The evaluation of welfare state performance: modelling a counterfactual world

Ennen, Jens

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Ennen, J. (2009). The evaluation of welfare state performance: modelling a counterfactual world. *Historical Social Research*, *34*(2), 129-146. https://doi.org/10.12759/hsr.34.2009.2.129-146

The Evaluation of Welfare State Performance: Modelling a Counterfactual World

Jens Ennen [*]

Abstract: *»Was wäre gewesen, wenn...? Evaluation sozialstaatlicher Maßnahmen mithilfe kontrafaktischer Modelle«*. The evaluation of welfare state performance is an important issue in times of tight government budgets, high unemployment and growing inequality. Policymakers and taxpayers want to know if a specific programme has led to the intended effect, and with no excessive waste of resources. For such evaluations to be thorough and robust, appropriate methods and the right counterfactuals are important. It is difficult to say what would have happened if a certain policy had not been implemented or implemented differently. This holds even more for the impact on a single individual than for aggregate results. This article will highlight some examples and possibilities of how to deal with counterfactual questions in the context of the Hartz reforms, probably the most far-reaching welfare state and labour market changes in the history of the Federal Republic of Germany. Furthermore this reform was the first big attempt of systematic welfare state evaluation in Germany.

Keywords: Welfare state, evaluation, counterfactuals, Hartz.

Introduction

Governments of industrialised countries are providing welfare schemes for citizens in need. Of course the specific organisation, financing and generosity of such mechanisms differ between economies, but the common motivation is that all people are subject to risks like illness, old-age and unemployment. The extent of exposure varies with age, gender, class and personal characteristics, but by pooling individual risk over large parts of society, compensation schemes reduce individual exposure in terms of the maximum potential financial loss.

As a consequence of the substantial spending on welfare systems, taxpayers want to know if the money has indeed a positive impact on the targeted population. Furthermore, the question arises if the same or even better results could have been achieved with fewer resources. Imposing a tax on the public or introducing compulsory contributions in order to use them for the welfare system

[*] Address all communications to: Jens Ennen, Centre for British Studies, Humboldt-Universität zu Berlin, Mohrenstr. 60, 13353 Berlin, Germany;
e-mail: jens.ennen@staff.hu-berlin.de.

means that spending – on the personal level as well as by the state – is not available for alternative uses any more.

Finding out if spending has indeed reached the target without too substantial adverse effects elsewhere requires asking the question what the outcome would have been in the absence of the welfare programme or under an alternative scheme. This is in fact a counterfactual question because we cannot precisely know what would have happened if at a certain point in time we had not decided to set up the scheme. Time goes on and we cannot rewind the passage of time like a video tape.

Answering counterfactual questions is not easy and economists have to come up with methods to estimate the outcomes of paths that were not taken in reality. Such estimations are often not very accurate, but due to the importance of such questions, it may be better to have crude estimates than none.

In which situations do economists ask counterfactual questions about welfare state performance and which methods do they use in order to answer them?

This article will illustrate the importance of counterfactual questions by taking the comprehensive labour market reforms in Germany as an example. These so-called *Hartz reforms* are interesting units of observation because on the one hand they can be considered as the most far reaching reforms of the German labour market and welfare state in the history of the Federal Republic. On the other they were the first big attempt to professionally evaluate the effectiveness and efficiency of the German welfare system (Jacobi and Kluve 2006). First of all the context and the details of the reform will be described. Then an overview over the different ways of counterfactual estimation will be provided before evaluation results will be presented.

The Hartz Reforms: Context

The main idea behind the comprehensive labour market and welfare reforms, which are usually called *Hartz reforms,* had the general purpose to reduce the extent of unemployment in Germany by improving the matching of jobseekers with vacancies, by putting all able-bodied welfare recipients into one category and by reducing the generosity and duration of maximum unemployment insurance benefits. Specifically pressure was increased on jobseekers to take up vacant positions (ibid).

A commission named "*Moderne Dienstleistungen am Arbeitsmarkt*" [modern labour market services] was established by the federal government and chaired by Peter Hartz, board member and HR director of Volkswagen (which explains the unofficial name *Hartz reforms*).

