The Danish Labor Market Experiments: Methods and Findings

Af
Jonas Maibom*
Aarhus University & DTMC

Keywords: Active labor market policy; Randomized Social Experiment, Economic Models
JEL codes: J64, J68, C9, C5

Abstract

In this paper I discuss recent contributions to a literature which combines the use of structural economic models and (quasi-)experimental variation in the data to evaluate labor market policies. I refer to this approach as Structural and Empirical Policy Evaluation (SPE) and relate it to methodological discussions in the literature on policy evaluation. I then introduce a series of recent Danish labor market experiments and applications of SPE to discuss the potential value added of this approach.

* I am grateful for helpful comments from Michael Svarer and Jonas Bødker. The research in relation to Maibom (2021) is supported by the Carlsberg Foundation (CF15-0647).

1. Introduction

The aim of this paper is to introduce a recent research agenda, Structural and Empirical Policy Evaluation (SPE), and illustrate the idea and value added of this approach by focusing on the Danish labor market experiments and related research. The idea of the SPE approach is to combine the use of economic models and (quasi-)experimental variation in the data. This framework can be seen as a synthesis of two different strands of literature: the causal effects framework and the building and estimation of economic models (see Heckman (2010)). By combining the state of the art from both literatures, the SPE approach aims to extend and improve the analysis and evalulation of policies. This combination is particularly powerful because each approach, to some extent, complements the other in addressing some of its shortcomings.

I focus on SPE applications using randomized controlled experiments.[1] The use of randomized controlled experiments is pervasive in economics (and in other fields), and they are generally referred to as a gold standard in policy evaluation. By developing and estimating economic models in combination with such experiments, the SPE approach extends the analysis of the experimental data to analyze a »new« series of questions which are useful in the process of understanding and interpreting the experimental results and in designing future follow-up experiments. The intention is thus to further capitalize from the experiments which can be quite costly to conduct. I illustrate this by examples from the literature with a particular emphasison Denmark. The focus of the paper is on the evaluation of labor market policies generally, however the overall methodological discussion, of course, extends beyond this area.[2]

The outline of this paper is as follows: first I provide a brief introduction to and overview of a part of the methodological literature relevant for this paper. This sets the stage and introduces some of the central ideas going forward. I then proceed with a selective review of the Danish labor market experiments particularly focused on the first wave of experiments (Quickly

1. Randomized controlled experiments have a long history in economics and date back to at least 1968 and the New Jersey Income Experiment in the U.S. The definition of randomized controlled experiments (or randomized controlled trials) sometimes also includes field experiments and social experiments. For a discussion of these concepts see List and Rasul (2011); Rothstein and von Wachter (2017).
2. It is not the aim of the paper to argue in favor of one approach over another nor to say that the SPE approach is superior in all dimensions. Instead, this paper argues that different approaches have different advantages and, thus, there are questions where a combination of theory and randomized experiments can be useful in order to make further progress.

Back To Work 1+2). Finally, I provide two examples of the SPE approach used in relation to the Danish labor market experiments.

2. Methods

> »Few topics in economics evoke more passion than discussions about the correct way to do empirical policy analysis.«
>
> James J. Heckman, 2010

The literature on the evaluation of (labor market) policy/programs can broadly and stylized be divided into two different strands which have developed independently through the last decades. They are sometimes distinguished by how they weight internal and external validity in their analysis. A study is said to have internal validity if the research design and findings are credible for the sample population under study. External validity refers to whether the observed impacts are valid outside the sample population and environment. This concerns whether the results generalize beyond the current setting, or more formally whether parameter invariance permits credible extrapolation beyond the population under study (Manski (2013)).[3] Below I briefly introduce these two different strands of the literature and some of the arguments in favor of and against each approach, respectively.

One part of the literature can be referred to as Empirical Policy Evaluation (EPE). This is a comprehensive research agenda centered around the analysis of the causal impact of a policy (see reviews by Angrist and Krueger (1999); DiNardo and Lee (2011)). This literature is focused on the identifying assumptions and in particular the identifying variation in the data exploited in the analysis of the policy. Randomized controlled experiments are often referred to as the gold standard and guide how the analyst thinks about the empirical research design – in lack of experimental variation, the analyst tries to find variation in the data that mimics the variation of an experiment.[4]

3. In the words of Manski (2013): »... analysts regularly assume that the distribution of outcomes that would occur under the policy of interest would be the same as the distribution of outcomes realized by a specific experimental treatment group.«
4. The value added of experiments in modern policy making should not be understated (neither the complexity involved in designing and conducting them appropriately), see for instance List (2011) and Duflo (2017). In the words of Duflo (2017) »similarly, in the service of fitting policies for the real world, field experiments are a natural complement to theory and economic intuition. Evaluation in the field is necessary for the plumbers (red. economists) because intuition, however sophisticated and however well grounded in existing theory and prior related evidence, is often a very poor guide of what will happen in reality.«

The primary and powerful advantage of randomized controlled experiments is that they, by design, solve the classical selection problem into treatment. Let Y_{1i} denote the potential outcome for individual i under treatment (T – treated) and Y_{0i} the outcome under non-treatment (NT) also referred to as the control. The average causal effect of a policy $E(Y_{1i} - Y_{0i}|T)$ is the parameter of interest (this is known as the average treatment effect of the treated). What we can construct in the data, however, is the difference in outcomes for treated and controls, this can be written as:

$$E(Y_{1i}|T) - E(Y_{0i}|NT) = E(Y_{1i} - Y_{0i}|T) + \{E(Y_{0i}|T) - E(Y_{0i}|NT)\} \quad (1)$$

This decomposition illustrates how a standard comparison of treated and controls in the absence of random assignment will consist of both the parameter of interest and an additional component referred to as selection bias (the part in {}). Selection bias arises if the assignment process, the selection into treatment, is not random/controlled. It reflects that differences in out-comes may simply arise because individuals in the treatment and control groups are different due to (self-) selection into treatment. Randomization ensures that the selection bias is 0 in expectation, see e.g. Heckman et al. (1999). The analyst creates an assignment rule (by randomization) into treatment which is uncorrelated with observable and unobservable characteristics of potential participants on average. This reduces the evaluation problem to one of comparing the average outcome in the treatment group to the control group.

