, the 12 months eligibility criterion is not strictly binding (in any of the two regimes) so the treatment group does include some firms which hire workers after less than 12 months of unemployment. To ensure that these choices are not driving our results, we present a robustness analysis where we control for the elapsed unemployment duration (and other characteristics) of the hired workers, leading to very similar results.

In the staff-selection system, the comparison group includes both firms to which the caseworkers actively deny a subsidy and firms which hire a long-term worker without making a subsidy claim, potentially because the preceding spell was too short. We cannot separate between these groups of firms. Similarly, in the rules-selection system the comparison group includes firms that do not use the subsidy despite being entitled to do so (e.g. because of not

[12] Note that the number of subsidies are slightly higher in the rules-selection regime (1700 per year) than for the staff-selection regime (960 per year). However, note that these numbers are small compared to the total number of firms in Sweden, so that it is unlikely that this difference between the two regime lead to differential general equilibrium effects.

[13] Note that the comparison group is made up by workers who are not formally entitled (yet) and those that are formally entitled but are not selected by caseworkers in the caseworker regime and those that choose not to participate in the rules selection regime. One reason for failing to use the subsidy when entitled is the stigma effect discussed by, e.g. Neumark (2013).

understanding the rules, or in case the hired worker does not disclose the duration of joblessness) and firms that hire a worker whose preceding spell was too short. In most of our specifications, we exclude disappearing firms from the year they disappear, but we also examine effects on firm survival and we are careful to take such effects into account when we interpreting our results.

3.3.3 Raw sample statistics

Table 1 provides summary statistics for the firms in our sample, but it also contains one of the key findings of this paper. In fact, the most striking feature of the table, in our view, is that with very few exceptions the treated firms (hiring with subsidies) are quite similar to the firms that hire unsubsidized long-term unemployed workers. Moreover, with one exception only (age of the hired worker), selection (on observables) is very similar between the staff-selection and rules-selection regimes.

Panel A of the table shows the industry composition. The treated firms are somewhat more likely to be in the manufacturing industry and wholesale/retail but for other industries, differences are small. Selection on all variables is very similar between the staff-selection and rules-selection regimes. Panel B turns to the employee-composition of the hiring firms. These statistics are again remarkably similar between the treated and controls considering that these are raw data generated by self-selection. The one statistic where there are some differences the share of high educated, which is somewhat lower within the treated firms. The time trends of increasing education and increasing shares of immigrants *between* the two regimes are visible but the within-period selection is very similar for the two regimes.

Panel C shows statistics for the hired workers. The main difference between treated and controls is that the subsidies target workers with much longer unemployment spells on average. This is true by design since we only require the control firms to hire workers who have been unemployed for at least 6 months. But despite this difference, we find rather similar age profiles and shares of immigrants (although higher in the second regime as expected due to low skilled immigration; we will return to this issue). The only notable difference between the treated and comparison firms is that the share of workers below 25 is higher among the treated firms in both regimes. We also see a shift from under-representation among the treated within the oldest group (55-64) to an over-representation of treatment within the same age group. We will explore these differences in several ways.[14] Panel C also shows that education is somewhat lower and the share of males is higher for the treated.

[14] In one robustness analysis, we match on all worker characteristics, and in another robustness analysis we exclude the oldest and the youngest workers. In both cases without any change in results.

Table 1. *Sample statistics for treated and comparison firms in the two regimes*

	Staff selection		Rules selection	
Firms group	Treated	Control	Treated	Control
Group size	8,679	25,322	3,411	4,798
Panel A: Industries				
Agriculture	0.04	0.03	0.02	0.03
Manufacturing	0.19	0.13	0.20	0.12
Construction	0.11	0.10	0.11	0.10
Wholesale and retail trade; repair	0.28	0.24	0.26	0.22
Accommodation and food service	0.07	0.09	0.10	0.13
Transport and storage	0.06	0.09	0.06	0.08
Real estate activities	0.16	0.20	0.16	0.19
Education	0.02	0.02	0.01	0.04
Human health and social work	0.02	0.03	0.02	0.03
Panel B: Pre-treatment average workers' characteristics				
Married	0.38	0.38	0.38	0.37
Male	0.68	0.62	0.68	0.62
Immigration to Sweden	0.12	0.16	0.21	0.24
Education: Compulsory	0.25	0.24	0.22	0.22
Education: Secondary	0.58	0.55	0.57	0.53
Education: Upper	0.17	0.22	0.21	0.26
Age: 24 or less	0.19	0.18	0.18	0.19
Age: 25–34	0.28	0.28	0.23	0.26
Age: 35–44	0.24	0.23	0.25	0.24
Age: 45–54	0.18	0.18	0.18	0.17
Age: 55–64	0.10	0.11	0.13	0.12
Age: 65 or more	0.01	0.01	0.02	0.02
Panel C: Hired workers' characteristics				
Age: 24 or less	0.26	0.11	0.17	0.11
Age: 25–34	0.32	0.33	0.25	0.31
Age: 35–44	0.21	0.26	0.23	0.26
Age: 45–54	0.14	0.18	0.18	0.19
Age: 55–64	0.07	0.12	0.18	0.13
Age: 65 or more	0.00	0.00	0.00	0.01
Immigration to Sweden	0.21	0.22	0.32	0.30
Married	0.29	0.35	0.36	0.35
Male	0.69	0.59	0.66	0.58
Education: Compulsory	0.20	0.20	0.21	0.20
Education: Secondary	0.66	0.59	0.59	0.55
Education: Upper	0.15	0.20	0.21	0.25
Average unemployment (days)	660.88	410.98	638.13	371.02

Continue to next page

Table 1 – continued from previous page

	Staff selection		Rules selection	
Firms group	Treated	Control	Treated	Control
Panel D: Pre-treatment firm outcomes				
No. of workers	9.70	11.09	10.08	11.66
Wage sum per worker	109.11	107.19	124.38	119.93
Hirings rate	0.28	0.30	0.30	0.31
Separations rate	0.23	0.26	0.22	0.25
Value added per worker	385.35	410.55	426.90	439.16
Operating profit per worker	74.96	76.78	91.56	79.71
Total investments	228.71	206.45	163.53	175.60
Investments per worker	52.69	44.53	37.39	41.21

Notes: Sample statistics for treated and comparison firms before matching. Panel A: share of firms hiring in each industry; Panels B, D: pre-hiring averaged workers' characteristics; Panel C: hired workers' demographics and residual time in unemployment before exiting to job. Panel D: all monetary values are inflation-adjusted (base year: 2000), and all outcomes normalized by firm size. Wage sum is the yearly sum of wages paid by the firm. Value added is total revenues minus costs of intermediate goods. Operating profit is the difference between operating revenues and expenses, minus depreciation and amortization.

The statistics in Panel C are relative to the subsidized workers and the long-term unemployed workers hired by the treated and comparison firms, respectively. Besides these workers, the two groups of firms may also hire other workers (non-subsidized) during the treatment year. Sample statistics for these workers are presented in Table A2 in the appendix. Here, we find very similar age and education profiles for the treated and comparison firms in both subsidy systems.

Panel D shows the pre-treatment outcomes of the hiring firms. Here we see somewhat larger differences, but as we will show in the results section below, they all essentially reflect the same underlying variable, namely that treated firms tend to be smaller than the comparison ones. Note that we focus on firms with fewer than 30 employees, which explains why the average firm size is rather small. Figure 1 shows the average number of workers in the treated and comparison firms within five years since the start of the subsidy, in the staff-selection system. Year zero is the year the subsidy starts or, for the comparison firms, the year they hire a long-term unemployed worker without a subsidy. From the figure we see that although the comparison firms are on average somewhat larger than the treated firms, the trends for the two groups are very similar. For both treated and comparison firms, the average number of workers remains roughly constant before the subsidy. Since we sample firms hiring at least one worker in year zero, we observe a jump in firm size in year zero for both groups. After this, firm size decreases over time, consistently with regression towards the mean. Figure 2 shows similar patterns for the rules-selection system.

Figure 1. Number of workers for treated and comparison firms, before matching (staff-selection system)

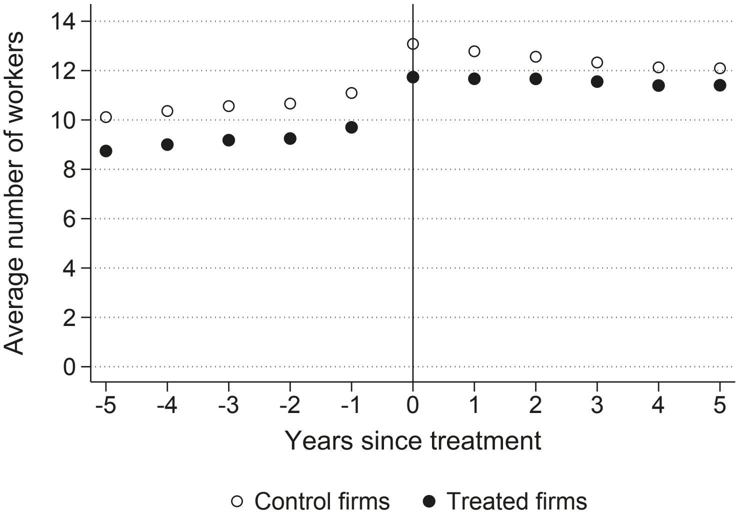
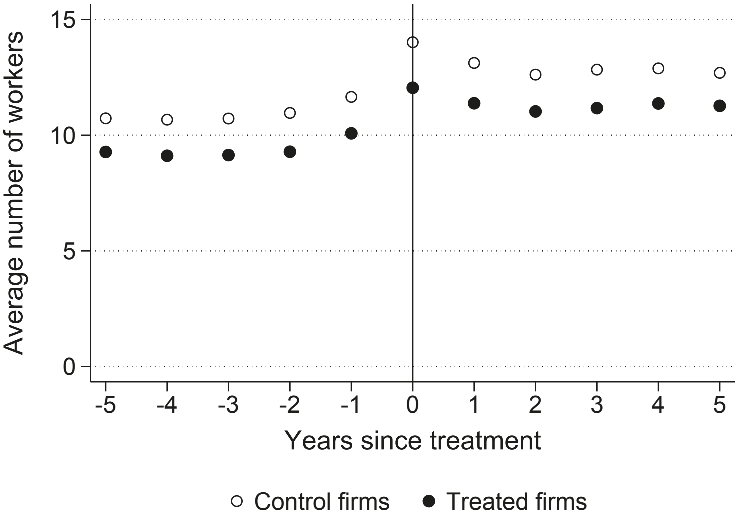


Figure 2. Number of workers for treated and comparison firms, before matching (rules-selection system)

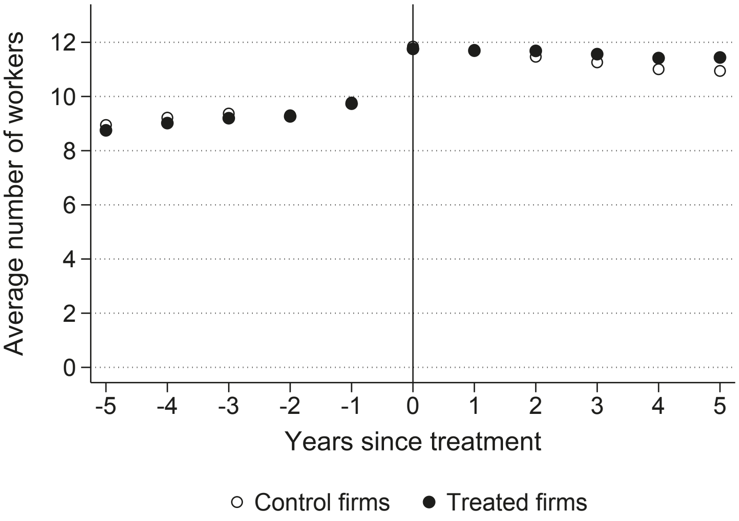


3.3.4 Matched samples

We believe that the statistics presented above (in particular, the size trends) are reassuring in terms of the basic approach of comparing treated and comparison firms to assess the impact of the subsidies. However, to ensure that we purge our comparison from any additional differences in observables, we use a matching algorithm. We select one comparison observation for each treated observation using nearest-neighbor propensity-score matching. Our matching vector includes the following variables (described in Table 1): industry dummies (8 categories), firm size, wage sum, number of separations as well as firm-level employee composition as captured by the variables in Table 1, Panel B. We perform the matching procedure separately for each calendar year (thus, also by subsidy scheme), and aggregate the data into two matched samples, one for the staff-selection system and one for the rules-selection system.

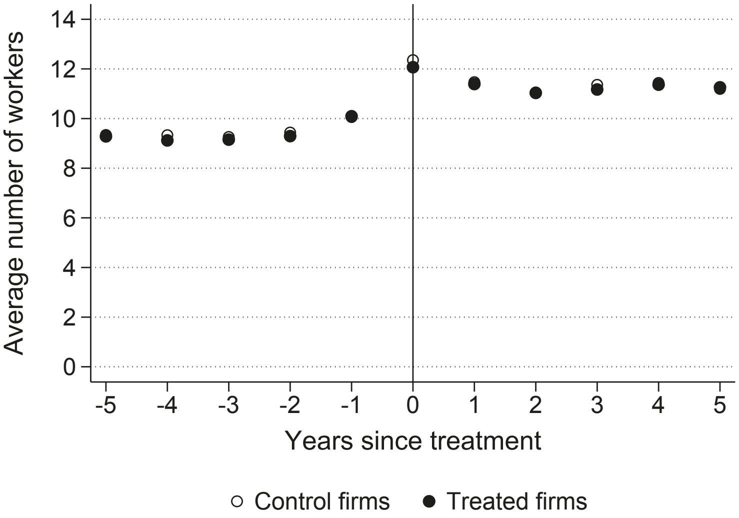
Figure 3 illustrates the matched treated and comparison firms in the staff-selection system. Note that we match on the average number of workers in year -1, which explains why firm size is almost exactly the same for the two groups in that year.[15] More importantly, the average number of workers is very well aligned for all pre-treatment years, despite the fact that we only match on the number of workers in year -1. We obtain similar results for the rules-selection system (Figure 4).

Figure 3. Number of workers for treated and comparison firms, after matching (staff-selection system)



[15] We have also examined the balance for the other firm characteristics used in the matching, and as expected they are all well-balanced.

Figure 4. Number of workers for treated and comparison firms, after matching (rules-selection system)



Differences between treated and matched controls in number of employees, wage sum and separations within a 5-year pre-match period are shown in Table A.1. To assess the usefulness of the matching protocol, we also check for pre-treatment differences in firm-performance measures that we do *not* match on. To this end, Table A.1 reports balancing tests for average profits, log value added and investments, as well as these three outcomes measured per worker, in the pre-hiring period up to five years before the subsidy. We also report statistics on the fraction of the firms that existed in the 5-year period before the treatment. Even if we do not match on these variables, we find very small differences between the treated and the comparison firms. This holds both for the staff-selection system (columns 1-3) and the rules-selection system (columns 4-6). The fact that we find similar pre-treatment trends also for these variables suggests that our matching protocol does produce control firms with a very similar history as the treated firms, also in terms of unobserved dimensions. In the robustness section, we provide estimates when matching on the characteristics of the hired worker, and when matching on a broader set of firm outcomes in levels and trends. As expected from the balancing tests described here, results are robust.