Prior to the reform there had been concerns about the big amount and persistence of unemployment in Germany. Since the beginning of the 1970s the Federal Republic has suffered from an increase in the number of people out of

job. Of course there were recurring economic upturns with resulting falls in unemployment, but the general development of joblessness was step-shaped. After strong business cycles unemployment did not fall to the level that prevailed during the prior boom. This does not only hold for Germany but also for other (continental) European economies. According to economic theory the resulting structural unemployment can be explained by the existence of institutions that prevent market forces from reducing real wages to a level that would dampen the extent of unemployment. The purpose of this paper is not to present theories that explain unemployment and its sources, but the gradual increase in natural unemployment should be kept in mind when discussing the motivation for introducing the reforms. The increase in joblessness since the 1970s is not particular to the Federal Republic. What is indeed special about the German economy is the reunification with the former socialist East. The GDR's economy turned out to be less efficient than expected by experts, with factory closures and resulting mass layoffs that contributed to the breakdown of the old industrial centres of the East. The other former planned economies did not have the possibility to get support from a relatively rich West, but for Germany a quick approximation of living standards in the East to the level of the West was considered to be necessary for political and ethical reasons (ibid.). Therefore the relatively well-developed West-German welfare state was – with some special rules – imposed on the East. Avoiding too dramatic differences in material well-being between both parts was certainly a good intention, but turned out to be difficult and expensive for the reunified economy. Many researchers like Hagen and Steiner (2000) or Caliendo and Steiner (2005) have investigated the development of spending on active employment policy over time. The motivation for such policies is the belief that measures like spending on education for disadvantaged people or public job creation schemes fight the underlying causes of unemployment. Policymakers decided to introduce such policies especially in the East where the share of active policies amounted to 30.9 per cent of total spending on labour market policy, in contrast to only 23.9 per cent in the West (Caliendo and Schneider 2005). In 1991 more than 3 out of ten members of the East German labour force participated in active measures like training and public job creation (Wunsch 2005). In addition to this quantitative importance of active measures in terms of participants, these programmes lasted relatively long. On the other hand the welfare state created a situation in which unemployed individuals had no strong incentives to take up jobs, provided that they existed in some areas of the country. Engels (2001) came to the conclusion that in the year 1990 the average family income of a 4-person-household with three children was only 15.3 (11.5) per cent lower in the West (East) than of a similar household with a single earner who had an unskilled job.

The West-German economy would probably also have needed reforms in the absence of reunification, but this article will not deal with this counterfac-

tual question. Nevertheless it is safe to say that the German reunification has, in spite of its multiple benefits, intensified the pressures on the welfare state.

An important factor for the growing willingness to reform the labour market was a scandal within the *Bundesanstalt für Arbeit (BA)* [the federal labour office], whose civil servants had published falsified figures in order to provide an overoptimistic picture about the success in placing unemployed people with vacancies. This incident and the media coverage of the scandal then incited the federal government under Gerhard Schröder to accelerate efforts that lead to a reform of the BA and its policy. The proposals of the Hartz commission were then implemented by the red-green coalition. Here, I will only highlight the most important components of the reforms.

Hartz Reforms: The Changes

The multitude of changes introduced between 2003 and 2005 can be put into three categories, as described by Jacobi and Kluve (2006).

The first pillar of reforms had the purpose to improve the services provided by the government authorities. Given the BA's extremely poor placement results, one component of the reform was to change the organisational structure of the public employment services. A stronger customer-orientation with fixed personal caseworkers and a monitoring of the BA's placement results services was the idea behind the change. In spite of the stricter evaluation of their work, the employment services obtained a greater discretion in deciding how to put their objectives into reality. In order to underline the stronger service orientation the BA was rebranded *Bundesagentur für Arbeit* [Federal labour agency], which makes it sound more modern and not so much like an office run by civil servants. Furthermore, the government decided to introduce a wider range of market mechanisms into the placement and training activities. After six weeks of unsuccessful placement jobseekers can ask for vouchers that make them eligible for support by private firms that are compensated if they succeed in finding an appropriate position. Private firms can also apply for offering training to unemployed persons. In order to make sure that an unemployed individual gets the support that corresponds to his or her needs, caseworkers put jobseekers into several categories corresponding to the likelihood of being placed quickly. This stronger targeting is meant to avoid situations in which people who are likely to find jobs within a short time horizon anyway fill training places that should rather be occupied by jobseekers who really need them.