Note that the key idea is to use the average outcome in the control group ($E(Y_{0i}|NT)$) as the counterfactual outcome for the treatment group ($E(Y_{0i}|T)$). But this comparison is only meaningful if the control group is not affected by the presence of treatment – formally this is referred to as the stable stable unit treatment assumption (SUTVA) (see DiNardo and Lee (2011)). The assumption implies that the outcome of the control group should not change with the presence of treatment. If the SUTVA assumption is not satisfied, we are wrongly interpreting any difference as an effect of treatment on the treatment group, whereas in fact this was partly an effect on the control group. I return to the role of the SUTVA assumption in Section 4.1.

The EPE approach has been extremely influential and has led to a »credibility revolution« of empirical research (see Angrist and Pischke (2010)). The focus of this literature is on the internal validity of a given study. In the search for internal validity, external validity can be sacrificed and is viewed as of second order importance because the analysis centers around the part of the sample where the identifying variation is credible.[5]

5. See e.g Imbens (2010) where it is argued that: »better LATE than nothing ...«, meaning it is better to have high internal validity on a subset of the sample than low internal validity ge-

While the starting point in the EPE approach is finding or constructing appropriate identifying variation, another approach exists where decision theoretic parameters and the formulation of the individual decision problem is the starting point. This literature is here refered to as TPE (theoretical policy evaluation). In TPE economic models are developed and estimated, and a policy can be evaluated by analyzing predictions of the model (Keane (2010); Boone et al. (2007); Blundell (2010); Spinnewijn (2015)). Policy impacts arise from individual behavior which is modelled through a set of equations and an individual optimization problem.[6] By estimating the model, the link to individual behavior/decisions fundamentals which drive observed outcomes in data is quantified. The role of structure and the assumption of universal (or invariant) behavioral relationships allow the TPE structural framework to analyze how impacts generalize outside specific settings (e.g. the specific sample). As a result, the quality of subsequent predictions is closely linked to whether the behavioral problem is well-specified initially.

Generally, a discussion of the pros and cons of each approach is ongoing in the literature (see e.g. Angrist and Pischke (2010); Sims (2010); Banerjee and Duflo (2009); Deaton (2010) and the references above). The advocates of EPE criticize the amount of untestable assumptions used in the TPE framework (the internal validity) including the required structure and parametric assumptions which are often needed to make the solution of the models possible and tractable. Advocates of TPE criticize the EPE framework for limited external validity.[7]

Heckman (2010) argues that while EPE is well suited for answering questions about the (local) impact of an already existing policy, other questions are also crucial for the design of future policies. In particular, Heckman (2010) defines three layers of learning by separate questions:

nerally. External validity is still valued in this literature, but the focus is on establishing a credible causal impact on some subset of the sample. Subsequently the analyst can then discuss external validity of this estimate by, e.g., discussing how representative the individuals under study are and potentially extrapolate from there (Manski (2013)).

6. The parameters guiding these relations are also called structural or policy invariant parameters, see Heckman (2010): »This concept received its clearest statement in a classic paper by Leonid Hurwicz (1962). A structural relationship in its original usage is a relationship invariant to a class of policy interventions and can be used to make valid policy forecasts for policies in that class. The explicit parametrizations used in the modern version of the structural literature are intended to represent policy invariant parameters.«
7. This is also discussed in Rothstein and von Wachter (2017) as: »In contrast, structural approaches that explicitly specify all aspects of the choice problem and resulting outcomes can in principle resolve both assignment and other design issues simultaneously. However, this approach hinges on the model being correctly specified, and hence may come at a substantial cost to internal validity.« and »So-called structural methods generally trade off internal validity in pursuit of more external validity, but a study that fails to solve the assignment problem is unlikely to be any more generalizable than it is internally valid.«

– What is the effect of an already implemented policy? (ex post policy evaluation)
– What will be the effect of a previously implemented policy in the future? (can the findings be generalized?)
– What would be the effect of a new policy in the future? (ex ante policy evaluation)

The first layer concerns establishing the »true« effect of the policy and thus requires internal validity. The second and third layers concern the external validity of a given finding, and Heckman argues that economic models are particularly useful here (see also Todd and Wolpin (2008); Wolpin (2013, 2007)).[8] This argument hinges on the idea that models specify economic behavioral relationships which are policy invariant (see also Keane (2010)). The use of models implies further assumptions in the analysis, but these assumptions are argued to be necessary in order to generate further progress and approach questions from layer 2 and 3.

As explained earlier, this review concentrates on applications of the SPE approach using randomized controlled experiments only.[9] Related to the three layers above, Rothstein and von Wachter (2017) outline some limitations, or areas where randomized experiments in themselves may not provide sufficient information. Some of these limitations are: 1) Spill-overs and market level impacts, 2) Questions about heterogeneity, 3) Questions about generalisability, and 4) Questions about mechanisms. In the words of Rothstein and von Wachter (2017): »But it (the experiment) cannot solve all identification problems faced by program evaluators, nor answer all questions posed by labor economists seeking to understand the workings of the labor market.«

2.1. SPE

During the last 15-20 years a new literature has emerged which can be seen as a mixture of the EPE and TPE approach. I refer to this literature as »Structural and Empirical Policy Evaluation« (SPE). The strength of the SPE design is an eclectic use of both approaches presented above (Blundell (2010)). This implies that economic models are specified and estimated in connection with, for example, policy reforms or randomized experiments. The aim of the SPE approach is to combine the best from both strands of the literature: economic theory extends the range of