3.3.5 Empirical model

Our analysis relies on comparing treated and comparison firms' outcomes using the matched samples. However, since we observe each cross-sectional unit over time, we can further strengthen the analysis by applying panel data methods to control for any group-specific differences not accounted for in the matching step. Thus we can adjust for all observed and unobserved fixed characteristics by estimating the following baseline model for firm i in year t:

$$y_{it} = \lambda_t + \beta D_i + \gamma(D_i \cdot T_{it}) + \varepsilon_{it}, \tag{3.1}$$

where λ_t is a year dummy, D_i is an indicator variable for firms in the treated samples and T_{it} is an indicator variable taking the value 1 after the start of the subsidy in this set of firms. Thus, D_i captures any remaining time-constant pre-existing differences between matched treated and comparison firms. In our robustness analyses, we also use firm fixed effects. The interaction $D_i \cdot T_{it}$ reflects any difference between the two groups after the start of the subsidy. We allow this difference to vary by time since the start of the subsidy. Model (3.1) is estimated separately for each subsidy system. Standard errors are clustered at the firm level.[16]

3.4 Results

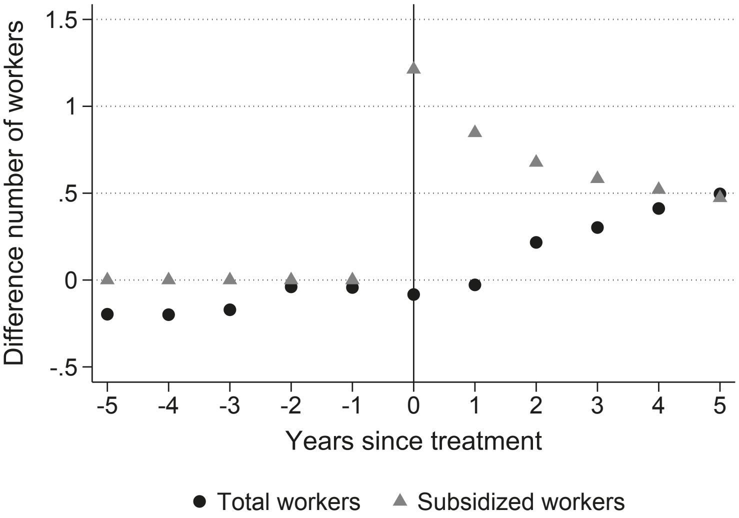
3.4.1 The staff-selection system

We first focus on the staff-selection system, during which all subsidies need to be approved by caseworkers. Figure 5 shows the difference between treated and comparison firms in the total number of workers (dots) and in the number of subsidized workers (triangles). As already noted, there are virtually no differences between treated and comparison firms in the pre-subsidy period. In the subsidy year, the number of subsidized workers increases by slightly more than one, which reflects the fact that some firms hire more than one subsidized worker at once. At the same time, the total number workers is almost unaffected. This happens because the comparison firms also hire at least one worker in year zero. After this, we see a gradually increasing positive difference between the average firm size of the treated and comparison firms. Five years after the start of the subsidy, the difference is around 0.5 workers. Since the average firm size in our sample is just below ten workers, the magnitude of this difference is far from trivial.

Figure 5 also reveals to what extent the observed differences between treated and comparison firms are due to the number of subsidized workers and/or due to the number of non-subsidized workers. Individuals hired with a subsidy are

[16] Note that this procedure does not into account that there is sampling variation in the matching step. Addressing this issue properly involves a large computational burden. We therefore provide estimates for given matched samples and have validated the most important results by performing genuine conditional difference-in-differences, using nearest neighbour Mahalanobis metric matching and the Abadie and Imbens (2006) estimator of the standard errors. This resulted in very similar standard errors.

Figure 5. Difference treated and comparison firms, staff-selection system



counted as subsidized workers throughout the remainder of their job spell.[17] Unsurprisingly, over time the number of subsidized workers decreases since some of them leave the firm, reflecting standard firm turnover in the labor market. Five years after the subsidy start, roughly 50% of the workers remain in the firm (around 0.5 workers). This number is almost identical to the difference in the total number of workers between treated and comparison firms. We conclude that the subsidies in the staff-selection system create *net* employment, and that the subsidized workers who remain in the firm do not replace other workers.

In Panel A of Table 2, we analyze the impact on various outcomes using the regression model presented in equation (1). In Column 1, we first examine if the effects on firm size are driven by differential firm survival, but we see no impact on the probability to remain in business. The table also reports estimates for several other firm performance outcomes. Column 2 repeats the results for firm size already highlighted in Figure 5. As expected, we obtain a similar pattern and the differences between treated and comparison firms both 1-2 and 3-5 years after the start of the subsidy are statistically significant. In Column 3, we study effects on the yearly wage sum. Although estimates are less precise, we obtain a similar pattern as for the number of workers.

A reasonable concern at this stage is that the increased number of workers could have a negative impact on productivity. We therefore turn to the

[17] The number of subsidized workers includes everyone hired using a subsidy, including both currently subsidized workers and workers who remain in the firm after the subsidy has expired. Very few firms in our sample use the subsidies more than once.

Table 2. *Estimates for firm–level outcomes by years since treatment*

	Firm survival	No. workers	Wage sum	Profits	Value added	Value added per worker
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Staff selection						
Year of	–	0.03	26*	35	0.06***	0.01
treatment		(0.11)	(14)	(33)	(0.01)	(0.01)
1–2 yrs. after	-0.0001	0.21*	20	64*	0.09***	0.03***
treatment	(0.0026)	(0.12)	(20)	(34)	(0.02)	(0.01)
3–5 yrs. after	0.0039	0.52***	55*	116***	0.09***	0.03***
treatment	(0.0034)	(0.15)	(29)	(42)	(0.02)	(0.01)
Average	0.7721	11.48	1731	482	7.56	6.02
No. observations	86,020	157,758	157,758	121,376	127,104	119,580
Panel B: Rules selection						
Year of	–	-0.20	28	-7	-0.01	-0.02
treatment		(0.19)	(30)	(62)	(0.03)	(0.02)
1–2 yrs. after	0.0200***	0.05	13	66	0.03	-0.00
treatment	(0.0039)	(0.21)	(41)	(65)	(0.03)	(0.02)
3–5 yrs. after	0.0433***	0.02	7	-60	0.03	-0.01
treatment	(0.0053)	(0.26)	(63)	(76)	(0.04)	(0.02)
Average	0.7826	11.26	2081	584	7.68	6.13
No. observations	33,970	62,807	62,807	52,139	50,741	47,195

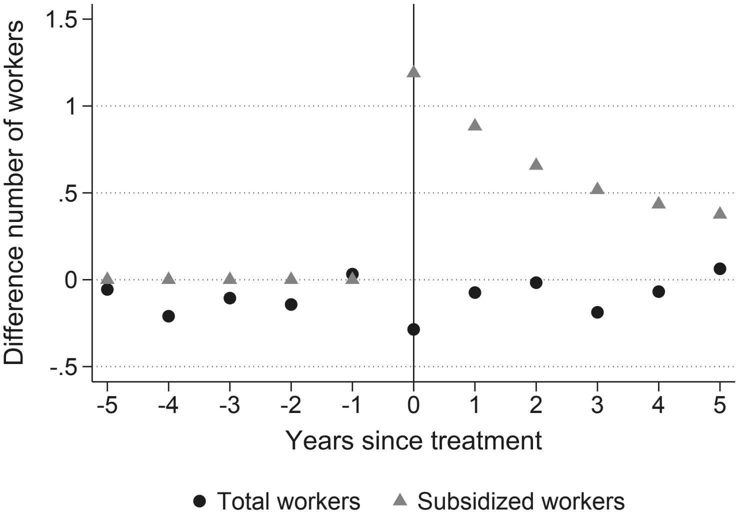
Notes: Estimates using the matched samples. Each model includes calender time fixed effects and indicators for treatment status. Average outcomes computed 3–5 years after treatment. Number of observations corresponds to the observed firm history years for Columns 2–5 and to the post–treatment period years for Column 1. Wage sum (in 1000 SEK) is the sum of all wages paid by the firm during the calendar year. Total value added (in log 1000 SEK) is total revenues minus intermediate consumption of goods and services. Profits (in 1000 SEK) are the difference between operating revenues and operating expenses, minus depreciation and amortization. Value added per worker is the logarithm of total value added divided by firm size. Standard errors clustered at firm level in parentheses. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

impact on firm performance measures. Column 4 reveals significant positive effects on profits. This may partly be a mechanical effect due to the subsidy. In Column 5, we show that the size effect is also visible in terms of production (log value added), which is reassuring. But, more importantly, we also want to assess the impact on *productivity per worker*. To this end, Column 6 studies the impact on log value added per worker. The results in fact suggest that productivity increases by 3 percent as a result of the subsidy. Thus, the faster size growth in the treated firms does not come at the cost of decreased per-worker productivity, but rather the reverse. This is perhaps even more surprising considering that the treated firms hire workers with twice as long elapsed unemployment duration as the control firms (see Table 1).

3.4.2 The rules-selection system

In the rules-selection system, caseworkers' involvement in the match creation greatly diminishes. First, we show in Figure 6 the difference in the total number of workers and the number of subsidized workers between treated and comparison firms. As for the staff-selection system, the number of subsidized workers increase by roughly one unit in the subsidy year and subsequently declines to about 0.5 workers five years after the subsidy. In contrast to the staff-selection results, we find no differences in size between treated and comparison firms during the follow-up period.

Figure 6. Difference treated and comparison firms, rules-selection system



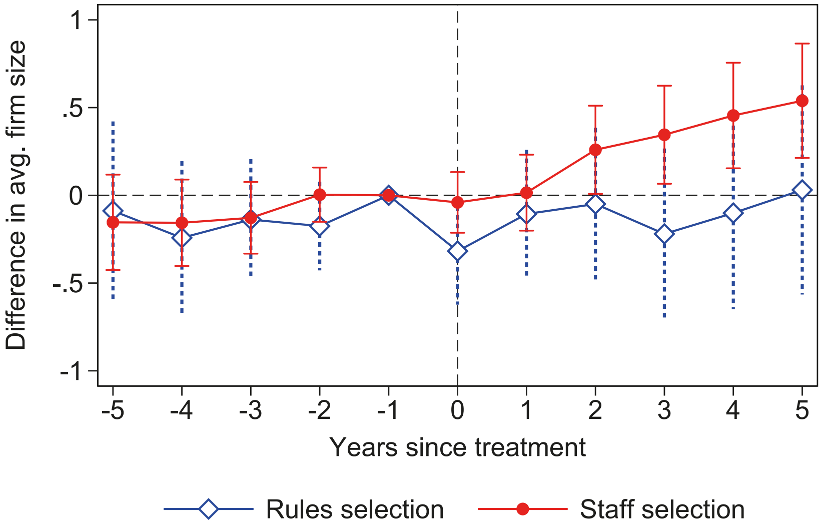
Results in table format are presented in Panel B of Table 2. Interestingly, we find a significant positive effect on firm survival that we do not see for the staff-selection system. Two years after the subsidy, the treated firms are 2 percentage points more likely to remain in business than the control firms. Since we find no evidence in this direction during the staff-selection period, the results appear to suggest that the caseworkers may have reduced the exposure to wage subsidies of firms that are on the verge of collapsing. It also suggests that the rules-selection subsidies have a positive effect on employment through reduced firm closures whereas the staff-selection subsidies had a positive employment effect through the performance of the survivors.

In Column 2, we repeat the analysis for number of employees, finding very small (insignificant) estimates both 1-2 and 3-5 years after the treatment. This pattern holds for all the other outcome variables shown in the table (Column 3 wage sum, Column 4 profits, Column 5 production and Column 6 productivity).

3.4.3 Comparison between the two subsidy schemes

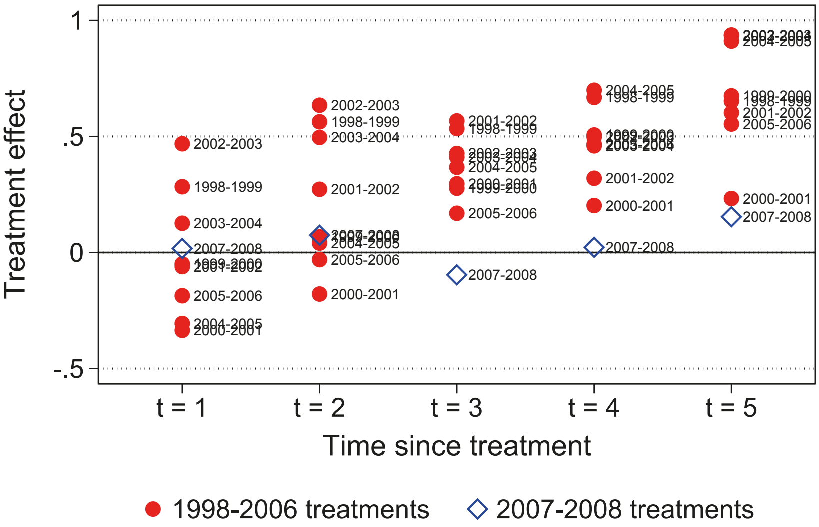
We now turn to a more explicit comparison between the two systems. We use the matched samples and show separate estimates for each year before and after the long-term unemployed hire. Figure 7 shows the estimates for the total number of workers for each system (with 95% confidence intervals). As already stressed, for both systems there are no significant pre-treatment trends. Moreover, the figure confirms the striking differences between the two systems. During the staff-selection system, the subsidies lead to increased employment, while during the rules-selection system there is no effect on the total number of workers. This pattern holds despite the fact that the subsidized workers tend to stay in the firm to the same degree in the two systems. In both Figures 5 and 6, we showed that around half of the subsidized workers remain in the firm five years after the start of the subsidy. Thus, the differences in total number of workers across systems is due to the number of non-subsidized workers.

Figure 7. Estimates for total number of workers, comparison of the two systems



To highlight that this result is unlikely to be due to random variation, Figure 8 shows estimates for each pair of two contiguous calendar years (using number of employees as the outcome). As expected, we observe some nontrivial variability in the estimates but the long-run staff-selection estimates remain distinctly positive (red dots), whereas the estimate is at zero for the rules-selection hirings (blue triangle).

Figure 8. Estimates for number of workers by calendar year



3.4.4 Robustness and alternative interpretations

Table 3 presents results from several robustness analyses with our baseline results for the number of workers in Column 1. Column 2 reports estimates when we add firm fixed effects to the baseline specification, instead of fixed effects for the two groups (treated and comparison firms). For neither of the two systems does this change our conclusions. There are positive effects for the staff-selection system but not for the rules-selection system. In Column 3, we include characteristics of the hired worker when we match treated and comparison firms (we use the characteristics shown in Panel C of Table 1, except for unemployment duration). When we in these ways adjust for differences in workers characteristics we still obtain very similar results as in our main analyses. Next, Column 4 adjusts for unemployed workers' time in unemployment before the start of the job in the matching step. That is, treated firms hiring a subsidized worker after 7 months are compared to comparison firms hiring a long-term unemployed worker after 7 months of unemployment, and so on. This adjusts for any additional differences between the subsidized workers hired by the treated firms and the long-term unemployed workers hired by the comparison firms. Again, this leads to similar results as on our baseline analysis. All this suggests that the composition of workers does not drive our results.