8. Frisch (1933): »... No amount of statistical information, however complete and exact, can by itself explain economic phenomena... we need guidance of a powerful theoretical framework. Without this no significant interpretation and coordination of our observations will be possible.«
9. One additional advantage of randomized experiments compared to other quasi-experimental variation (e.g. labor market reforms) is that it is often easier to specify the information structure and expectations prior to »treatment«. Using labor market reforms also requires thinking about the role of anticipation effects, see Blundell et al. (2011).

questions we can address, while the focus on identification and exogenous variation should decrease concerns about results being driven by, for example, model misspecification.[10]

The literature related to the SPE paradigm presents different ways of using the experimental variation. Two different approaches can be distinguished by whether the experimental variation is used for model validation (out-of-sample predictions) or model identification. I illustrate these approaches by two examples below, see also a formal discussion on the use of randomised experiments and model validation in Schorfheide and Wolpin (2016, 2012).

An early example of the model validation approach is Todd and Wolpin (2006), who use data from the PROGRESA experiment in Mexico an experiment designed to evaluate the effectiveness of conditional cash transfers on, e.g., school attendance of children. They estimate and validate a dynamic model of parental decisions about fertility and child schooling. Their aim is to use the model to analyze whether better treatments exist besides the one implemented in the experiment. The model is estimated on the control group and the data from the treatment group is a hold-out sample and is subsequently used as a benchmark for an »out of sample prediction«. The idea is that this validation exercise should increase the credibility of the model and thus the various counterfactual experiments that the authors carry out in order to generate advice on alternative treatments.

Attanasio et al. (2012) provide an early example of the model identification approach where data from both treatment and control groups are instead directly used in estimation to identify parameters or mechanisms that would otherwise be hard to identify. They also use the PROGRESA experiment and analyze how the type of household income (e.g. school grants versus child labor income) matters for the decisions of the household. The response could generally differ depending on the source of income because the marginal utility associated with different income sources may differ due to, for example, preferences for child labor, within household allocations/bargaining, or other preference shifters. Furthermore, Attanasio et al. (2012) use the experiment to estimate equilibrium effects of the conditional cash transfers on child wages in treated villages. To estimate both equilibrium effects and preferences for the subsidy, Attanasio et al. (2012) use the data from both the treatment group and the control group in the estimation.

Other examples of the model validation approach include Lise et al. (2015, 2004); Duflo et al. (2012), while Ferrall (2012) is an example of the model identification approach. The unifying theme in all these contributions is the use of economic models to further capitalize on the experimental variation and analyze

10. Another related literature is the development of sufficient statistic formulars such as, for example, the Baily Chetty formula, see Chetty (2009) and Kolsrud et al. (2018). This literature focuses on evaluating already existing policies and is not reviewed here, but see Kleven (2020).

questions in areas where the experimental variation in itself is not sufficient (see Rothstein and von Wachter (2017)). In Section 4 I discuss two applications of the SPE approach in the context of the Danish labor market experiments in further detail. First, however, I provide some background for the experiments.

3. The Danish Experiments

Active Labor Market Programs (ALMP) are considered an important part of the Danish flexicurity model (see Andersen and Svarer (2007)). Since resources spent on ALMP activities are many[11], there is always a focus on prioritization of resources and the effectiveness of different elements. The Danish National Labor Market Authorities has a strategy of »evidence based policy making«. As a part of this strategy, randomized controlled experiments were introduced, and they were designed partly in collaboration with researchers and conducted by local employment authorities. Below I review the danish labor market experiments. I focus in particular on the »first wave« of experiments called (Quickly Back To Work 1+2) since data from these experiments is used in two applications of the SPE approach which I discuss in the next section. Lastly, I also (selectively) review some of the experiments from later waves, these experiments are so far unexploited in the SPE approach and therefore discussed in less detail.[12]

3.1. Quickly back to work – first wave

The first labor market experiment in the Quickly Back To Work series was conducted in 2005. The experiment was called Quickly Back to Work 1 and was targeted newly unemployed individuals eligible for unemployment benefits. It was conducted at selected locations in two Danish regions: Storstrøm and Sønderjylland.[13]

Formally, treated individuals received a letter at inflow into the experiment (typically the week they start on unemployment benefits) which informed them about their participation in a pilot study and that their participation in the ex-

11. Direct costs of active labor market policies amount to close to 0.7 % of GDP in 2012 in Denmark (see Ministry of Employment Expert Panel (2014)).
12. For a complete overview of these experiments see: https:// www.star.dk/viden-og-tal/hvad-virker-i-beskaeftigelsesindsatsen/randomiserede-kontrollerede-forsoeg-rct/, see also Rosholm (2014). Finally some of the lessons learned from the experiments below have also influenced policy (see Ministry of Employment Expert Panel (2014)).
13. See Graversen and van Ours (2008b,a) for details beyond what is presented below (including a list of the specific participating job centers). Note, that in the first wave of experiments »randomization« was based on individual's date of birth (born between 1-15th would be assigned to the treatment group) the idea being that this was easier implementable than an actual randomization procedure.

periment served as a conditionality for receiving unemployment benefits. The treatment consisted of three stages: First, treated individuals would participate in a two-week job search assistance course. Second, treated individuals participated in either weekly or biweekly meetings with a caseworker for the next 7 weeks. Third, the unemployed would enter an early activation program (either job search assistance, subsidized employment, classroom training, or vocational training) for at least 13 weeks. After approximately 40 weeks, remaining unemployed individuals in the treatment group would finish their participation in the experiment.

Empirical evaluations of the experiment (see e.g. Graversen and van Ours (2008b,a)) showed quite large employment effects, and when contrasted with the costs of running the experiment, the results were still favorable. In fact, direct cost savings (saved income transfers subtracted the costs of running the program) from the experiment were estimated to 500-600 Euros per individual see the Danish Economic Council (2007). Furthermore, the Danish Economic Council (2007) estimates that the economic gain to society from the experiment amounts to around 2000 Euro per individual. This calculation contrasts costs from running the experiment to the increase in earnings in the treatment group.[14] This calculation ignores the existence of potential individual level costs for participants (e.g. loss of leisure or costs of production) as I discuss in Section 4.2.

While the first experiment showed remarkable employment effects, it was less clear exactly why this was the case. In particular, and by design, the evaluation allowed for an immediate evaluation of the combined interventions, whereas a more detailed analysis of each specific element was not immediately possible.[15]

To learn more, a second experiment Quickly back to work 2 was designed where some of the interventions focused on elements of the full treatment in Quickly Back To Work 1.

This second experiment was implemented in 2008, and the impact of 3 different interventions (group meetings, individual meetings and early activation) was evaluated separately in 3 different regions (Mid Jutland, Northern Jutland and Zealand).[16] The empirical evaluation in Maibom et al. (2017) shows that after around 4.5 years the treatment group has accumulated more weeks in employ-

14. The increase in earnings is thus assumed to represent the value of increased production. The calculation also »corrects« for distortionary taxation (marginal costs of providing public funds). Saved income transfers are not directly included in the calculation of the economic gain to society as they simply represent a redistribution of income, however the gain achieved through less distortive taxation (because less income transfers are needed) is included.
15. Due to dynamic selection from exits out of unemployment, the remaining unemployed are a non-random sub-sample of those initially unemployed and randomized. Explicitly modeling this selection process would be an alternative to running a new experiment.
16. Further details about the experiment, the implementation and the employment effects are provided in Maibom et al. (2017).

ment across all three experiments, but the effect is largest in the region with meetings. The effect is close to 8 weeks (and statistically significant) and corresponds to an increase in the employment rate of around 5% over this period. Note further that comparing impacts across regions should generally be done with caution as it requires the additional assumption that the effect of treatment does not vary with regions (more below).

Maibom et al. (2017) calculate the impact from the experiment on the government budget – the different components in this calculation are given in Table 4 in Maibom et al. (2017). The gains part includes saved income transfers and an increase in tax revenue due to the increase in employment. The saved income transfers are adjusted to account for the impact of both direct and indirect taxes. The cost side involves costs associated with the experiment, but costs are also adjusted for the reduction in regular activities at the job center after the experiment due to the decrease in unemployment. Maibom et al. (2017) estimate that the impact on the government budget is a surplus of 5104 Euro per unemployed for meetings, again a substantial effect. Note that this calculation focuses on the impact on the government budget only. It is not an attempt to quantify the total economic gains or welfare (see more below).

3.2. Addendum – Later experiments

While the first two experiments were focused on the (average) impact on a broad group of unemployed individuals, later experiments became more targeted at »smaller« and also more vulnerable subgroups. The interventions studied were still similar, i.e., early and frequent interactions with caseworkers, but the results were not as encouraging in the second wave of experiments (perhaps partly due to the fact that the experiments now targeted a weaker group of unemployed).

For instance, Rosholm (2014) reports from two experiments (»Aktive hurtigere i gang« and »Alle i gang«): »In this case, the results were negative. While there was a marked increase in the exit probability from social assistance, there was no corresponding exit into employment. Rather, many social assistance recipients went onto disability benefits, which was hardly the program's intention. For the sick-listed, there was a similar effect.« This experiment primarily targeted recipients of social assistance (welfare benefits for workers who are not (or no longer) eligible for unemployment benefits) without a job for at least 6 months and some unemployed on sickness benefits, thus generally a weaker group of unemployed compared to the earlier experiments. Maibom et al. (2014) report similar findings from a Youth experiment (Quickly back to work 5) which was targeted unemployed youth (below 30) either on social assistance or unemployment benefits (both stock and inflow) in November 2009. Maibom et al. (2014) find very small and insignificant effects on employment and education for educated youth, while uneducated youth had negative employment effects and slightly positive (but in-

significant) education effects. Both groups showed an increase in sickness benefits.

More recent experiments have involved treatments other than the »usual« ones: meetings or early activation. Instead treatments were specifically designed to the targeted subgroups. One example is the »mentor trial«, where youth in risk of marginalization were assigned a personal mentor for up to a whole year. This experiment showed both positive employment and education outcomes (see Svarer et al. (2014) for a presentation).

All of these later experiments so far remain unexploited in an SPE setup, although an economic model would potentially be useful here as well (more below).

4. Quickly back to work and SPE

Below, I focus on two contributions that are examples of the SPE approach using data from the first wave of the experiments. They address two retrospective concerns, or intentions, to further learn about the driver of experimental impacts, which rose in response to the relatively large impacts on employment.

The first set of questions concerns the importance of congestion and equilibrium effects. The concern is that the large effects on employment are partly »spurious« as they arise because individuals in the treatment groups get jobs at the expense of members in the control group. If such congestion effects are important, this implies that the SUTVA assumption in equation 1 is violated, and treatment-control differences are as such uninformative about the impact of the experiment. The reason is that both groups are affected by the experiment, and the difference in outcomes is only informative on the differential effect of the policy but not the policy itself. Further, if congestion effects are important, any experimental results will be of limited policy relevance because the impact of the policy depends on how many individuals are treated in a given labor market. Hence to evaluate the effect on a full implementation of the policy, you need to treat the full population.[17]

The second set of questions concerns the mechanisms behind the experimental impacts and the role of ALMPs more generally. One hypothesis is that the experimental impacts arise because that individuals value unemployment less due