In Columns 5–7, we match on larger sets of firm outcomes. Column 5 shows estimates when we add profits and value added, as well as all firm outcomes two years before the subsidy. In Column 6, we also match on pre-treatment investments one and two years before the subsidy, and Column 7

Table 3. *Estimates for number of workers, different specifications*

	Baseline	Firm FE	Match workers characteristics	Match time unemployed	More controls	Match investments	Match hirings	Age 25–54	Small firms	Medium firms
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Staff selection										
Year of treatment	0.03 (0.11)	0.08 (0.10)	0.00 (0.12)	0.12 (0.12)	0.03 (0.11)	0.01 (0.11)	-0.03 (0.11)	0.02 (0.14)	0.06 (0.11)	0.10 (0.21)
1–2 yrs. after treatment	0.21* (0.12)	0.21* (0.12)	0.33** (0.13)	0.30** (0.14)	0.17 (0.12)	0.25** (0.13)	0.19 (0.13)	0.24 (0.15)	0.25** (0.12)	0.19 (0.23)
3–5 yrs. after treatment	0.52*** (0.15)	0.52*** (0.14)	0.57*** (0.15)	0.45*** (0.17)	0.56*** (0.15)	0.50*** (0.15)	0.47*** (0.15)	0.44** (0.19)	0.63*** (0.14)	0.45* (0.27)
Average	11.48	11.48	11.48	11.76	11.48	11.48	11.48	11.53	7.05	17.92
No. observations	157,758	157,758	157,372	112,291	157,926	158,162	158,139	101,797	91,455	64,934

Continue to next page

Notes: Robustness of estimates for firm size regressions. Column (1): estimation with baseline Propensity Score (PS) specification used for the main results of the paper; Column (2): baseline specification augmented with firm fixed effects; Column (3): individual–level demographics of the hired workers added to the baseline PS specification; Column (4): results when sampling all PES unemployment spells longer than 30 days – with exit to either unsubsidized or subsidized job – and then matching treated and controls hirings based on residual time registered at the PES as unemployed (discretized through 36 monthly- and 4 biannual-dummies); Column (5): baseline PS specification augmented with (i) profits, log value added and per–worker productivity measured the pre–treatment year and with (ii) the change in these quantities as well as in firm size and wage sum between -2 and -1.; Column (6): baseline PS specification augmented with log net investments per worker in -1; Column (7): baseline PS specification augmented with hirings level from $t = -1$ to $t = 0$; Column (8): restrict to firms hiring 25 to 54-year old unemployed; Column (9): matched sample of firms having less than 10 employees the pre–treatment year; Column (10): matched sample of firms having 10 to 30 employees the pre–treatment year. Standard errors clustered at firm level in parentheses. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

Table 3 – continued from previous page

	Baseline	Firm FE	Match workers characteristics	Match time unemployed	More controls	Match investments	Match hirings	Age 25–54	Small firms	Medium firms
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel B: Rules selection										
Year of treatment	-0.20 (0.19)	-0.07 (0.18)	-0.15 (0.19)	-0.01 (0.21)	-0.16 (0.19)	-0.23 (0.20)	0.03 (0.19)	-0.18 (0.22)	-0.20 (0.19)	-0.30 (0.33)
1–2 yrs. after treatment	0.05 (0.21)	0.20 (0.19)	-0.02 (0.21)	-0.12 (0.25)	-0.12 (0.20)	-0.18 (0.20)	0.20 (0.20)	-0.02 (0.24)	-0.23 (0.21)	0.41 (0.37)
3–5 yrs. after treatment	0.02 (0.26)	0.21 (0.24)	-0.03 (0.28)	-0.08 (0.32)	-0.09 (0.27)	-0.14 (0.27)	0.17 (0.26)	-0.22 (0.34)	-0.14 (0.27)	0.23 (0.45)
Average	11.26	11.26	11.27	11.54	11.26	11.26	11.26	11.16	6.45	17.63
No. observations	62,807	62,807	62,952	42,041	63,261	62,657	62,841	39,324	35,410	27,271

Notes: Robustness of estimates for firm size regressions. Column (1): estimation with baseline Propensity Score (PS) specification used for the main results of the paper; Column (2): baseline specification augmented with firm fixed effects; Column (3): individual–level demographics of the hired workers added to the baseline PS specification; Column (4): results when sampling all PES unemployment spells longer than 30 days – with exit to either unsubsidized or subsidized job – and then matching treated and controls hirings based on residual time registered at the PES as unemployed (discretized through 36 monthly- and 4 biannual-dummies); Column (5): baseline PS specification augmented with (i) profits, log value added and per–worker productivity measured the pre–treatment year and with (ii) the change in these quantities as well as in firm size and wage sum between -2 and -1.; Column (6): baseline PS specification augmented with log net investments per worker in -1; Column (7): baseline PS specification augmented with hirings level from $t = -1$ to $t = 0$; Column (8): restrict to firms hiring 25 to 54-year old unemployed; Column (9): matched sample of firms having less than 10 employees the pre–treatment year; Column (10): matched sample of firms having 10 to 30 employees the pre–treatment year. Standard errors clustered at firm level in parentheses. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

reports estimates adjusting for the pre-treatment hiring rate one and two years before the subsidy. Here, the hiring rate is defined as the number of workers hired in the treatment year.[18] This way, we adjust for a large set of pre-treatment levels and trends in key dimensions. These robustness estimates are all similar to our baseline specification.

In a final robustness analysis, we exclude the oldest workers (above 54) and the youngest workers (below 25), because the sample statistics showed differences between treated and comparison firms in the fraction of young and old workers. Again, the results are very similar to our baseline results.

In order to shed more light on the allocation process, we split the sample into small firms (fewer than 10 workers in year zero) and medium-sized firms (10-30 workers). Interestingly, the results presented in Columns 9 and 10 of 3 reveal somewhat larger effects for the small firms; 3-5 years after the subsidy the effect is 0.63 workers for the small firms and 0.45 workers for the medium-sized firms. Notably, in relative terms, the difference is even more pronounced as small firms by definition have fewer employees to start with.

We now turn to exploring additional alternative interpretations of the differences between the two systems. In our analyses, we compare periods with partly different business cycle conditions and the unemployment rate at the time of the subsidized hiring may affect the impact of wage subsidy. To examine whether this affects our findings, Columns 1 and 2 of Table 4 report estimates for firms hiring under different business cycle conditions defined by high (above the median) vs. low (below the median) national-level unemployment rates during our sampling period. The results for the two systems are similar to those in our main results.

Another interesting aspect is the Great Recession which lead to increased unemployment rates in Sweden from the first quarter of 2009. Even though the impact of the Great Recession on the Swedish labor markets was, in fact, not particularly great (unemployment rate was 6.2% in 2007 and 8.6% in 2010), it may still affect our results since the effects in the medium-run for rules-selection system (firms sampled in 2007–08) are identified during the recession. To explore this, we split the sample by the unemployment rate four years after the treatment year. The idea is to compare the effects of subsidized jobs in firms in the two systems that face the same type of business conditions in the medium-run. The estimates from the two different samples reported in Columns 3 and 4 of Table 4 show that this does not change the interpretation of our results. The only notable difference is that for the staff-selection regime the effect for high unemployment periods are insignificant but the point estimate is very close to that for the low unemployment period.[19]

[18] We have also explored specifications where we adjust for for the share of hired workers and the number of ineligible workers (number of non-subsidized workers). Again, this leads to similar results.

[19] We have also divided the sample by the unemployment rate three and five years after the treatment leading to the same conclusions.

Table 4. *Firm size regressions by unemployment rate and immigrant status*

	Unemployment rate hiring year		Unemployment rate 4 years since hire		Immigrant status	
	Low	High	Low	High	Native	Immigrant
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Staff selection						
Year of	0.14	0.02	0.04	0.03	0.09	-0.26
treatment	(0.24)	(0.12)	(0.12)	(0.30)	(0.12)	(0.25)
1–2 yrs. after	0.31	0.19	0.20	0.30	0.40^{***}	0.03
treatment	(0.27)	(0.14)	(0.13)	(0.37)	(0.14)	(0.29)
3–5 yrs. after	0.51	0.52^{***}	0.52^{***}	0.55	0.59^{***}	0.34
treatment	(0.34)	(0.16)	(0.16)	(0.41)	(0.17)	(0.36)
Panel B: Rules selection						
Year of	-0.25	0.07	-0.10	-0.21	0.03	0.12
treatment	(0.21)	(0.45)	(0.51)	(0.20)	(0.21)	(0.37)
1–2 yrs. after	0.04	0.23	-0.24	0.03	-0.00	0.27
treatment	(0.23)	(0.51)	(0.60)	(0.21)	(0.25)	(0.42)
3–5 yrs. after	0.05	0.02	-0.36	0.05	0.27	0.23
treatment	(0.29)	(0.60)	(0.96)	(0.25)	(0.33)	(0.53)

Notes: Columns (1) and (2) show results for firm size regressions by partitioning firms as hiring when the monthly unemployment rate is above or below the 1998–2008 median national level, respectively. In columns (3) and (4) firms are partitioned according to the yearly unemployment rate 4 years since treatment as compared to the 1996–2012 median national level. Columns (5) and (6) report the coefficients for firms hiring native or immigrant long–term unemployed. All regressions include year fixed effects and use the matched sample. Standard errors clustered at firm level in parentheses. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

This provides suggestive evidence that the Great Recession cannot explain our findings. Also, note that we compare firms that hire workers with and without subsidies (equally affected by seniority rules), and that our outcomes are at firm (not worker) level.

Next, we explore the impact of the rising share of immigrants amongst the unemployed. Since there are more immigrants in the unemployed pool during the more recent rules-selection system, our findings may be sensitive to differences in effects between immigrants and natives. To test for this, we split the samples into firms hiring natives and immigrants.[20] The results are presented in Columns 5 and 6 of Table 4. The patterns appear robust, in particular if we focus on the impact on natives. As expected, the estimates become very

[20] For the few cases of multiple hirings, we use the modal immigrant status type of the hires, giving priority to migrants in case of ties.

imprecise for immigrants, in particular during rules-selection when the sample gets very small.

One possible interpretation of our main findings is that caseworkers are able to target firms that are or expect to grow faster in the future, despite identical pre-treatment trends. If so, the subsidies are allocated to firms that would have outperformed the comparison firms regardless of the subsidy. As already documented in both Table A.1 and Figures 3 and 4, the treated and comparison firms have very similar pre-treatment trends (both before and after matching the data), including in dimensions that we do not match on (most importantly profits and production). But this does not completely rule out the possibility of differences in forward-looking expectations. Hence, instead of solely focusing on pre-treatment trajectories, for a much more direct test we use data on investments. The idea is that investments capture expectations about future outcomes and forward-looking attitudes. We therefore study how the yearly net investments in machinery and land differ between the treated and comparison firms in the subsidy year and the year before the subsidy.

Table 5. *Effects on firms investments; matched sample*

	Net investments per worker	
	logs	level
Panel A: Staff selection		
Pre–treatment year	0.02 (0.04)	-2.39 (5.82)
Year of treatment	0.06 (0.04)	12.78 (11.17)
Panel B: Rules selection		
Pre–treatment year	-0.10 (0.07)	-15.63 (9.65)
Year of treatment	-0.18** (0.07)	-17.12* (9.01)

Notes: Firm investments regressions using the matched sample. The outcomes are defined considering the yearly amount invested in machinery and land net of disinvestments, both in logs and in levels. The Propensity Score specification did not include investments among the pre–treatment controls. Standard errors clustered at firm level in parentheses. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

The estimates, provided in Table 5, reveal no significant differences in the staff-selection system. This result suggests that the fact that treated firms outperform comparison firms is not explained by caseworkers targeting firms with better forward-looking expectations, as captured by investments at least. In the rules-selection system the results are very different, however. The evidence suggests *lower* investments among the treated firms in the subsidy year. Comparing across the two regimes, the results thus suggest that caseworkers are able to select away firms with lower-than-average future expectations and investment rates. These businesses are instead more likely to use the subsidies during the rules-selection setting.

3.5 Summary and conclusions

In this paper we study how targeted wage subsidies schemes are related to firm performance. We find that subsidies can have a very positive sustained effect on a range of firm production and productivity measures, including firm size, wage sum, profits, value added and per-worker productivity. This is robustly true in the setting (before 2007) when caseworkers needed to approve all subsidies.

However, the patterns are less robust after 2007 when caseworkers no longer were involved in the allocation process. Instead, results turn much smaller and, with two exceptions, statistically insignificant for subsidies falling under the rules-selection regime. In this period, the impact on firm survival is positive. In addition, treated firms have lower-than-average investments. A possible interpretation of these changed patterns is that caseworkers during the staff-selection regime prevented firms with poor expectations from receiving subsidies, a process which may have reduced the impact on the firm-survival margin if this process kept marginal firms from seeking treatment as a last resort. We try to test for alternative explanations, including those related to the business cycle (although the “Great recession" was quite mild in Sweden) and find no support for the alternative explanations, but we acknowledge that we cannot fully rule out that other factors contributed to the change in responses.

Overall, however, we do believe that our results should be interpreted as suggesting that our Swedish targeted wage subsidies in fact have not allocated subsidies to poor performing firms, at least during the period when caseworkers acted as gatekeepers. The starkest result of our paper is the relatively strong post-match performance of the treated firms during this period. But it should also be noted that surviving firms who hire through subsidies, even during the period *without* caseworker approval, appear to perform at least as well as other firms that hire unemployed workers.

Our paper adds to the growing, but still relatively scarce, literature on how ALMPs affect firm-level performance, employer-employee sorting, and the interplay between the two. Thus providing evidence in line with the recent call by Card et al. (2018) for more research on how public policies affect the allocation of workers across firms. The policy relevance of the results is apparent. The results suggest that i) concerns that targeted wage subsidies allocate resources to bad firms may be unwarranted and that ii) policy-makers who are worried about displacement effects may want to consider ensuring caseworkers’ approval of targeted wage subsidies since our results were unanimously positive during the period with caseworker approval.

Appendix

Table A.1. *Sample statistics for pre–treatment outcomes for the matched samples*

	Staff selection			Rules selection		
	Treated	Control	Difference	Treated	Control	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: outcomes matched in $t-1$						
No. of workers						
$t-5$	8.75	8.95	−0.20	9.28	9.34	−0.06
$t-4$	9.02	9.22	−0.20	9.12	9.33	−0.21
$t-3$	9.20	9.37	−0.17	9.15	9.26	−0.11
$t-2$	9.27	9.30	−0.04	9.29	9.44	−0.14
$t-1$	9.73	9.78	−0.04	10.09	10.06	0.03
Wage sum per worker						
$t-5$	104.44	104.11	0.33	122.43	124.80	−2.36
$t-4$	106.13	106.81	−0.68	123.36	124.25	−0.89
$t-3$	110.03	108.01	2.02	125.00	125.75	−0.75
$t-2$	111.53	110.31	1.22	126.00	126.46	−0.46
$t-1$	109.59	107.82	1.76	124.62	122.14	2.48
Separations						
$t-5$	0.24	0.26	−0.02	0.26	0.28	−0.02
$t-4$	0.25	0.25	0.00	0.26	0.27	−0.01
$t-3$	0.24	0.24	−0.01	0.26	0.24	0.01
$t-2$	0.24	0.26	-0.02^{*}	0.23	0.26	-0.02^{*}
$t-1$	0.23	0.23	−0.01	0.22	0.23	−0.01
Panel B: outcomes not matched						
Profits (Th. SEK)						
$t-5$	368.85	401.72	−32.86	322.62	344.51	−21.90
$t-4$	302.34	382.75	−80.41	323.72	314.14	9.58
$t-3$	281.52	371.97	−90.46	344.66	369.42	−24.75
$t-2$	299.03	337.15	−38.12	403.10	355.92	47.18
$t-1$	332.20	386.26	−54.07	465.85	437.44	28.42

Continue to next page

Pre–treatment average outcomes matched in $t-1$ (Panel A) or not matched (Panel B), where t is the time when the firm hires. All firm–level outcomes computed using the matched samples. Monetary values are inflation-adjusted (base year: 2000). Separations are normalized by firm size. Wage sum and total profits measured in 1000 SEK, total firm value added in log 1000 SEK. Columns (3) and (6) report the differences in the averages for treated and control firms (hiring with or without a subsidy, respectively) in the two regimes. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

Table A.1 – continued from previous page

	Staff selection			Rules selection		
	Treated	Control	Difference	Treated	Control	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
Profits per worker						
$t-5$	96.59	100.73	−4.14	79.22	95.40	−16.18
$t-4$	80.81	88.55	−7.74	75.07	76.25	−1.18
$t-3$	70.67	89.00	−18.33	84.36	87.27	−2.91
$t-2$	74.96	77.32	−2.35	88.93	77.96	10.98
$t-1$	74.83	83.94	−9.11	91.56	84.11	7.45
Log value added						
$t-5$	7.11	7.14	−0.03	7.19	7.12	0.06
$t-4$	7.08	7.14	−0.06	7.15	7.13	0.01
$t-3$	7.11	7.12	−0.01	7.14	7.16	−0.02
$t-2$	7.12	7.12	0.00	7.18	7.18	0.00
$t-1$	7.14	7.12	0.02*	7.28	7.23	0.05*
Value added per worker						
$t-5$	389.97	408.84	−18.87	445.58	427.75	17.83
$t-4$	378.48	402.19	−23.70	420.14	428.43	−8.29
$t-3$	375.53	401.39	−25.86	416.45	427.23	−10.78
$t-2$	371.87	384.97	−13.10	420.86	423.11	−2.25
$t-1$	385.25	409.74	−24.48	426.90	442.15	−15.25
Tot. investments						
$t-5$	206.43	257.03	−50.60	211.38	208.33	3.05
$t-4$	204.51	206.39	−1.88	137.22	126.79	10.43
$t-3$	208.25	200.21	8.05	118.22	147.59	−29.37
$t-2$	191.41	172.72	18.69*	145.67	113.83	31.84*
$t-1$	228.83	222.96	5.87	163.58	178.20	−14.62
Tot. investments per worker						
$t-5$	44.21	60.28	−16.07	49.07	48.58	0.49
$t-4$	47.21	39.56	7.65	38.22	17.83	20.39
$t-3$	54.85	49.02	5.83	32.61	33.94	−1.33
$t-2$	46.55	42.30	4.25	34.22	33.00	1.22
$t-1$	52.92	52.74	0.18*	37.39	46.82	−9.43*
Firm survival						
$t-5$	0.64	0.65	−0.01	0.65	0.65	−0.00
$t-4$	0.71	0.72	−0.01	0.72	0.72	−0.00
$t-3$	0.79	0.79	−0.00	0.80	0.79	0.01
$t-2$	0.88	0.87	0.01**	0.90	0.89	0.01

Pre–treatment average outcomes matched in $t-1$ (Panel A) or not matched (Panel B), where t is the time when the firm hires. All firm–level outcomes computed using the matched samples. Monetary values are inflation-adjusted (base year: 2000). Separations are normalized by firm size. Wage sum and total profits measured in 1000 SEK, total firm value added in log 1000 SEK. Columns (3) and (6) report the differences in the averages for treated and control firms (hiring with or without a subsidy, respectively) in the two regimes. *, ** and *** denote significance at the 10, 5 and 1 percent levels.