17. Crepon et al. (2013) study displacement of jobs in a two stage experimental setting where both the intensity of treatment (the fraction treated) and the randomization into treatment is varied across local labor markets. The difference in the outcomes of control groups across labor markets with a high and low intensity can then be used to quantify the total congestion effect in a specific setting. In the Danish labor market setting two stage randomization is harder to implement as the labor market is limited in size – there are simply not that many separated local labor markets to randomize intensities to.

to the experimental treatment. From the empirical literature we know that ALMP activities can be associated with threat effects, i.e., employment effects already prior to actual program participation.[18] Threat effects would arise if future participants view participation as costly and in response change job search to avoid future participation. Such individual program costs are important, as their existence opens up a question of whether it is still beneficial for society to have these programs or whether these individual costs are too high compared to the benefits of the program. These costs are unaccounted for in previous work. In fact, evaluations of social programs and the literature on policy evaluation more generally often ignores the loss of leisure and other nonpecuniary costs (e.g. stigma) associated with program participation (see discussions in, e.g., Heckman et al. (1999); Greenberg and Robins (2008)).

Below, I will present two attempts to answer these questions and discuss how these answers involve introducing »more structure«.

4.1. Congestion effects

Gautier et al. (2018) analyze the importance of congestion and equilibrium effects associated with ALMPs using Quickly Back to Work 1. Ultimately they are interested in analyzing what the impact would be of the intervention if it was rolled out as a permanent policy affecting all individuals.

The SUTVA assumption (see Section 2) would be violated if, for instance, the labor market has a fixed number of vacancies (inelastic labor demand). If treatment makes the treated individuals search more or better for jobs than the controls, they may get some of the jobs at the expense of the control group (who would get some of these jobs in absence of the experiment).

Broadly speaking, the importance of such congestion effects depends on: 1) How large the increase is in applications or competition for the control group. This depends on how many are treated and how individuals respond to treatment. 2) The extent to which labor demand is elastic, i.e., whether and to what extent firms respond by posting more (or less) vacancies. The first part is a direct (partial) competition effect, and the second part is a labor demand effect. The full size of the congestion effect is the combination of the two, and in equilibrium the relative importance of these channels would depend on the share of individuals who are treated.

Using a difference-in-difference procedure, Gautier et al. (2018) first show that the members of the control group find jobs at a slower pace after the start of the experiment compared to workers in other regions (where there was no experiment). This suggests that the control group is in fact adversely affected by the experiment. They also look at the supply of vacancies and find some evidence of an

18. See for instance Black et al. (2003); Hagglund (2011); Geerdsen (2006).

increase in the supply of vacancies. Hence, it looks as if both competition and labor demand effects arise due to the experiment. To disentangle these differential channels and analyze the role of congestion effects in a fully rolled-out policy, we need an organizing framework which contains a description of how these channels interact and an (empirically based) quantification of the different elements such that we can analyze how the combined effect varies with the fraction of individuals treated.

Gautier et al. (2018) set up an equilibrium search and matching model which can be seen as an extended version of the Diamond-Mortensen-Pissarides model (see Mortensen and Pissarides (1999)). The key feature is that the success of an application, i.e., the offer of a job, depends on the search behavior of other unemployed and the number of vacant jobs. The experiment enters the model as a reduction in the cost of sending out applications for treated individuals and a change in the effectiveness of their applications.

The model allows firms to respond to an increase in applications by changing the number of vacancies they post. Firms take into account that searching for workers, or new matches, is a frictional and time-consuming process and that the probability of a successful match also depends on the hiring policy of other firms. The model has an endogenous matching function with two key coordination frictions: i) workers do not know where other workers apply and ii) firms do not know if any of the applicants they meet are also candidates considered by other firms. If an unemployed worker receives multiple job offers, the worker will randomly choose one of them. Therefore, if the unemployed workers respond to the experiment by sending many more applications they are also making it more likely that they will get multiple job offers and hence turn one down – from the perspective of the firm, this is not attractive and they may respond to an increase in applications per worker by posting fewer vacancies. The results in Gautier et al. (2018) suggest that this later channel is important.

The model is estimated on data from the experiment. The imposed structure, i.e., the estimated equilibrium search and matching model, allows the authors to analyze the role of congestion effects in the results from Quickly back to work 1. The results in Gautier et al. (2018) imply that congestion effects play an important role in explaining the large employment effects in the experiment. They find that around one third of the employment effects of the experiment is a result of decreased job finding in the control group. They use the structural model to analyze what happens to the equilibrium if the policy applies to everyone in the labor market, i.e. the share of treated workers is one. They find that that the equilibrium unemployment rate is largely unchanged (it is actually marginally increasing) in response to the experiment (see Figure 6 in Gautier et al. (2018)). This is clearly opposite to the results of the simple treatment control comparison, and illustrates the importance of analyzing the mechanisms behind the observed impacts in the experiment. The results suggest that labor demand is simply not elastic enough to

absorb the increase in applications from the experiment. In fact, firms may actually post fewer vacancies.

Overall, there is a clear link back to the SPE paradigm in this work: the authors use Quickly back to work 1 as the benchmark of what the model should predict at a certain treatment intensity.[19] This link improves the internal validity of the exercise, while the imposed structure allows the authors to extrapolate the results and also simulate other policies.

4.2. Welfare effects

Maibom (2021) uses data from Quickly Back To Work 2 to estimate the effect of the experimental treatment and more generally the role of ALMPs on individual decision making. A goal of the analysis is to quantify any utility costs associated with program participation and analyze the implications for overall welfare.

The empirical problem is that utility costs, which may be both pecuniary and non-pecuniary as in e.g. Moffitt (1983), are generally unobserved and hard to measure.[20] [21] Maibom (2021) follows a revealed preference approach and estimates utility costs (and benefits) using data on, for example, employment outcomes in combination with a structural model of job search. The model generates a link between observed outcomes in the data and decision theoretic parameters such as utility costs and benefits, while the experiment generates exogenous variation in the treatment intensity. The internal validity of estimates on utility costs are closely linked to whether the counterfactual, i.e., the hypothetical behavior in absence of treatment, is correctly specified. This explains why the experiment is important for the analysis and why data on both the control and treatment group is used for estimation and identification of utility costs and benefits.

The analysis takes the mandatory element of ALMP participation into account. In the Danish setting (and in many other contexts, see Maibom (2021)), program participation is a conditionality for receiving benefits. Individuals may actually participate in programs they dislike simply because they do not have available

19. In the words of Gautier et al. (2018) »We exploit the fact that, due to the experimental design, the increase in search intensity of participants in the activation program is truly exogenous. This makes the identification of the structural parameters more convincing than in typical calibration exercises«.
20. The program participation cost could consist of several parts such as direct costs from the loss of leisure, costs related to an increase in effort (visiting the job center or preparing for meetings etc.) or even stigma associated with program participation or visiting the public employment service. Maibom (2021) does not distinguish between these different sources of costs.
21. It may also be challenging to extract these costs ex ante (in e.g. surveys) primarily because there are some challenges in giving an appropriate metric and quantification. Further, evidence in the literature suggests that program participants are not correctly assessing whether a (labor market) program is beneficial or not, see Smith et al. (2020). Lastly, there are some obvious incentive problems for participating individuals who may not state the true costs.

job offers (hence employment is not an alternative) and they only qualify for unemployment benefits if they participate.[22] Therefore, data on actual ALMP participation is not in itself informative about the size and relevance of utility costs. As a further complication, the programs may also have beneficial elements, which implies that direct utility costs are not the only thing that drives the impact of the experiment and individual decision making.

Job seekers instead indirectly reveal their preferences for ALMP participation through job search. By searching more or less the job seekers can influence the risk of future participation in ALMPs. The model in Maibom (2021) maps (indirectly) revealed preferences such as differences in, for example, employment rates between treatment and control groups into differences in behavior such as job search and reservation wages. These later differences are then directly informative about the valuation of participation in the programs.

The model in Maibom (2021) is set up to represent the central decision problem participants face while being enrolled in the experiment. The model is an infinite horizon, discrete states, discrete choices, dynamic program. Each period unemployed job seekers choose a level of search activity and whether to accept a given job offer or not (if a job offer is present). By searching more or less, the participants influence the likelihood of receiving a job offer in the future and thus also the probability of future participation in ALMPs.

In the model there are two groups of participants – treated and controls. At the beginning of the experiment treated and controls are identically distributed across states, and the only difference is that during the next periods the treatment group ... treatment groups progress through a set of experimental stages: First, they enter a waiting/threat stage were they know of the future treatment but are not treated yet. Second, they enter the treatment stage where the actual treatment occurs. Lastly, they enter a post-treatment environment which is identical to the environment controls »live« in.

In the model »treatment« has two effects. First, if utility costs are important, the threat of future program participation raises incentives for individuals to search for a job and thus leave unemployment. Therefore experimental impacts can arise already from the point where treated individuals learn about the experiment, i.e., the threat stage. Secondly, if benefits from program participation are important, participation in the treatment stage is associated with an increase in the probability of getting a job offer. There is therefore a useful source of identification of costs versus benefits from the programs in the time profile of impacts – if impacts arise very early, and even prior to treatment, the model will interpret this as evidence in favor of utility costs. If impacts arise after program participati-

22. Of course if utility costs are large enough to dominate the expected sanction (or loss of benefits) from non-participation in ALMPs, the job seekers may simply choose not to participate. The estimates in Maibom (2021) suggest this is not the case.

on the model will interpret this as evidence in favor of program gains. The explicit inclusion of the differential treatment stages implies that model predictions can be contrasted to data over the period of the experiment and thus leverage identification of utility costs and benefit (see Maibom (2021)).

The estimates from the model suggest that utility costs associated with program participation are sizable. For low and medium educated workers the direct utility costs associated with activation are only 15-20 percent lower than the costs of working which seems reasonable given that activation is often nearly as time-consuming as full-time employment.

The estimates imply that low educated workers would be willing to give up close to 80 (50) percent of their UI in a given week to avoid participation in activation (meetings). From a welfare perspective utility costs are important to include in evaluating whether a given level of ALMP is actually beneficial. However, direct utility costs associated with unavoidable program participation are not the appropriate metric if we want to evaluate the overall effect of the experimental treatment and calculate the social benefits of the experiment in a full welfare analysis, the reason is that utility costs can be avoided through job finding and hence not realized.

To quantify the effect on worker welfare of the experimental treatment, Maibom (2021) calculates the compensating variation. The compensating variation is the monetary payment which makes individuals in the treatment group indifferent between belonging to the treatment or the control group at the beginning of the experiment. The average compensating variation amounts to around 1.5 weeks of UI benefits per individual in the treatment group. By comparing the compensating variation and actual operating costs of the experimental treatment to the gains from increased job finding, Maibom (2021) performs a welfare analysis and shows that accounting for the compensating variation and the cost of working, two objects which are obtained using the job search model, substantially reduces the social benefits of the experiment.[23] Ovverall the analysis illustrates that accounting for the effect of ALMPs on worker welfare is important in assessing whether a given level of ALMPs is appropriate from a welfare perspective.[24]

23. Note that this analysis is different to that of Maibom et al. (2017) which focus on the direct impacts on the government budget only and abstains from an analysis of the welfare effects of the experiment as such.
24. The analysis works under the assumption that the role of congestion effects is limited in creating the experimental impacts. Note that there may be reasons to expect congestion effects to be lower in Quickly Back To Work 2 than Quickly Back To Work 1. Primarily because the intervention is not as intense and because the regional labor markets are larger labor markets. Note further that there is a difference in the question studied: while Gautier et al. (2018) use the experimental variation to create predictions about the impacts of a full roll out of the program, Maibom (2019) focuses on the implications for worker welfare of the experiment in its current form.