Table A.2. *Hired workers' characteristics before matching*

	Treated firms		Control firms	
	Subsidized hires (1)	All hires (2)	Unsubsidized hires (3)	All hires (4)
Panel A: Staff selection				
Age: 24 or less	0.26	0.31	0.11	0.27
Age: 25–34	0.32	0.29	0.33	0.30
Age: 35–44	0.21	0.20	0.26	0.22
Age: 45–54	0.14	0.13	0.18	0.14
Age: 55–64	0.07	0.06	0.12	0.08
Age: 65 or more	0.00	0.00	0.00	0.01
Immigrant	0.21	0.17	0.22	0.19
Married	0.29	0.27	0.35	0.29
Male	0.69	0.67	0.59	0.59
Education: Compulsory	0.20	0.24	0.20	0.24
Education: Secondary	0.66	0.59	0.59	0.55
Education: Upper	0.15	0.17	0.20	0.22
Firm hirings	1.06	4.80	1.05	5.56
Panel B: Rules selection				
Age: 24 or less	0.17	0.30	0.11	0.29
Age: 25–34	0.25	0.24	0.31	0.27
Age: 35–44	0.23	0.20	0.26	0.21
Age: 45–54	0.18	0.14	0.19	0.14
Age: 55–64	0.18	0.10	0.13	0.08
Age: 65 or more	0.00	0.01	0.01	0.01
Immigrant	0.32	0.25	0.30	0.26
Married	0.36	0.29	0.35	0.29
Male	0.66	0.64	0.58	0.59
Education: Compulsory	0.21	0.25	0.20	0.24
Education: Secondary	0.59	0.54	0.55	0.52
Education: Upper	0.21	0.21	0.25	0.25
Firm hirings	1.05	4.97	1.04	6.15

Notes: Characteristics of workers hired by the treated and control firms before matching. Columns (1) and (3) report the characteristics of the long-term unemployed workers hired in the first quarter with or without a subsidy, respectively. Columns (2) and (4) show the characteristics of all workers hired the same year in which the long-term unemployed were hired with or without a subsidy.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abowd, J. M., Kramarz, F., and Margolis, D. N. (1999). High wage workers and high wage firms. *Econometrica*, 67(2):251–333.

Andersson, F., Holzer, H., Lane, J., Rosenblum, D., and Smith, J. A. (2016). Does federally-funded job training work? Nonexperimental estimates of WIA training impacts using longitudinal data on workers and firms. CESifo Center for Economic Studies and ifo Institute, Working Paper No. 6071.

Bishop, J. H. and Montgomery, M. (1993). Does the targeted jobs tax credit create jobs at subsidized firms? *Industrial Relations*, 32(3):289–306.

Blasco, S. and Pertold-Gebicka, B. (2013). Employment policies, hiring practices and firm performance. *Labour Economics*, 25:12–24.

Blundell, R., Dias, M. C., Meghir, C., and van Reenen, J. (2004). Evaluating the employment impact of a mandatory job search program. *Journal of the European Economic Association*, 2(4):569–606.

Cahuc, P., Carcillo, S., and Barbanchon, T. L. (2018). The effectiveness of hiring credits. *The Review of Economic Studies*, 86(2):593–626.

Calmfors, L., Forslund, A., and Hemström, M. (2002). Does active labour market policy work? Lessons from the Swedish experiences. IFAU working paper, 2002:4.

Card, D., Cardoso, A. R., Heining, J., and Kline, P. (2018). Firms and labor market inequality: Evidence and some theory. *Journal of Labor Economics*, 36(S1):S13–S70.

Card, D., Heining, J., and Kline, P. (2013). Workplace heterogeneity and the rise of West German wage inequality. *The Quarterly Journal of Economics*, 128(3):967–1015.

Card, D., Kluve, J., and Weber, A. (2010). Active labour market policy evaluations: A meta-analysis. *The Economic Journal*, 120(548):F452–F477.

Card, D., Kluve, J., and Weber, A. (2017). What works? A meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association*, 16(3):894–931.

Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., and Zamora, P. (2013). Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *The Quarterly Journal of Economics*, 128(2):531–580.

Egebark, J. and Kaunitz, N. (2018). Payroll taxes and youth labor demand. *Labour Economics*, 55:163–177.

Engström, P. and Johansson, P. (2012). The medical doctors as gatekeepers in the sickness insurance? *Applied Economics*, 44(28):3615–3625.

Ferracci, M., Jolivet, G., and van den Berg, G. J. (2014). Evidence of treatment spillovers within markets. *Review of Economics and Statistics*, 96(5):812–823.

Gautier, P., Muller, P., van der Klaauw, B., Rosholm, M., and Svarer, M. (2018). Estimating equilibrium effects of job search assistance. *Journal of Labor Economics*, 36(4):1073–1125.

Harkman, A. (2002). Vilka motiv styr deltagandet i arbetsmarknadspolitiska program? (What determines participation in labor market programs?). IFAU working paper, 2002:9.

Kangasharju, A. (2007). Do wage subsidies increase employment in subsidized firms? *Economica*, 74(293):51–67.

Kluve, J. (2010). The effectiveness of European active labor market programs. *Labour Economics*, 17(6):904–918.

Lalive, R., Landais, C., and Zweimüller, J. (2015). Market externalities of large unemployment insurance extension programs. *American Economic Review*, 105(12):3564–3596.

Lechner, M., Wunsch, C., and Scioch, P. (2013). Do firms benefit from active labour market policies? University of Basel Discussion Paper.

Lise, J., Seitz, S., and Smith, J. A. (2004). Equilibrium policy experiments and the evaluation of social programs. NBER Working Paper, No. 10283.

Lundin, M. (2000). Anställningsstödens implementering vid arbetsförmedlingarna (Employment support implementation at the public employment service offices). IFAU working paper, 2000:4.

Markussen, S., Røed, K., and Røgeberg, O. (2013). The changing of the guards: Can physicians contain social insurance costs? IZA Discussion Papers, No. 7122.

Neumark, D. (2013). Spurring job creation in response to severe recessions: Reconsidering hiring credits. *Journal of Policy Analysis and Management*, 32(1):142–171.

Pallais, A. (2014). Inefficient hiring in entry-level labor markets. *American Economic Review*, 104(11):3565–3599.

Saez, E., Schoefer, B., and Seim, D. (2017). Payroll taxes, firm behavior, and rent sharing: Evidence from a young workers' tax cut in Sweden. NBER Working Paper, No. 23976.

Sjögren, A. and Vikström, J. (2015). How long and how much? Learning about the design of wage subsidies from policy changes and discontinuities. *Labour Economics*, 34:127–137.

Song, J., Price, D. J., Guvenen, F., and Bloom, N. (2018). Firming up inequality. *The Quarterly Journal of Economics*, 134(1):1–50.

4. Comparing Discrete-Time Multi-State Models Using Dissimilarities

with Raffaella Piccarreta and Marco Bonetti

Acknowledgments: We thank Liming Cai for providing the SAS scripts to estimate the MSLT model, and for suggestions about its extension to the analysis of the FFS data. We are grateful to Pietro Tebaldi for having extended his user-written Stata commands *mstat* and *mtest* to allow for the direct input of dissimilarity matrices.

4.1 Introduction

We consider the general case where we observe n subjects and the activities (or states) that they experience over time, so that a trajectory – a finite *sequence* or ordered collection of states – is observed for each cross-sectional unit. There are many areas where data of this type is collected and analyzed. For example, in epidemiology the conditions of treated individuals are typically observed over time, and in each period the patient can experience alternative focal events such as remission, occurrence of one or more types of diseases, or death. In demography, one may be interested in studying the transition of individuals to adulthood with respect to family formation or employment career. in economics, typical event history sequences concern transitions between employment, unemployment and out-of-labor force.

In this context, common objects of interest are the event of having experienced a state, the timing of the transition towards a state, or the length of the permanence in a state (see e.g., Lawless, 2003). Here we adopt an "holistic" perspective and focus on the trajectory as a whole rather than on the timing or occurrence of specific events.

The states that an individual may experience can be classified in different ways. For instance, they can be recurring (when one state can be visited more than once), transient (when a subsequent transition to another state can be observed), or absorbing (when transitions to other states cannot be further observed after visiting the current one). Here we consider state trajectories that include all of these kinds, and such that they will in general be right censored, with evolution over a specified period (e.g., for a period of 10 years after having received a treatment, or between the ages of 15 and 30).

We focus on parametric models describing the evolution of individual trajectories with respect to a set of covariates. Multi-state models are a popular approach to describe the occurrence of events of different kinds over time. For a review of multi-state models and their implementation, see for instance Putter et al. (2007) and Beyersmann et al. (2012), who focus on models for the hazard of transitioning to specific states within the context of (unidirectional) multi-state models with no recurrent events. Note that traditional survival analysis is a special case of a multi-state model with just one absorbing state beyond the initial state. Competing risks models can also be seen as special cases of multi-state model with a common initial state and two or more final absorbing states.

Such models can prove useful from the descriptive point of view to identify relevant covariates and/or to assess the effect of covariates on the evolution of trajectories. Here we are interested in studying and comparing the predictive power of competing models, that is their ability to generate trajectories that are "similar" to those observed. Specifically, our main goal is to propose criteria to suitably compare collections of pairwise dissimilarities computed across observed and model-generated (i.e. simulated) sequences. In particular, we propose three alternative distance-based criteria for the implementation of these comparisons.

In our analyses we proceed as follows. First, we estimate competing multistate models by using observed sequences. We use data collected as part of the Fertility and Family Surveys (FFS) study, conducted in the 1990's in selected member States of the United Nations Economic Commission for Europe (Latten and De Graaf, 1997). We focus on women from the Netherlands, and in particular on the relationship between their childbearing and family formation trajectories and a set of time-fixed background variables. Next, we use the estimated model functional forms to simulate event histories. Finally, we implement alternative distance-based criteria to assess the dissimilarity between simulated and observed sequences. Since the models will in general relate trajectories to discrete covariates, we will compare the dissimilarities between the observed trajectories with the corresponding predicted sequences conditionally on specific combinations of the covariates levels observed in the data. To make such comparison meaningful, we will restrict our analyses to combinations of covariate values with high enough frequencies. Moreover, the comparisons will be kept separated rather than pooled into an overall measure across covariates values, since the performance of a model may differ across the covariate space.

The paper is structured as follows. In Section 4.2 we describe the data and two alternative semi-Markov models for the probability of transitioning from one state to another while accounting for a set of covariates, the previously visited state and the time spent in the previous state. Such models are the Multi-State Life Table (MSLT) approach, described in Cai et al. (2006, 2010) and the State Change model (SCM) introduced by Bonetti et al. (2013). Both approaches model the probability of transitioning towards the next state of the event history in discrete time, and they were applied to the analysis of the FFS data in Lombardi (2012). In Section 4.3 we describe alternative distance-based methods that can be used to compare event histories. Section 4.4 illustrates the assessment of the predictive accuracy of the MSLT and SCM models when applied to the FFS data. We conclude with some comments in Section 4.5.

4.2 Data and two event history analysis models

4.2.1 The Fertility and Family Survey data

We consider data collected as part of the Fertility and Family Surveys (FFS) study, conducted in the 1990's in selected member States of the United Nations Economic Commission for Europe (Latten and De Graaf, 1997). The same data was analyzed in Bonetti et al. (2013) and in Lombardi (2012), who focused on 1897 women from the Netherlands born between 1953 and 1962. In particular, the interest is on women's childbearing and family formation patterns. For each woman the ordered collection of the monthly states experienced between 18 and 30 years of age can be summarized by the sequence s $= (s_1, \ldots, s_P)$ ($P = 144$ months for all women).

Specifically, the states taken into account are: living without a partner and having no children (single, N), married without children (M), in unmarried cohabitation without children (U), single with at least one child (NC), married with at least one child (MC), and cohabiting and having at least one child (UC).

A more compact representation of a woman's trajectory can be obtained by listing the distinct visited states $\mathbf{v} = (v_1, v_2, \ldots, v_h)$ (*states* sequence) and the durations $\mathbf{t} = (t_1, t_2, \ldots, t_h)$ of the permanence in each state (*durations* sequence), with h indicating the observed total number of states visited. For example, for a woman who lived without a partner for 22 months, then cohabited for 27 months, then lived as single again for 31 months, and finally married and remained in that state for 64 more months, it is $h = 4$, $\mathbf{v} =$ (N, U, N, M), and $\mathbf{t} = (22, 27, 31, 64)$. Although some states can be visited more than once, the "children" state is an *absorbing* state: after the first child is born, the woman cannot return to any of the "no children" states (deaths of children are not considered).

The goal in the analysis of Bonetti et al. (2013) and Lombardi (2012) was to relate the sequences to a set of categorical (baseline) socio-demographic characteristics: birth cohort, level of education, religious status, and having or not separated or divorced parents. We distinguish between the two cohorts 1953–1957 and 1958–1962. *Education* is based on the years of education received after the age of 15, and it is grouped into three classes: women who interrupted their studies (none), those who proceeded with an additional 3 years of education, and those who received more than 3 years of additional education. *Religion* indicates whether a woman declared herself as being religious or not. Finally, *Divorce* indicates whether a woman's parents are separated or divorced.

Note that the survey collected information on women at the time of the interview, i.e. after the age of 30. As a consequence, the use of *Divorce* to explain or predict the sequences may be questionable. Given that most of parental divorces take place during adolescence, however, the assessment of the effect of parents' being divorced is likely to be informative even if we consider the variable as time-fixed. The use of *Religion* might also be problematic because of the possibility of (rare) changes of religious status during one's life. Education, on the other hand, is a baseline variable as it refers to events that occurred before the age of 18.

4.2.2 Two event history analysis models

As we have seen, the data consist of state transitions in discrete time. Even if transitions occur continuously, when available data is interval censored, a common approach is to treat time as discrete, provided that the time intervals are sufficiently narrow and transitions are not too frequent. These two informal conditions are consistent with the implicit assumptions that: (a) only one transition can happen within each time interval; and (b) at the beginning of each interval individuals are at risk of experiencing the allowed transitions, which may occur in correspondence of an unknown but random point within the time interval (Cai et al., 2010).