The estimate of the model further show that the compensating variation is heterogeneous across individuals due to differences in their ability to leave unemployment and the value of alternatives. These results are important in a discussion about the optimal usage of ALMPs and can be linked back to research about optimal design of transfer systems more generally (see e.g. Besley and Coate (1992); Nichols et al. (1982); Kreiner and Tranæs (2005)).

Overall, the analysis in Maibom (2019) also links to the SPE approach. The analysis specifies an economic model and uses it in combination with credible exogenous variation to advance the analysis from assessing immediate impacts of the treatment to using the experiment as a source of variation to analyze a more fundamental question about the underlying mechanism and welfare implications of the experimental impacts.

4.3. SPE and later waves of experiments

As described above, later waves of the Danish labor market experiments were less successful in terms of promoting employment for unemployed job seekers. These experiments so far remain unexplored with the SPE approach, although such analysis could be useful in analyzing how impacts (or lack of) generalize or are shaped by, for example, different constraints facing different subgroups or different elements of the treatment »package«. Another related issue is how regional differences, and differences in the business cycle, affect different experimental impacts.

More broadly using economic models to think about future and past experiments may also prove usefull in organizing the accumulation of knowledge across different experiments further. For instance, the results in Maibom (2019) suggest that simply re-targeting the interventions to weaker groups of unemployed should not generate the same employment effects as in (Quickly Back To Work 2. The reason is that the primary mechanism underlying the experimental impacts is that treated individuals search harder for employment, and less so change their preference about the type of jobs they are willing to accept. In consequence, impacts depend on the return to such increased job search and the treatment effects are heterogeneous. With this in mind a follow up experiment could perhaps focus on testing this prediction more directly by measuring job search and help in identifying key constraints in why the return to search is low for specific group of workers, e.g., are they searching in-optimally. Of course, such results and predictions are almost surely not the final word but could represent useful steps moving further forward. As summarized in Blundell (2017): »... There is a strong complementarity between these approaches. Quasi-experimental evaluations can provide robust measures of certain policy impacts but are necessarily local and limited in scope. Structural estimation allows counterfactual policy simulations which can then feed into a policy (re-)design analysis«

5. Conclusion

Labor market experiments are an important part of the strategy for evidence based policy making set forward by the national labor market authorities in Denmark. Since 2005 several randomized experiments have been conducted in order to obtain new insights about appropriate and cost effective policy measures that help secure a well functioning labor market. The prevalence of these labor market experiments have also stimulated research.

This paper has presented a part of this research focused on a new research agenda »structural and empirical policy evaluation« (SPE). Behind this agenda lies an ambition to achieve both high internal and external validity. This is achieved by combining economic models and credible empirical research designs. Internal validity is increased by relying on credible exogenous variation such that the data patterns are causal and not driven by unobserved confounders. Economic models increase the external validity of the analysis as they provide a framework in which to extend the analysis from assesing immediate impacts to questions regarding optimality, generalizability, and effects of alternative policies without historical precedent. Of course, the cost of imposed structure and assumptions is not innocent and this paper does not argue that the SPE approach should be followed in all research. More specifically, it depends on the type of questions which the analysis aims to address.

The SPE approach has been illustrated in two examples in relation to the Danish labor market experiments. In particular, the approach has been used to analyze the importance of congestion and equilibrium effects, and to analyze the impacts on welfare for society from ALMPs. Both examples illustrate that important progress can be made using this approach.

References

Andersen, T. M. and Svarer, M. (2007). Flexicurity – Labour market performance in Denmark. CESifo Economic Stndies.

Angrist, J. D. and Krueger, A. B. (1999). Chapter 23 Empirical strategies in labor economics. In Ashenfelter, O. and Card, D., editors, Handbook of Labor Economics, volume 3 PART of Handbook of Labor Economics, chapter 23, pages 1277-1366. Elsevier.

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. Jonrnal of Economic Perspectives, 24(2):3-30.

Attanasio, O. P., Meghir, C., and Santiago, A. (2012). Education choices in Mexico: Using a structural model and a randomized experiment to evaluate PROGRESA. Review of Economic Studies, 79(1):37-66.

Banerjee, A. V. and Duflo, E. (2009). The Experimental Approach to Development Economics. Annual Review of Economics, 1.

Besley, T. and Coate, S. (1992). Workfare versus Welfare: Incentive Arguments for Work Requirements in Poverty-Alleviation Programs. American Economic Review.

Black, D. A., Smith, J. A., Berger, M. C., and Noel, B. J. (2003). Is the threat of reemployment services more effective than the services themselves? Evidence from random assignment in the UI system. American Economic Review.

Blundell, R. (2010). Comments on: Michael P. Keane Structural vs. atheoretic approaches to econometrics . Jonrnal of Econometrics, 156(1):25-26.

Blundell, R. (2017). What have we learned from structural models? American Economic Review, 107(5):287-292.

Blundell, R., Francesconi, M., and van der Klaauw, W. (2011). Anatomy of Welfare Reform Evaluation: Announcement and Implementation Effects. IZA Discussion Paper.

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

Chetty, R. (2009). Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods. Annnal Review of Economics, 1(1):451-488.

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

Danish Economic Council (2007). Danish Economy, Spring Report. Technical report.

Deaton, A. (2010). Instruments, randomization, and learning about development. Jonrnal of Economic Literature, 48(2):424-455.

DiNardo, J. and Lee, D. S. (2011). Program Evaluation and Research Designs, volume 4 of Handbook of Labor Economics, chapter 5, pages 463-536. Elsevier.