The standard approach to model $P_{q \to r}$, the probability of transitioning from state q to state r, between times t and $(t+1)$, is the generalized multinomial logistic regression (see e.g., Agresti, 2002):

$$\log\left\{\frac{P_{q\to r}(\mathbf{X}_t)}{P_{q\to M}(\mathbf{X}_t)}\right\} = \mathbf{Y}_t^T\alpha_r + \mathbf{X}_t^T\beta_r, \quad q = 1,\ldots,M;\ r = 1,\ldots,M-1, \tag{4.1}$$

where M is a reference state, $\mathbf{X}_t$ is the vector of explanatory variables at time t (also including an intercept term), and $\mathbf{Y}_t$ is a vector of $(M-1)$ dummy variables, whose m-th element indicates whether or not at time t the visited state was the m-th one.

The transition probabilities can then be written as:

$$P_{q\to r}(\mathbf{X}_t) = \exp\left\{\mathbf{Y}_t^T\alpha_r + \mathbf{X}_t^T\beta_r\right\}\left(1 + \sum_{m=1}^{M-1}\exp\left\{\mathbf{Y}_t^T\alpha_m + \mathbf{X}_t^T\beta_m\right\}\right)^{-1}, \tag{4.2}$$

for $q = 1,\ldots,M;\ r = 1,\ldots,M-1$. The probability of transitioning to the reference state is $P_{q\to M}(\mathbf{X}_t) = \left(1+\sum_{m=1}^{M-1}\exp\left\{\mathbf{Y}_t^T\alpha_m + \mathbf{X}_t^T\beta_m\right\}\right)^{-1}$.

Note that in (4.1) we have $M-1$ response categories for the *arrival* state, r, each paired with the reference one. Thus, the *starting* state q is defined through $\mathbf{Y}_t$ and there are $(M-1)$ non-redundant logits, each characterized by the vector parameter $\theta_r = \left(\alpha_r^T, \beta_r^T\right)^T$, which is therefore specific to the arrival state, r. More parsimonious specifications can be obtained by constraining some of the model parameters to be equal across transitions.

The specification in (4.1), with the *current state* term included, is flexible since it allows one to *test* the significance of the partial effect of being in state q on the transition probabilities. An alternative strategy is to omit the $\mathbf{Y}_t$ term in (4.1) and to fit M separate multinomial logistic models applied to the sub-samples of cases having $Y_{qt} = 1$ (e.g., Laditka and Wolf, 1998). Such alternative strategy is preferable when one wants to allow a different parameter set for each *starting* state.

In (4.1), the probabilities of transitioning across states depend on the past history only through the current state or, equivalently, the state visited after time t only depends on the state experienced at t. In some cases, it is more realistic to let the transition probabilities $P_{q\to r}(\cdot)$ depend not only on q, but also on the length of the permanence in the current state (*duration* of the state) since the most recent entry in that state.

The resulting less restrictive models are referred to with different names, but they are all characterized by this *semi-Markov* property. Two specific such models were applied to the FFS data in Lombardi (2012) and in Bonetti et al. (2013).

The first model is the Multi-State Life Table (MSLT) model proposed by Cai et al. (2006, 2010). The probability of transitioning from state q to r is modeled as:

$$\log\left\{\frac{P_{q\to r}(\mathbf{X}_t)}{P_{q\to M}(\mathbf{X}_t)}\right\} = \mathbf{Y}_t^T\alpha_r + \mathbf{X}_t^T\beta_r + \gamma_r\delta_t \tag{4.3}$$

for $q = 1,\dots,M$; $r = 1,\dots,M-1$, and where δ_t is the time spent in the current state (since the most recent entrance into it). The duration effect γ_r is assumed to be specific for the different *arrival* states, and constant across the current state. If the current state is an absorbing state, then the probability of transitioning to another state is set equal to zero.

A feature of MSLT is that it adjusts for the time spent in the current state, but it does not directly model such durations. Notice that the duration effect can enter the model through, say, polynomial terms, and that it is also possible to allow the duration effect to vary with one or more covariates by adding interaction terms.

A possible alternative to the MSLT model is the State Change Model (SCM) described in Bonetti et al. (2013). SCM separately models the time to the next *generic* transition to a different state, and the probability of transitioning to specific states conditionally on a transition occurring. Regression models are built for the two parts of the model: time-to-event regression models for the duration part, and conditional multinomial regression models relating the probabilities of transitioning to the different arrival states to a set of covariates and to the observed duration up to the transition.

Specifically, the time to the next transition may be assumed to follow the geometric distribution with a parameter p_j that depends on the covariates through the logit link:

$$p_j = \exp\left\{\mathbf{Z}_j^T\delta\right\}\left(1+\exp\left\{\mathbf{Z}_j^T\delta\right\}\right)^{-1}, \tag{4.4}$$

where $\mathbf{Z}_j$ summarizes the information available when the j-th state is entered, and δ is the vector of regression parameters. The covariates in $\mathbf{Z}_j$ can be time-varying, and $\mathbf{Z}_j$ may include also the last state visited before entering the j-th one.

As for the probability of transitioning from one state to another, a variation of generalized conditional multinomial regression models is used. Specifically, the probability of transitioning from state q to state r at the j-th transition is modelled as:

$$P_{q\to r,j} = \exp\left\{\mathbf{X}_j^T \beta_{qr}\right\}\left(1+\sum_{m=1,m\neq q}^{M-1} \exp\left\{\mathbf{X}_j^T \beta_{qm}\right\}\right)^{-1}, \tag{4.5}$$

with $q = 1,\ldots,M$; $r = 1,\ldots,M-1$ and $r \neq q$. Here $\mathbf{X}_j$ summarizes the information available at the moment when the j-th state is entered, and it may or may not coincide with the $\mathbf{Z}_j$ used in (4.4). The probability of transitioning into the same state, $P_{q\to q,j}$ is set to zero for all q, j.

A possible drawback of SCM is an excessive number of parameters to estimate. This can be mitigated by constraining some of them to be equal to zero. Bonetti et al. (2013) suggest a preliminary nonparametric screening procedure to select the most promising explanatory variables.

The main difference between MSLT and SCM is that while the former allows for an effective description of covariate effects on the transition probabilities, the latter allows for a direct interpretation of covariate effects on the time-to-event distribution for the time until the next transition.

The FFS sequences describing childbearing and family formation patterns were analyzed in Bonetti et al. (2013) and in Lombardi (2012) using the SMC and the MSLT, respectively. In both cases, the sparseness of the data (within combinations of covariate values) did not allow fitting the models on the various "Children" states. Therefore, the three states NC, MC, and UC were grouped into a unique absorbing state "C". Since transitions from states with children to states without children are not possible, all the parameters regarding these transitions were set to zero. The marginal frequencies of all transitions in the observed data are shown in Table 1.

Table 1. *Frequencies of transitions from row state to column state in the FFS data*

	N	U	M	C
N	0	911	920	40
U	178	0	554	47
M	32	8	0	1140

The initial explanatory variables were Cohort, Education, Religion, and Divorce. Also, the transition from one state to another at a given moment was related to the previously visited state (*Previous* $=$ M, U), and to the *Age* at the time of the transition. In addition, SCM also included the time spent in the state before the transition (*TVal*). Note that the latter covariates change at each visited state.

Table A.1 in the Appendix reports the maximum likelihood estimates for the MSLT model, whereas Tables A.2 and A.3 report results obtained for the duration and for the transition components of the SCM model. In both cases, only the variables which turned out to be significant for at least one of the conditional transitional probabilities are reported.

In particular, for the MSLT model the Wald-based backward elimination procedure led to selecting the entire set of regressors, including *Age*Age* and the duration *Tval*. For the SCM, the results of the variable selection procedure yielded for the duration part of the model the covariates *Age*, *Previous state*, and *Cohort* (binary variable indicating whether a woman is in the younger cohort, 1958–1962). For the transition component of the SCM the following covariates were significant for at least some of the conditional transitional probabilities: *Tval* (Time spent in the previous state), *Age*, *Education*, *Religion*, *Divorce* and *Cohort*. Note that the overall interpretation of the effects of the two time-varying covariates *Tval* and *Age* is rather complicated, since they enter both the components of the SCM.

Both SCM and MSLT can be used to describe the trajectories' generating mechanism. In particular, for both models it is possible to generate, via micro-simulation, event histories starting from the estimated parameters and conditionally on the observed combinations of covariates. The simulated trajectories can then be compared to the observed event histories to evaluate the model's appropriateness in terms of prediction from the sample at hand.

4.3 Comparing observed and simulated event histories

In what follows, we indicate by S the set of all n observed sequences, and by $\mathrm{S}(\mathbf{x})$ the set of the sequences observed for the vector $\mathbf{x}$ of covariate values. Similarly, we let $\widehat{\mathrm{S}}$ and $\widehat{\mathrm{S}}(\mathbf{x})$ indicate the corresponding sets for sequences simulated from an estimated model. Our goal is to compare the observed and the predicted trajectories through their level of *similarity*.

to this aim, a preliminary step is to describe how close the simulated sequences are to the observed ones with respect to the frequencies and the durations of visits. This may be done qualitatively, either conditionally within values of the explanatory variables or marginally for a specific level of just one of the covariates.

A further step is to use dissimilarity-based approaches to compare observed and simulated sequences. Measuring the dissimilarity between two sequences is a standard problem in sequence analysis (SA), one of the most common approaches to describe life courses when adopting an holistic perspective (i.e. when focusing on event histories as a whole, rather than only on the timing of events or on the states visited). Note that the models described in the previous section do not necessarily apply only to life courses data, and the techniques introduced in the context of SA can indeed be fruitfully applied whenever analyzing sequence data.

In SA, the starting point is to choose a suitable measure of the pairwise distance (or dissimilarity) between trajectories A variety of methods exist. We focus on Optimal Matching (OM), an alignment technique that was originally introduced in molecular biology to study protein or DNA sequences (Sankoff and Kruskal, 1983), and that was later extended to the study of life courses in sociology (Abbott, 1995). OM explicitly quantifies the *effort* needed to transform one sequence into another. There exist three elementary transformation operations: (i) insertion of a state; (ii) deletion of a state; (iii) substitution of a state with another one. Each operation has a respective cost, and the dissimilarity between two sequences is defined as the minimum total cost needed to transform a sequence into the other one. Substitution costs may be assigned subjectively on the basis of theory or a priori knowledge (see, among the others, McVicar and Anyadike-Danes, 2002). Alternatively, one may follow a data-driven approach and relate substitution costs to transition frequencies, so that frequent transitions are considered less costly than rare transitions (Rohwer and Pötter, 2004).[1]

See the next section for further details on the specification of the dissimilarity measure that we use in our analyses. Note that the distance-based criteria that we propose below can be implemented with any distance measure. Hence, to simplify the discussion of the comparison methods that we propose, assume that a properly defined dissimilarity measure exists. We now turn to the methods used to compare observed and model-generated sequences.

Using the $\mathbf{S}(\mathbf{x})$ and $\widehat{\mathbf{S}}(\mathbf{x})$ notation, a first approach is to evaluate, for a given combination of the covariates levels, the dissimilarities between the sequences in $\widehat{\mathbf{S}}(\mathbf{x})$ and a properly defined *summary* of those in $\mathbf{S}(\mathbf{x})$ i.e. a "typical" sequence observed in the sample corresponding to $\mathbf{x}$. To summarize $\mathbf{S}(\mathbf{x})$, we suggest to use the medoid $\overline{\mathbf{S}}(\mathbf{x})$, defined as the sequence having the minimum total dissimilarity from all the other sequences in $\mathbf{S}(\mathbf{x})$. As such, it can be considered a meaningful central tendency measure and summary of $\mathbf{S}(\mathbf{x})$ (Sheikh et al., 2007; Aassve et al., 2007). One then measures how close the simulated sequences are from the medoid of the observed sequences, separately for the two models as well as for the observed sequences. The degree of similarity can then be assessed based on summary statistics (e.g., mean or standard deviation) of all the dissimilarities between the sequences and the medoid.

Other possibilities arise when one considers entire distributions of between-sequence dissimilarities, the so-called *interpoint distance distributions* (IDDs). This approach is based on the estimated cumulative distribution function of the dissimilarities between sequences within a group or across two groups. Specifically, consider a distribution $F_{\mathbf{Y}}$ taking values in a possibly highly dimensional (as in this case) space $\mathcal{Y}$. Define the random variable $D = d(\mathbf{Y}_1, \mathbf{Y}_2)$ as the dissimilarity between the two *i.i.d.* elements $\mathbf{Y}_1$ and $\mathbf{Y}_2$ extracted from $F_{\mathbf{Y}}$. The IDD is the distribution $F_D(d) = P(D \leq d)$ of D.

[1] Even if criticized (see Aisenbrey and Fasang, 2010, for an in-depth review of the most relevant criticisms and related proposals to overcome them) OM remains the most common way to measure dis/similarity between trajectories in sequence analysis.

Let D indicate the "distance" between two observations as measured by any symmetric (non-negative) function of the two observations, and in particular by any dissimilarity measure that may be relevant for the particular problem at hand. To estimate $F_D(d)$, an *i.i.d.* sample of n cases $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ can be drawn, and inference can be based upon the set of the $\binom{n}{2}$ pairwise (dependent) dissimilarities between them. In particular, Bonetti and Pagano (2005) consider the empirical cumulative density function (ECDF):

$$F_n(d) = \frac{2}{n(n-1)} \sum_{1 \leq i < h \leq n} 1(d_{ih} \leq d),$$

where $d_{ih} = d(\mathbf{y}_i, \mathbf{y}_h)$ indicates the distance or dissimilarity between the i-th and the h-th sample observations.

The ECDF of all the pairwise distances evaluated at a finite number of values along the distance axis has an asymptotic multivariate normal distribution. If one considers a grid of bins-defining points $d_1, \ldots, d_K$ (bins) along the distance axis, and the vector of the ECDF evaluated at the end of each bin, then such vector may be written as $F_n(\mathbf{d}) = \{F_n(d_1), ..., F_n(d_K)\}$. In particular, the comparison between $F_n(\mathbf{d})$ and a null hypothesis distribution for D (say, $F_0(\mathbf{d})$) can be based on the quadratic form:

$$M = \{F_n(\mathbf{d}) - F_0(\mathbf{d})\}^{\mathrm{T}} \widehat{\mathbf{\Sigma}}^{-} \{F_n(\mathbf{d}) - F_0(\mathbf{d})\},$$

where $\widehat{\mathbf{\Sigma}}^{-}$ is a generalized inverse of the estimated variance-covariance matrix of $F_n(\mathbf{d})$. This statistic can be described as a Mahalanobis distance between the observed and the expected distribution of the distances discretized to the K bins. Under the null hypothesis that $D \sim F_0(\mathbf{d})$, the vector

$$\sqrt{n}\,[F_n(d_1) - F_0(d_1), \ldots, F_n(d_K) - F_0(d_K)]$$

converges in distribution as to a zero-mean multivariate normal random variable. While M converges in distribution to a chi-squared random variable as n diverges to infinity, empirical experience shows that the convergence is slow. For this reason, it is often preferable to use empirical testing routines, such as Monte Carlo or permutation testing (Bonetti and Pagano, 2005).

Manjourides (2009) extends this approach to the two-sample case, that is the situation when one wants to test whether two groups of multivariate observations follow the same distribution by comparing their interpoint distance distributions. This extension is most relevant here. For two groups G_1 and G_2 of observations (with sizes n_1 and n_2 respectively), let $F_{n_g}(\mathbf{d}) = [F_{n_g}(d_1), ..., F_{n_g}(d_k)]$, where $F_{n_g}(d)$ is the ECDF computed using only the subjects in group g, with $g = 1, 2$. The test statistic to verify the null hypothesis that the distribution of the distances is the same in the two groups is:

$$\tilde{M} = [F_{n_1}(\mathbf{d}) - F_{n_2}(\mathbf{d})]^{\mathrm{T}} \widehat{\mathbf{\Sigma}}^{-} [F_{n_1}(\mathbf{d}) - F_{n_2}(\mathbf{d})], \tag{4.6}$$

where $\widehat{\boldsymbol{\Sigma}}^{-}$ is the Moore-Penrose generalized inverse of the estimated variance-covariance matrix $\widehat{\boldsymbol{\Sigma}}$ of the vector $[F_{n_1}(\mathbf{d}) - F_{n_2}(\mathbf{d})]$. Inference can be based on the permutation distribution obtained by permuting the group labels of the observations.

The $\tilde{M}$ statistic can therefore be used to compare the dissimilarities of sequences in $\mathbf{S}(\mathbf{x})$ with those in $\widehat{\mathbf{S}}(\mathbf{x})$. Here we use this approach to determine whether two models predict "similar" sequences, i.e. sequences having the same IDD. The $\tilde{M}$ test can be implemented using permutation-based inference with the Stata functions `mstat` and `mtest` (Tebaldi et al., 2011).

As a third and last possibility, also based on interpoint distances, we suggest the application of a Wilcoxon-like test, as first suggested in Mosler (2002). The idea is to contrast the within-group dissimilarities (relative only to sequences in the same group) to the between-group dissimilarities (computed between sequences belonging to different groups) with a rank-based test statistic. Consider again the two groups of observations G_1 and G_2, and let $d_{i,h}^{(g)}$ (with $g = 1, 2$) denote the distance between the i–th and the h–th case in group g. Thus, there are a total of $\binom{n_1}{2}$ and $\binom{n_2}{2}$ *intra*-sample distances computed for cases within the same group. Also, let the distances $d_{i,h}^{(1,2)}$ with $i \in G_1$ and $h \in G_2$ be the *inter*-sample distances, calculated for cases that belong to different groups. After having sorted the set that includes *all* the inter- and intra-sample distances in ascending order and having assigned ranks to these distances, we can define the test statistic

$$T = \sum_{i=1}^{n_1} \sum_{h=1}^{n_2} R(d_{i,h}^{(1,2)}), \tag{4.7}$$

where $R(d_{i,h}^{(1,2)})$ denotes the rank of $d_{i,h}^{(1,2)}$ in the collection of all distances taken together.

If both samples come from the same distribution, then the generic inter-sample distance $d_{i,h}^{(1,2)}$ should follow the same marginal distribution as the generic intra-sample distances, $d_{i,h}^{(1)}$ or $d_{i,h}^{(2)}$. Mosler (2002) formalized the related hypothesis test, according to which we reject the null hypothesis when T is large. In particular, he derived the first two moments of T, and proposed an exact permutation-based non-parametric approach for testing with T. Such approach is also based on permuting the group labels of the observations.

4.4 An application to life course event histories

In this section we apply the methods previously introduced to compare the MSLT and SCM models described in Section 4.2. Note that the same methods proposed here can be used to compare the predictive performance of any two competing multi-state models. All sequences were generated from the transition probability structure implied by the MSLT and SCM models, by using the baseline covariates values and by plugging in the estimated parameter values reported in Tables A.1–A.3 in the Appendix.[2] Next, we assess the MSLT and SCM predictive performance in terms of the similarity between the sequences observed in the sample and those simulated through each of the two models.[3] The simulations were set up so that for each sequence type observed in the sample, we generated 100 sequences based on the individual's covariates values. This returns the two sets of simulated sequences $\widehat{\mathsf{S}}_{SCM}$ and $\widehat{\mathsf{S}}_{MSLT}$.

As a preliminary step, we compare the global features of the sequences in S and in the two sets of simulated sequences. We first focus on the state sequences, $\mathbf{v}$, in the three sets. In particular we compare the frequencies of the most frequent $\mathbf{v}$'s, the distribution of the visited states (irrespective of their duration), and the average duration of each visited state, as reported in Figure 1. Overall, 5732 states were visited (recall that each state can be visited more than once). The most visited state are N, M and C; a similar ordering is observed for the states durations. The traditional family formation pattern (N, M, C) is the most frequent state sequence, even though a relatively high proportion of women experienced cohabitation before marriage.

Moving to the simulated sequences, the characteristics of the MSLT-based sequences appear to be more similar to those of the observed ones, compared to the sequences generated using the SCM. In the latter case, the frequency and the average duration of the state N are larger compared to the observed ones, and the reverse holds for the state C. As for the visited states, the most frequent observed sequence is also the most frequent simulated sequence. Nonetheless, in the simulated sets some sequences have a different relevance compared to the sample. For example, in accordance with the distribution of the visited states, the SCM over-represents the sequence N by overestimating its relevance, and in general the same holds for the sequences including N; on the other hand, it under-represents the sequences including the state C. The MSLT model, instead, slightly over-represents the sequences including U.

To analyze the differences between observed and simulated sequences, it is also useful to consider the plot of the transversal state distributions, reported in Figure 2, i.e. the sequence of the distribution of the states for each month of observation (note that these graphs do *not* describe the transitions from one state to another). The same considerations made above can be drawn based on these plots: the MSLT model appears to better reproduce the distributions of the states in the sample, whereas the SCM tends to over-represent state N and to somewhat under-represent state C.

We now evaluate the differences in performance of the two models using the three dissimilarities-based criteria introduced in the previous section. In

[2] See Bonetti et al. (2013) and in Cai et al. (2010) for additional information. For simplicity we did not take into account sampling variability of the preliminary model estimation step.

[3] Optimal matching and sequence analysis were implemented with the R package *TraMineR*.

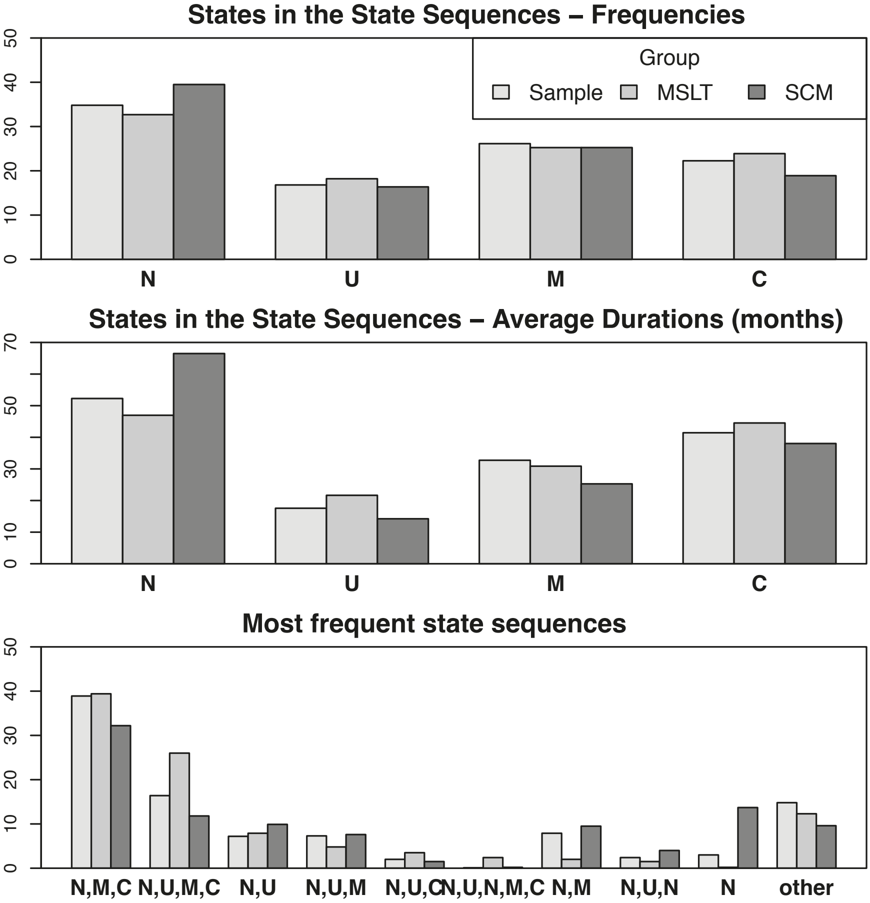


Figure 1. Description of sequences in the FFS dataset and in the simulated datasets

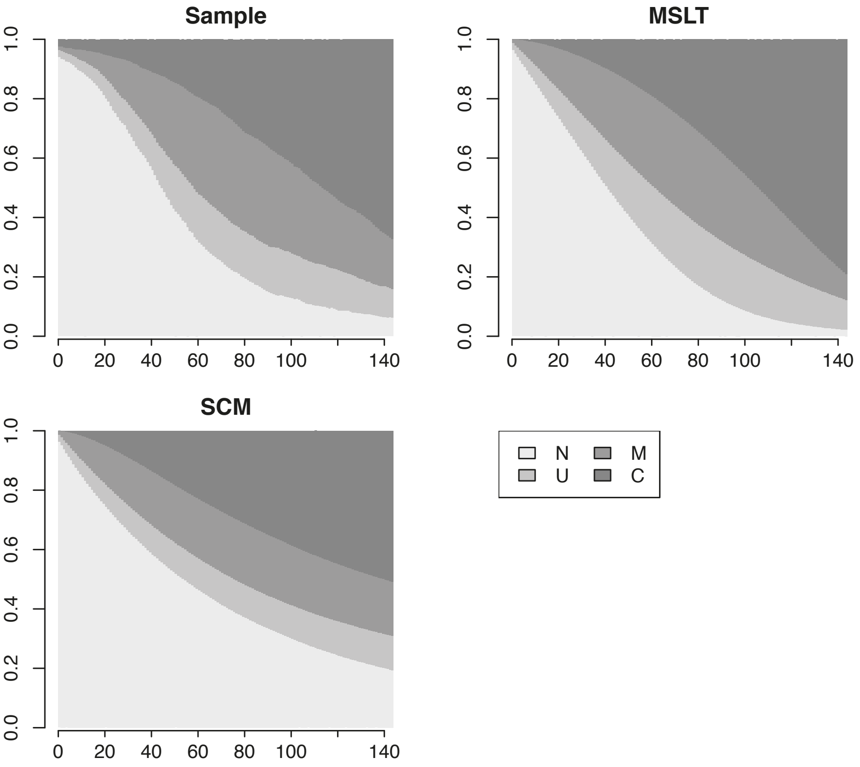


Figure 2. Transversal state distributions in the FFS data and in the simulated data

what follows, we focus on the sequences with combinations of covariates with a relatively high observed frequency. These are defined in Table 2, that reports the 5 most frequent combinations of covariates values in the FSS sample, characterizing a total of 1175 individuals (63.4% of the sample size).

Table 2. *Most frequent combinations of covariates values*

$\mathbf{x}$	Cohort	Education	Religion	Divorce	Initial state	Frequency
$\mathbf{x}_1$	53-57	0-3 Yrs	Yes	No	N	227
$\mathbf{x}_2$	53-57	>3 Yrs	Yes	No	N	247
$\mathbf{x}_3$	58-62	0-3 Yrs	Yes	No	N	195
$\mathbf{x}_4$	58-62	>3 Yrs	No	No	N	177
$\mathbf{x}_5$	58-62	>3 Yrs	Yes	No	N	329

In order to apply the procedures described in Section 4.3, we need to choose a pairwise dissimilarity measure. Here we use Optimal Matching (OM), but other alternatives may be adopted. In particular, by following a standard approach in the literature, the insertion and the deletion costs were both set equal to 1, while the substitution cost between two states was chosen to be inversely related to the transition frequency (Rohwer and Pötter, 2004).

Table 3. *Summaries of dissimilarities of the sequences from the medoid: averages, coefficients of variation, minimum and maximum* $\overline{\mathbf{S}}$

	Summary	Combination of covariates				
	Statistic	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$d(\mathbf{S}(\mathbf{x}), \bar{\mathbf{S}}(\mathbf{x}))$	Mean	90.5	101.0	95.1	104.9	101.1
	CV	0.60	0.49	0.50	0.41	0.42
	Min	3.9	6.0	10.0	11.9	23.8
	Max	259.5	270.4	219.6	209.4	212.7
$d(\widehat{\mathbf{S}}_{SCM}(\mathbf{x}), \bar{\mathbf{S}}(\mathbf{x}))$	Mean	117.2	121.7	124.1	128.2	135.0
	CV	0.48	0.33	0.41	0.39	0.31
	Min	7.9	9.4	2.0	8.0	13.9
	Max	200.3	230.7	280.5	255.6	229.1
$d(\widehat{\mathbf{S}}_{MSLT}(\mathbf{x}), \bar{\mathbf{S}}(\mathbf{x}))$	Mean	87.8	101.9	94.0	117.7	109.7
	CV	0.48	0.45	0.48	0.43	0.43
	Min	7.9	5.9	7.9	13.9	15.9
	Max	193.7	237.3	252.7	247.6	254.4

Notes: See Table 2 for the definition of the combination of covariates.

For each combination of covariate values in Table 2, we consider the observed sequences, $\mathbf{S}(\mathbf{x})$, and the simulated ones, $\widehat{\mathbf{S}}_{SCM}(\mathbf{x})$ and $\widehat{\mathbf{S}}_{MSLT}(\mathbf{x})$. Sequences in $\mathbf{S}(\mathbf{x})$ are summarized using their medoid $\overline{\mathbf{S}}(\mathbf{x})$, i.e. the sequence minimizing the sum of its dissimilarities with all other sequences in the group. The first approach that we proposed with goal of assessing the predictive ability of a model is the computation of the dissimilarities between the sequences in $\widehat{\mathbf{S}}(\mathbf{x})$ and $\overline{\mathbf{S}}(\mathbf{x})$, $d(\widehat{\mathbf{S}}(\mathbf{x}), \overline{\mathbf{S}}(\mathbf{x}))$. The distribution of such dissimilarities can be analyzed graphically or by using summary statistics.

Table 3 reports selected summaries (mean, coefficient of variation, minimum and maximum) of the distributions obtained for the most frequent covariate combinations. For the sake of comparison, in the table we also consider the summary of the dissimilarities between sequences in $\mathbf{S}(\mathbf{x})$ and their medoid. Table results in the table suggest that the MSLT model leads to simulated sequences which, compared to those generated using the SCM, tend to be closer to the medoids-sequences, and also have dissimilarities with smaller coefficients of variation (i.e. characterized by less variability).

The descriptive analysis presented so far can be helpful to assess the predictive performance of alternative models. Even so, one might be interested to test such differences in a more formal way. Moreover, instead of focusing on the dissimilarities between the simulated sequences and a summary of the observed ones, it could be sensible to refer to criteria taking into account *all* the interpoint distances. This is what we do now, as we examine the results of applying the tests based on the two statistics in equations (4.6) and (4.7) introduced in the previous section.

Table 4. *Tests on the differences between observed and model-based dissimilarities*

Panel A: M-statistic test results (two-sided p-values)

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\mathbf{S}(\mathbf{x})$ vs. $\mathbf{S}_{SCM}(\mathbf{x})$	0.0005	0.0049	0.0006	0.0114	<0.00001
$\mathbf{S}(\mathbf{x})$ vs. $\mathbf{S}_{MSLT}(\mathbf{x})$	0.0008	0.00001	0.2969	0.1874	0.0004
$\mathbf{S}_{SCM}(\mathbf{x})$ vs. $\mathbf{S}_{MSLT}(\mathbf{x})$	<0.00001	<0.00001	<0.00001	0.0270	<0.00001

Panel B: Mosler-Wilcoxon test results (two-sided p-values)

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\mathbf{S}(\mathbf{x})$ vs. $\mathbf{S}_{SCM}(\mathbf{x})$	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
$\mathbf{S}(\mathbf{x})$ vs. $\mathbf{S}_{MSLT}(\mathbf{x})$	0.164	0.002	0.136	0.004	<0.0001
$\mathbf{S}_{SCM}(\mathbf{x})$ vs. $\mathbf{S}_{MSLT}(\mathbf{x})$	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

Notes: See Section 4.3 for the definition of the two dissimilarity-based tests.