Duflo, E. (2017). The economist as plumber. American Economic Review, 107(5):1-26.

Duflo, E., Hanna, R., and Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. American Economic Review, 102(4):1241-1278.

Ferrall, C. (2012). Explaining and forecasting results of the self-sufficiency project. Review of Economic Studies, 79(4):1495-1526.

Frisch, R. (1933). Editor s Note. Econometrica, 1(1):1-4.

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

Geerdsen, L. P. (2006). Is there a threat effect of labour market programmes? A study of ALMP in the Danish UI system. Economic Journal, 116(513):738-750.

Graversen, B. K. and van Ours, J. C. (2008a). Activating unemployed workers works; Experimental evidence from Denmark. Economics Letters, 100(2):308-310.

Graversen, B. K. and van Ours, J. C. (2008b). How to help unemployed find jobs quickly: Experimental evidence from a mandatory activation program. Jonrnal of Pnblic Economics, 92(10-11):2020-2035.

Greenberg, D. H. and Robins, P. K. (2008). Incorporating nonmarket time into benefit-cost analyses of social programs: An application to the selfsufficiency project. Journal of Public Economics, 92(3-4):766-794.

Hagglund, P. (2011). Are there pre-programme effects of active placement efforts? Evidence from a social experiment. Economics Letters, 112(1):91-93.

Heckman, J. J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. Jonrnal of Economic Literatnre, 48(2):356-398.

Heckman, J. J., Lalonde, R. J., and Smith, J. A. (1999). Chapter 31 The economics and econometrics of active labor market programs. In Handbook of Labor Economics, volume 3 PART, pages 1865-2097.

Imbens, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). Jonrnal of Economic Literature, 48(2):399-423.

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

Kleven, H. (2020). Sufficient Statistics Revisited. Forthcomming in Annual Review of Economics.

Kolsrud, J., Landais, C., Nilsson, P., and Spinnewijn, J. (2018). The optimal timing of unemployment benefits: Theory and evidence from Sweden. American Economic Review, 108(4-5):985-1033.

Kreiner, C. T. and Tranæs, T. (2005). Optimal workfare with voluntary and involuntary unemployment. Scandinavian Journal of Economics, 107(3):459-474.

Lise, J., Seitz, S., and Smith, J. (2004). Equilibrium Policy Experiments and the Evaluation of Social Programs. Working Paper 758, National Bureau of Economic Research.

Lise, J., Seitz, S., and Smith, J. (2015). Evaluating search and matching models using experimental data. IZA Jonrnal of Labor Economics, 4(1).

List, J. A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. Journal of Economic Perspectives, 25(3):3-16.

List, J. A. and Rasul, I. (2011). Field Experiments in Labor Economics, volume 4 of Handbook of Labor Economics, chapter 2, pages 103-228. Elsevier.

Maibom, J. (2021) The Welfare Effects of Mandatory Reemployment Programs: Combining a Structural Model and Experimental Data. Working Paper.

Maibom, J., Rosholm, M., and Svarer, M. (2017). Experimental Evidence on the Effects of Early Meetings and Activation. Scandinavian Journal of Economics.

Maibom, J., Rosholm, M., Svarer, M., and Rosholm, M. (2014). Can Active Labour Market Policies Combat Youth Unemployment? Nordic Economic Policy Review, (Discussion paper no. 7912).

Manski, C. F. (2013). Public Policy in an Uncertain World.

Ministry of Employment Expert Panel (2014). Veje Til Job – en arbejdsmarkedsindsats med mening. Technical report, Ministry of Employment.

Moffitt, R. (1983). American Economic Association An Economic Model of Welfare Stigma: The American Economic Review, 73(5):1023-1035.

Mortensen, D. T. and Pissarides, C. A. (1999). Chapter 39 New developments in models of search in the labor market. In Ashenfelter, O. and Card, D., editors, Handbook of Labor Economics, volume 3 PART of Handbook of Labor Economics, chapter 39, pages 2567-2627. Elsevier.

Nichols, A., Zeckhauser, R., and Zeckhauser, R. (1982). Targeting Transfers through Restrictions on Recipients. American Economic Review, 72(2):372-77.

Rosholm, M. (2014). Do case workers help the unemployed? IZA World of Labor.

Rothstein, J. and von Wachter, T. (2017). Social Experiments in the Labor Market. Handbook of Economic Field Experiments, 2:555-637.

Schorfheide, F. and Wolpin, K. I. (2012). On the use of holdout samples for model selection. American Economic Review, 102(3):477-481.

Schorfheide, F. and Wolpin, K. I. (2016). To hold out or not to hold out. Research in Economics, 70(2):332-345.

Sims, C. A. (2010). Comment on angrist and pischke. Journal of Economic Perspectives, pages 1-9.

Smith, J. A., Whalley, A., and Wilcox, N. T. (2020). Are Program Participants Good Evaluators? IZA Working Paper, IZA DP No.

Spinnewijn, J. (2015). Unemployed but optimistic: Optimal insurance design with biased beliefs. Journal of the European Economic Association, 13(1):130-167.

Svarer, M., Rosholm, M., Havn, L., and Høeberg, L. (2014). Evaluering af mentorindsats til unge uden uddannelse og job. Rambøll Management Consulting.

Todd, P. and Wolpin, K. I. (2008). Ex Ante Evaluation of Social Programs. Annales d Economie et de Statistiqne, (91/92):263.

Todd, P. E. and Wolpin, K. I. (2006). Assessing the impact of a school subsidy program in mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. American Economic Review, 96(5):1384-1417.

Wolpin, K. I. (2007). Ex ante policy evaluation, structural estimation, and model selection. American Economic Review, 97(2):48-52.

Wolpin, K. I. (2013). The Limits of Inference Withont Theory. Tjalling C. Koopmans Memorial Lectures. The MIT Press.