We first compute the $\tilde{M}$ statistic to compare dissimilarities across groups of trajectories, separately for each of the most frequent covariate values. The interpoint distance distributions (IDDs) in each set were estimated based on $K = 20$ bins. The p-values characterizing the test statistics, based on 1000 permuted samples, are reported in Panel A of Table 4. The results in the table suggest once again that the trajectories generated using the MSLT model tend to be more similar to the trajectories in the data as opposed to the trajectories generated based on the SCM model. Consistently, the MSLT and the SCM trajectories appear to be different from each other when one looking at their interpoint distances.

Similar conclusions can be drawn also when using the Mosler-Wilcoxon test; results are reported in Panel B of Table 4. The table shows that the $\tilde{M}$ and the Mosler-Wilcoxon tests produce somewhat complementary results, in that they return (non) rejections and evidence against different null hypotheses for different sets of covariate values.

A general look at Table 4 suggests that, overall, the MSLT model tends to perform better than SCM one, at least for the FFS dataset and the dissimilarity measures and covariates levels considered. One possible reason behind this is that the MSLT model uses parameters specific for the different arrival states, whereas the SCM model uses parameters specific both for the arrival and for the departure states. Moreover, the two models treat durations differently.[4] An in-depth assessment of the reasons behind the different predictive performance of the two models is outside the scope of this paper. However, the results presented here suggest that our methods can help in identifying ways in which to modify the models so as to improve their predictive performance.

4.5 Conclusions

In this project we proposed three criteria for comparing alternative parametric multi-state models from the point of view of their ability to reproduce a sample of observed trajectories. In particular, we compared model-generated trajectories to the trajectories that were used to estimate the models parameters.

In general, the methods proposed here may be seen either as part of a strategy for model selection or as pure prediction comparison tools. In our analyses we focused on the former aspect,[5] and the *p*-values of the tests considered can be interpreted as a metric for goodness of fit of competing models. Moreover, they can provide guidance on the selection among alternative non-nested modeling approaches. Our illustration here was based on this in-sample prediction approach, and the proposed distance-based methods appear to be a promising tool to detect differences between the two models considered.

[4] In the MSLT, the durations affect the transitions probabilities at each time point. Instead, the SCM assumes a simple geometric regression model for the time until the next transition, and includes the duration in the part of the model that refers to the conditional transition probabilities at the time when a transition occurs.

[5] In the second case, the competing models may be used to produce trajectories that are compared to observed trajectories that were not used to estimate the models parameters.

Note that in our analyses we chose to keep the assessments separate rather than pooled across different combinations of covariates. We also decided to focus on covariates combinations with relatively high frequencies.[6] These and other aspects of our analyses can be easily extended.[7]

Finally, note that, despite the proposed criteria provide insights about the existence of discrepancies between observed and simulated sequences, clearly they do not directly suggest an "interpretation" of the reasons behind such discrepancies. This is because the criteria proposed here do not describe in what respects the model-generated sequences differ from the observed ones. On the other hand, the measures of dissimilarity used in SA typically penalize some differences more than others, so that hopefully the most substantial and problematic deviations are emphasized by a relatively a large dissimilarity. Whether this can be meaningfully interpreted, of course, depends on the definition of dissimilarity that is used in the specific application.

Relatedly, all comparison measures proposed here directly depend upon the chosen measure of dissimilarity. This can be seen as a limitation, and the robustness of results should be assessed by performing sensitivity analyses. On the other hand, the existence of different results when using alternative dissimilarity measures may actually shed light on relevant features of the sequences. Moreover, the measure of dissimilarity might be chosen to capture characteristics of the sequences that are deemed to be relevant according to theory or prior knowledge.

[6] The comparison of the predictive power of models for covariate values with low frequency requires particular care. For such cases, it might be preferable to generate a number of sequences larger than the number of observed sequences.

[7] As an additional example, the *p*-values of the comparison criteria will in general depend on the number of simulated sequences. As this number increases, one may expect the tests to be more likely to reject. Hence, the number of simulated sequences could also be used as a tuning parameter to let vary when implementing the comparison tests proposed here.

Appendix

Maximum likelihood estimates of the MSLT and SCM models as obtained in Lombardi (2012) and in Bonetti et al. (2013).

Table A.1. *MSLT estimates and p-values (Lombardi, 2012)*

Transition to N			
Parameter	Estimate	SE	*p*-value
Intercept	7.847	0.2126	<0.0001
Previous state = M	−11.560	0.2672	< 0.0001
Previous state = U	−6.735	0.2501	<0.0001
Tval	−0.015	0.0014	<0.0001
Age	0.019	0.0053	0.043
*Age*Age*	−0.0002	0.00005	0.0004
Educ2	0.261	0.1361	0.055
Educ3	1.086	0.1359	<0.0001
Religion	−0.192	0.0833	0.021
Divorce	−0.247	0.1559	0.11
Cohort	0.139	0.0786	0.077
Transition to M			
Parameter	Estimate	SE	*p*-value
Intercept	3.072	0.1929	<0.0001
Previous state = M	0.637	0.1837	0.0005
Previous state = U	0.847	0.2335	0.0003
Tval	−0.006	0.0011	<0.0001
Age	0.012	0.0034	0.0004
*Age*Age*	−0.0001	0.00003	<0.0001
Educ2	0.248	0.0985	0.012
Educ3	0.416	0.1003	<0.0001
Religion	0.043	0.0619	0.48
Divorce	−0.310	0.1141	0.007
Cohort	0.042	0.0591	0.47
Transition to U			
Parameter	Estimate	SE	*p*-value
Intercept	2.496	0.2252	<0.0001
Previous state = M	−8.554	0.4064	<0.0001
Previous state = U	2.894	0.2386	<0.0001
Tval	−0.003	0.0015	0.071
Age	0.023	0.0047	<0.0001
*Age*Age*	−0.0002	0.00004	<0.0001
Educ2	0.539	0.1510	0.0004
Educ3	1.202	0.1493	<0.0001
Religion	−0.507	0.0855	<0.0001
Divorce	−0.036	0.1548	0.82
Cohort	0.354	0.0828	<0.0001

Table A.2. *Parameter estimates and p-values for the duration component of the SCM model. Replication of Bonetti et al. (2013).*

Parameter	Estimate	SE	p-value
Intercept	−4.329	0.0301	<0.00001
Age	−0.012	0.0006	<0.00001
Previous state = M	0.900	0.0410	<0.00001
Previous state = U	1.077	0.0458	<0.00001
Cohort	0.079	0.0300	0.009

Table A.3. *Parameter estimates and p-values for the transition component of the SCM model. Replication of Bonetti et al. (2013).*

	Transition from N to M			Transition from N to U		
Parameter	Est.	SE	p-value	Est.	SE	p-value
Intercept	2.799	0.5736	<0.0001	−4.071	0.8806	<0.0001
Tval	−0.005	0.0064	0.22	0.013	0.0064	0.02
Age	−0.054	0.0080	<0.0001	−0.270	0.0086	0.0009
Educ2	0.319	0.5456	0.28	0.394	0.8540	0.32
Educ3	0.110	0.5274	0.42	1.546	0.8380	0.03
Religion	1.869	0.3674	<0.0001	−0.253	0.3769	0.25
Divorce	−2.259	0.4213	<0.0001	0.890	0.5493	0.05
Cohort	−0.145	0.3413	0.34	0.390	0.3711	0.15
	Transition from M to N			Transition from M to U		
Parameter	Est.	SE	p-value	Est.	SE	p-value
Intercept	2.553	0.7213	0.0002	1.652	0.5760	0.002
Tval	−0.013	0.0063	0.02	0.005	0.0063	0.22
Age	−0.015	0.0057	0.005	−0.010	0.0050	0.02
Educ2	−0.107	0.6485	0.43	0.647	0.5446	0.12
Educ3	0.695	0.6343	0.14	1.201	0.5248	0.01
Religion	−0.161	0.3416	0.32	0.672	0.3625	0.03
Divorce	−0.515	0.4290	0.11	−1.083	0.3762	0.002
Cohort	−0.384	0.3557	0.14	0.782	0.3356	0.01
	Transition from U to N			Transition from U to M		
Parameter	Est.	SE	p-value	Est.	SE	p-value
Intercept	−4.657	1.3044	0.0002	2.932	0.6849	<0.0001
Tval	0.007	0.0136	0.31	−0.018	0.0059	0.001
Age	−0.0003	0.0130	0.49	−0.008	0.0052	0.06
Educ2	0.567	1.1027	0.30	0.398	0.6048	0.25
Educ3	−1.088	1.3913	0.22	0.819	0.5963	0.08
Religion	−0.400	0.7203	0.29	0.101	0.3168	0.38
Divorce	−1.086	2.3363	0.32	−1.076	0.4019	0.004
Cohort	−0.473	0.7413	0.26	0.062	0.3342	0.43

References

Aassve, A., Billari, F. C., and Piccarreta, R. (2007). Strings of adulthood: A sequence analysis of young British women's work-family trajectories. *European Journal of Population*, 23:369–388.

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology*, 21:93–113.

Agresti, A. (2002). *Categorical data analysis*. John Wiley & Sons. Wiley, New York.

Aisenbrey, S. and Fasang, A. E. (2010). New life for old ideas: The "second wave" of sequence analysis bringing the "course" back into the life course. *Sociological Methods and Research*, 38:420–462.

Beyersmann, J., Allignol, A., and Schumacher, M. (2012). *Competing risks and multistate models with R*. Springer. Springer, New York, NY.

Bonetti, M. and Pagano, M. (2005). The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering. *Statistics in Medicine*, 24(5):753–773.

Bonetti, M., Piccarreta, R., and Salford, G. (2013). Parametric and nonparametric analysis of life courses: An application to family formation patterns. *Demography*, 50(3):881–902.

Cai, L., Hayward, M., Saito, Y., Lubitz, J., Hagedorn, A., and Crimmins, E. (2010). Estimation of multi-state life table functions and their variability from complex survey data using the SPACE program. *Demographic Research*, 22:129–158.

Cai, L., Schenker, N., and Lubitz, J. (2006). Analysis of functional status transitions by using a semi-Markov process model in the presence of left-censored spells. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(4):477–491.

Laditka, S. B. and Wolf, D. A. (1998). New methods for analyzing active life expectancy. *Journal of Aging and Health*, 10(2):214–241.

Latten, J. and De Graaf, A. (1997). *Fertility and family surveys in countries of the ECE region*, volume The Netherlands of *Standard Country Report*. New York, NY.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. John Wiley & Sons. New York, NY.

Lombardi, S. (2012). Multistate models for event-history data: Methods and applications to women's childbearing and family formation patterns. MSc thesis in Economics, Bocconi University.

Manjourides, J. D. (2009). Distance based methods for space and time modelling of the health of populations. PhD dissertation, Harvard School of Public Health, Department of Biostatistics.

McVicar, D. and Anyadike-Anes, M. (2002). Predicting successful and unsuccessful transitions from school to work using sequence methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 165:317–334.

Mosler, K. (2002). *Multivariate dispersion, central regions, and depth: The lift zonoid approach*. Springer. New York, NY.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430.

Rohwer, G. and Pötter, U. (2004). *TDA user's manual*. Ruhr-Universität Bochum. Bochum, Germany.

Sankoff, D. and Kruskal, J. B. (1983). *Time warps, string edits and macromolecules*. Addison-Wesley. Reading, Massachusetts.

Sheikh, Y. A., Khan, E. A., and Kanade, T. (2007). Mode-seeking by medoid shifts. Washington, DC. IEEE Computer Society.

Tebaldi, P., Bonetti, M., and Pagano, M. (2011). M statistic commands: Interpoint distance distribution analysis. *The Stata Journal*, 11(2):271–289.

Economic Studies

1987:1 Haraldson, Marty. To Care and To Cure. A linear programming approach to national health planning in developing countries. 98 pp.

1989:1 Chryssanthou, Nikos. The Portfolio Demand for the ECU. A Transaction Cost Approach. 42 pp.

1989:2 Hansson, Bengt. Construction of Swedish Capital Stocks, 1963-87. An Application of the Hulten-Wykoff Studies. 37 pp.

1989:3 Choe, Byung-Tae. Some Notes on Utility Functions Demand and Aggregation. 39 pp.

1989:4 Skedinger, Per. Studies of Wage and Employment Determination in the Swedish Wood Industry. 89 pp.

1990:1 Gustafson, Claes-Håkan. Inventory Investment in Manufacturing Firms. Theory and Evidence. 98 pp.

1990:2 Bantekas, Apostolos. The Demand for Male and Female Workers in Swedish Manufacturing. 56 pp.

1991:1 Lundholm, Michael. Compulsory Social Insurance. A Critical Review. 109 pp.

1992:1 Sundberg, Gun. The Demand for Health and Medical Care in Sweden. 58 pp.

1992:2 Gustavsson, Thomas. No Arbitrage Pricing and the Term Structure of Interest Rates. 47 pp.

1992:3 Elvander, Nils. Labour Market Relations in Sweden and Great Britain. A Comparative Study of Local Wage Formation in the Private Sector during the 1980s. 43 pp.

12 Dillén, Mats. Studies in Optimal Taxation, Stabilization, and Imperfect Competition. 1993. 143 pp.

13 Banks, Ferdinand E.. A Modern Introduction to International Money, Banking and Finance. 1993. 303 pp.

14 Mellander, Erik. Measuring Productivity and Inefficiency Without Quantitative Output Data. 1993. 140 pp.

15 Ackum Agell. Susanne. Essays on Work and Pay. 1993. 116 pp.

16 Eriksson, Claes. Essays on Growth and Distribution. 1994. 129 pp.

17 Banks, Ferdinand E.. A Modern Introduction to International Money, Banking and Finance. 2nd version, 1994. 313 pp.

18 Apel, Mikael. Essays on Taxation and Economic Behavior. 1994. 144 pp.

19 Dillén, Hans. Asset Prices in Open Monetary Economies. A Contingent Claims Approach. 1994. 100 pp.

20 Jansson, Per. Essays on Empirical Macroeconomics. 1994. 146 pp.

21 Banks, Ferdinand E.. A Modern Introduction to International Money, Banking, and Finance. 3rd version, 1995. 313 pp.

22 Dufwenberg, Martin. On Rationality and Belief Formation in Games. 1995. 93 pp.

23 Lindén, Johan. Job Search and Wage Bargaining. 1995. 127 pp.

24 Shahnazarian, Hovick. Three Essays on Corporate Taxation. 1996. 112 pp.

25 Svensson, Roger. Foreign Activities of Swedish Multinational Corporations. 1996. 166 pp.

26 Sundberg, Gun. Essays on Health Economics. 1996. 174 pp.

27 Sacklén, Hans. Essays on Empirical Models of Labor Supply. 1996. 168 pp.

28 Fredriksson, Peter. Education, Migration and Active Labor Market Policy. 1997. 106 pp.

29 Ekman, Erik. Household and Corporate Behaviour under Uncertainty. 1997. 160 pp.

30 Stoltz, Bo. Essays on Portfolio Behavior and Asset Pricing. 1997. 122 pp.

31 Dahlberg, Matz. Essays on Estimation Methods and Local Public Economics. 1997. 179 pp.

32 Kolm, Ann-Sofie. Taxation, Wage Formation, Unemployment and Welfare. 1997. 162 pp.

33 Boije, Robert. Capitalisation, Efficiency and the Demand for Local Public Services. 1997. 148 pp.

34 Hort, Katinka. On Price Formation and Quantity Adjustment in Swedish Housing Markets. 1997. 185 pp.

35 Lindström, Thomas. Studies in Empirical Macroeconomics. 1998. 113 pp.

36 Hemström, Maria. Salary Determination in Professional Labour Markets. 1998. 127 pp.

37 Forsling, Gunnar. Utilization of Tax Allowances and Corporate Borrowing. 1998. 96 pp.

38 Nydahl, Stefan. Essays on Stock Prices and Exchange Rates. 1998. 133 pp.

39 Bergström, Pål. Essays on Labour Economics and Econometrics. 1998. 163 pp.

40 Heiborn, Marie. Essays on Demographic Factors and Housing Markets. 1998. 138 pp.

41 Åsberg, Per. Four Essays in Housing Economics. 1998. 166 pp.

42 Hokkanen, Jyry. Interpreting Budget Deficits and Productivity Fluctuations. 1998. 146 pp.

43 Lunander, Anders. Bids and Values. 1999. 127 pp.

44 Eklöf, Matias. Studies in Empirical Microeconomics. 1999. 213 pp.

45 Johansson, Eva. Essays on Local Public Finance and Intergovernmental Grants. 1999. 156 pp.

46 Lundin, Douglas. Studies in Empirical Public Economics. 1999. 97 pp.

47 Hansen, Sten. Essays on Finance, Taxation and Corporate Investment. 1999. 140 pp.

48 Widmalm, Frida. Studies in Growth and Household Allocation. 2000. 100 pp.

49 Arslanogullari, Sebastian. Household Adjustment to Unemployment. 2000. 153 pp.

50 Lindberg, Sara. Studies in Credit Constraints and Economic Behavior. 2000. 135 pp.

51 Nordblom, Katarina. Essays on Fiscal Policy, Growth, and the Importance of Family Altruism. 2000. 105 pp.

52 Andersson, Björn. Growth, Saving, and Demography. 2000. 99 pp.

53 Åslund, Olof. Health, Immigration, and Settlement Policies. 2000. 224 pp.

54 Bali Swain, Ranjula. Demand, Segmentation and Rationing in the Rural Credit Markets of Puri. 2001. 160 pp.

55 Löfqvist, Richard. Tax Avoidance, Dividend Signaling and Shareholder Taxation in an Open Economy. 2001. 145 pp.

56 Vejsiu, Altin. Essays on Labor Market Dynamics. 2001. 209 pp.

57 Zetterström, Erik. Residential Mobility and Tenure Choice in the Swedish Housing Market. 2001. 125 pp.

58 Grahn, Sofia. Topics in Cooperative Game Theory. 2001. 106 pp.

59 Laséen, Stefan. Macroeconomic Fluctuations and Microeconomic Adjustments. Wages, Capital, and Labor Market Policy. 2001. 142 pp.

60 Arnek, Magnus. Empirical Essays on Procurement and Regulation. 2002. 155 pp.

61 Jordahl, Henrik. Essays on Voting Behavior, Labor Market Policy, and Taxation. 2002. 172 pp.

62 Lindhe, Tobias. Corporate Tax Integration and the Cost of Capital. 2002. 102 pp.

63 Hallberg, Daniel. Essays on Household Behavior and Time-Use. 2002. 170 pp.

64 Larsson, Laura. Evaluating Social Programs: Active Labor Market Policies and Social Insurance. 2002. 126 pp.

65 Bergvall, Anders. Essays on Exchange Rates and Macroeconomic Stability. 2002. 122 pp.

66 Nordström Skans, Oskar. Labour Market Effects of Working Time Reductions and Demographic Changes. 2002. 118 pp.

67 Jansson, Joakim. Empirical Studies in Corporate Finance, Taxation and Investment. 2002. 132 pp.

68 Carlsson, Mikael. Macroeconomic Fluctuations and Firm Dynamics: Technology, Production and Capital Formation. 2002. 149 pp.

69 Eriksson, Stefan. The Persistence of Unemployment: Does Competition between Employed and Unemployed Job Applicants Matter? 2002. 154 pp.

70 Huitfeldt, Henrik. Labour Market Behaviour in a Transition Economy: The Czech Experience. 2003. 110 pp.

71 Johnsson, Richard. Transport Tax Policy Simulations and Satellite Accounting within a CGE Framework. 2003. 84 pp.

72 Öberg, Ann. Essays on Capital Income Taxation in the Corporate and Housing Sectors. 2003. 183 pp.

73 Andersson, Fredrik. Causes and Labor Market Consequences of Producer Heterogeneity. 2003. 197 pp.

74 Engström, Per. Optimal Taxation in Search Equilibrium. 2003. 127 pp.

75 Lundin, Magnus. The Dynamic Behavior of Prices and Investment: Financial Constraints and Customer Markets. 2003. 125 pp.

76 Ekström, Erika. Essays on Inequality and Education. 2003. 166 pp.

77 Barot, Bharat. Empirical Studies in Consumption, House Prices and the Accuracy of European Growth and Inflation Forecasts. 2003. 137 pp.

78 Österholm, Pär. Time Series and Macroeconomics: Studies in Demography and Monetary Policy. 2004. 116 pp.

79 Bruér, Mattias. Empirical Studies in Demography and Macroeconomics. 2004. 113 pp.

80 Gustavsson, Magnus. Empirical Essays on Earnings Inequality. 2004. 154 pp.

81 Toll, Stefan. Studies in Mortgage Pricing and Finance Theory. 2004. 100 pp.

82 Hesselius, Patrik. Sickness Absence and Labour Market Outcomes. 2004. 109 pp.

83 Häkkinen, Iida. Essays on School Resources, Academic Achievement and Student Employment. 2004. 123 pp.

84 Armelius, Hanna. Distributional Side Effects of Tax Policies: An Analysis of Tax Avoidance and Congestion Tolls. 2004. 96 pp.

85 Ahlin, Åsa. Compulsory Schooling in a Decentralized Setting: Studies of the Swedish Case. 2004. 148 pp.

86 Heldt, Tobias. Sustainable Nature Tourism and the Nature of Tourists' Cooperative Behavior: Recreation Conflicts, Conditional Cooperation and the Public Good Problem. 2005. 148 pp.

87 Holmberg, Pär. Modelling Bidding Behaviour in Electricity Auctions: Supply Function Equilibria with Uncertain Demand and Capacity Constraints. 2005. 43 pp.

88 Welz, Peter. Quantitative new Keynesian macroeconomics and monetary policy 2005. 128 pp.

89 Ågren, Hanna. Essays on Political Representation, Electoral Accountability and Strategic Interactions. 2005. 147 pp.

90 Budh, Erika. Essays on environmental economics. 2005. 115 pp.

91 Chen, Jie. Empirical Essays on Housing Allowances, Housing Wealth and Aggregate Consumption. 2005. 192 pp.

92 Angelov, Nikolay. Essays on Unit-Root Testing and on Discrete-Response Modelling of Firm Mergers. 2006. 127 pp.

93 Savvidou, Eleni. Technology, Human Capital and Labor Demand. 2006. 151 pp.

94 Lindvall, Lars. Public Expenditures and Youth Crime. 2006. 112 pp.

95 Söderström, Martin. Evaluating Institutional Changes in Education and Wage Policy. 2006. 131 pp.

96 Lagerström, Jonas. Discrimination, Sickness Absence, and Labor Market Policy. 2006. 105 pp.

97 Johansson, Kerstin. Empirical essays on labor-force participation, matching, and trade. 2006. 168 pp.

98 Ågren, Martin. Essays on Prospect Theory and the Statistical Modeling of Financial Returns. 2006. 105 pp.

99 Nahum, Ruth-Aïda. Studies on the Determinants and Effects of Health, Inequality and Labour Supply: Micro and Macro Evidence. 2006. 153 pp.

100 Žamac, Jovan. Education, Pensions, and Demography. 2007. 105 pp.

101 Post, Erik. Macroeconomic Uncertainty and Exchange Rate Policy. 2007. 129 pp.

102 Nordberg, Mikael. Allies Yet Rivals: Input Joint Ventures and Their Competitive Effects. 2007. 122 pp.

103 Johansson, Fredrik. Essays on Measurement Error and Nonresponse. 2007. 130 pp.

104 Haraldsson, Mattias. Essays on Transport Economics. 2007. 104 pp.

105 Edmark, Karin. Strategic Interactions among Swedish Local Governments. 2007. 141 pp.

106 Oreland, Carl. Family Control in Swedish Public Companies. Implications for Firm Performance, Dividends and CEO Cash Compensation. 2007. 121 pp.

107 Andersson, Christian. Teachers and Student Outcomes: Evidence using Swedish Data. 2007. 154 pp.

108 Kjellberg, David. Expectations, Uncertainty, and Monetary Policy. 2007. 132 pp.

109 Nykvist, Jenny. Self-employment Entry and Survival - Evidence from Sweden. 2008. 94 pp.

110 Selin, Håkan. Four Empirical Essays on Responses to Income Taxation. 2008. 133 pp.

111 Lindahl, Erica. Empirical studies of public policies within the primary school and the sickness insurance. 2008. 143 pp.

112 Liang, Che-Yuan. Essays in Political Economics and Public Finance. 2008. 125 pp.

113 Elinder, Mikael. Essays on Economic Voting, Cognitive Dissonance, and Trust. 2008. 120 pp.

114 Grönqvist, Hans. Essays in Labor and Demographic Economics. 2009. 120 pp.

115 Bengtsson, Niklas. Essays in Development and Labor Economics. 2009. 93 pp.

116 Vikström, Johan. Incentives and Norms in Social Insurance: Applications, Identification and Inference. 2009. 205 pp.

117 Liu, Qian. Essays on Labor Economics: Education, Employment, and Gender. 2009. 133 pp.

118 Glans, Erik. Pension reforms and retirement behaviour. 2009. 126 pp.

119 Douhan, Robin. Development, Education and Entrepreneurship. 2009.

120 Nilsson, Peter. Essays on Social Interactions and the Long-term Effects of Early-life Conditions. 2009. 180 pp.

121 Johansson, Elly-Ann. Essays on schooling, gender, and parental leave. 2010. 131 pp.

122 Hall, Caroline. Empirical Essays on Education and Social Insurance Policies. 2010. 147 pp.

123 Enström-Öst, Cecilia. Housing policy and family formation. 2010. 98 pp.

124 Winstrand, Jakob. Essays on Valuation of Environmental Attributes. 2010. 96 pp.

125 Söderberg, Johan. Price Setting, Inflation Dynamics, and Monetary Policy. 2010. 102 pp.

126 Rickne, Johanna. Essays in Development, Institutions and Gender. 2011. 138 pp.

127 Hensvik, Lena. The effects of markets, managers and peers on worker outcomes. 2011. 179 pp.

128 Lundqvist, Heléne. Empirical Essays in Political and Public. 2011. 157 pp.

129 Bastani, Spencer. Essays on the Economics of Income Taxation. 2012. 257 pp.

130 Corbo, Vesna. Monetary Policy, Trade Dynamics, and Labor Markets in Open Economies. 2012. 262 pp.

131 Nordin, Mattias. Information, Voting Behavior and Electoral Accountability. 2012. 187 pp.

132 Vikman, Ulrika. Benefits or Work? Social Programs and Labor Supply. 2013. 161 pp.

133 Ek, Susanne. Essays on unemployment insurance design. 2013. 136 pp.

134 Österholm, Göran. Essays on Managerial Compensation. 2013. 143 pp.

135 Adermon, Adrian. Essays on the transmission of human capital and the impact of technological change. 2013. 138 pp.

136 Kolsrud, Jonas. Insuring Against Unemployment 2013. 140 pp.

137 Hanspers, Kajsa. Essays on Welfare Dependency and the Privatization of Welfare Services. 2013. 208 pp.

138 Persson, Anna. Activation Programs, Benefit Take-Up, and Labor Market Attachment. 2013. 164 pp.

139 Engdahl, Mattias. International Mobility and the Labor Market. 2013. 216 pp.

140 Krzysztof Karbownik. Essays in education and family economics. 2013. 182 pp.

141 Oscar Erixson. Economic Decisions and Social Norms in Life and Death Situations. 2013. 183 pp.

142 Pia Fromlet. Essays on Inflation Targeting and Export Price Dynamics. 2013. 145 pp.

143 Daniel Avdic. Microeconometric Analyses of Individual Behavior in Public Welfare Systems. Applications in Health and Education Economics. 2014. 176 pp.

144 Arizo Karimi. Impacts of Policies, Peers and Parenthood on Labor Market Outcomes. 2014. 221 pp.

145 Karolina Stadin. Employment Dynamics. 2014. 134 pp.

146 Haishan Yu. Essays on Environmental and Energy Economics. 132 pp.

147 Martin Nilsson. Essays on Health Shocks and Social Insurance. 139 pp.

148 Tove Eliasson. Empirical Essays on Wage Setting and Immigrant Labor Market Opportunities. 2014. 144 pp.

149 Erik Spector. Financial Frictions and Firm Dynamics. 2014. 129 pp.

150 Michihito Ando. Essays on the Evaluation of Public Policies. 2015. 193 pp.

151 Selva Bahar Baziki. Firms, International Competition, and the Labor Market. 2015. 183 pp.

152 Fredrik Sävje. What would have happened? Four essays investigating causality. 2015. 229 pp.

153 Ina Blind. Essays on Urban Economics. 2015. 197 pp.

154 Jonas Poulsen. Essays on Development and Politics in Sub-Saharan Africa. 2015. 240 pp.

155 Lovisa Persson. Essays on Politics, Fiscal Institutions, and Public Finance. 2015. 137 pp.

156 Gabriella Chirico Willstedt. Demand, Competition and Redistribution in Swedish Dental Care. 2015. 119 pp.

157 Yuwei Zhao de Gosson de Varennes. Benefit Design, Retirement Decisions and Welfare Within and Across Generations in Defined Contribution Pension Schemes. 2016. 148 pp.

158 Johannes Hagen. Essays on Pensions, Retirement and Tax Evasion. 2016. 195 pp.

159 Rachatar Nilavongse. Housing, Banking and the Macro Economy. 2016. 156 pp.

160 Linna Martén. Essays on Politics, Law, and Economics. 2016. 150 pp.

161 Olof Rosenqvist. Essays on Determinants of Individual Performance and Labor Market Outcomes. 2016. 151 pp.

162 Linuz Aggeborn. Essays on Politics and Health Economics. 2016. 203 pp.

163 Glenn Mickelsson. DSGE Model Estimation and Labor Market Dynamics. 2016. 166 pp.

164 Sebastian Axbard. Crime, Corruption and Development. 2016. 150 pp.

165 Mattias Öhman. Essays on Cognitive Development and Medical Care. 2016. 181 pp.

166 Jon Frank. Essays on Corporate Finance and Asset Pricing. 2017. 160 pp.

167 Ylva Moberg. Gender, Incentives, and the Division of Labor. 2017. 220 pp.

168 Sebastian Escobar. Essays on inheritance, small businesses and energy consumption. 2017. 194 pp.

169 Evelina Björkegren. Family, Neighborhoods, and Health. 2017. 226 pp.

170 Jenny Jans. Causes and Consequences of Early-life Conditions. Alcohol, Pollution and Parental Leave Policies. 2017. 209 pp.

171 Josefine Andersson. Insurances against job loss and disability. Private and public interventions and their effects on job search and labor supply. 2017. 175 pp.

172 Jacob Lundberg. Essays on Income Taxation and Wealth Inequality. 2017. 173 pp.

173 Anna Norén. Caring, Sharing, and Childbearing. Essays on Labor Supply, Infant Health, and Family Policies. 2017. 206 pp.

174 Irina Andone. Exchange Rates, Exports, Inflation, and International Monetary Cooperation. 2018. 174 pp.

175 Henrik Andersson. Immigration and the Neighborhood. Essays on the Causes and Consequences of International Migration. 2018. 181 pp.

176 Aino-Maija Aalto. Incentives and Inequalities in Family and Working Life. 2018. 131 pp.

177 Gunnar Brandén. Understanding Intergenerational Mobility. Inequality, Student Aid and Nature-Nurture Interactions. 2018. 125 pp.

178 Mohammad H. Sepahvand. Essays on Risk Attitudes in Sub-Saharan Africa. 2019. 215 pp.

179 Mathias von Buxhoeveden. Partial and General Equilibrium Effects of Unemployment Insurance. Identification, Estimation and Inference. 2019. 89 pp.

180 Stefano Lombardi. Essays on Event History Analysis and the Effects of Social Programs on Individuals and Firms. 2019. 150 pp.