The final step under the first pillar, as categorized by Jacobi and Kluve (2006), was the requirement that professional evaluators monitor the effectiveness of the reforms. This means that about 20 research institutes with about 100 experts closely follow the reforms. We will come to the theoretical foundations for the evaluation of the reforms later.

The second pillar of the reforms may be considered as the central component of the whole project. The notion that summarises the idea behind it is "fördern und fordern", which means that jobseekers receive support, but conversely the BA requires them to actively participate in the assistance process. This activation of jobseekers means that the whole benefit system changes in order to put jobseekers under more pressure. Before the reforms, jobseekers who had had a regular job were entitled to receive *Arbeitslosengeld* [unemployment benefit], which is part of the social insurance system and earnings-related, for an initial time period, and then they were still eligible for *Arbeitslosenhilfe* [unemployment aid], a tax-financed benefit. This second scheme was less generous than the first one, but still related to prior earnings. Jobseekers obtained entitlement to unemployment aid for time periods of one year, but it was usually extended without a maximum spell. The Hartz reforms abolished the unemployment aid scheme, which means that unemployed persons are only eligible for Arbeitslosengeld II (ALG II) [unemployment benefit II] after the expiration of Arbeitslosengeld I (ALG I) [unemployment benefit I]. ALG II is means-tested and is reduced to the level of social assistance. All able-bodied individuals are merged into one category and obtain a tax-financed benefit that is unrelated to prior earnings. Those people who had regular jobs subject to social insurance contribution first obtain ALG I whereas jobseekers without social insurance contribution directly move into ALG II. If the authorities have the impression that the recipient does not actively contribute to efforts to make future employment more probable (e.g. writing applications, participating in training or workfare), the BA is allowed to reduce benefit payments. This stronger pressure on the unemployed had the purpose to increase labour supply, accelerate the filling of vacancies and to reduce the abuse of the welfare system by individuals who are able but unwilling to make a living by working. On the other hand a side-effect of the reforms is the growing fear to fall through the welfare net, even for the middle classes who were – before Hartz – less prone to concerns about their financial well-being in the future. This issue is important to take into account when talking about German society due to its stronger emphasis on security than Anglo-Saxon cultures where risk aversion seems to be less pronounced. In spite of its broad coverage in the media this question is beyond the scope of this paper.

Other measures under the second pillar are the extended possibility for jobseekers to obtain start-up subsidies on the basis of a business plan approved by the BA, which then leads to the creation of an *Ich-AG* [Me, Inc.], or an extension of wage subsidies to employers under the condition they employ workers who are considered as hard to place on the regular labour market, like older people. In order to increase incentives to take up jobs for earners with only low incomes, the government decided to introduce so-called Midijobs. This means that people who earn between 400 and 800 Euros per month receive social security subsidies in such a form that the amount of the support decreases with

increases in earnings. Once an employee reaches the threshold of 800 Euros the subsidy is phased out (Jacobi and Kluve 2006).

The third pillar of changes constitutes comprehensive labour market deregulation. This refers to the rules related to temporary work, fixed-term contracts, and wage setting. It is safe to say that the changes related to this category are the least pronounced in the context of the *Hartz reforms*. Wage setting has remained relatively centralised in the Federal Republic and regulation of temporary work had already gradually become less intense in the years before Hartz. In the domain of dismissal law not so much has changed. After the reform also people who are older than 52 but younger than 58 years have the possibility to get their fixed contracts renewed repeatedly, whereas this was only possible for employees older than 58 years before. Furthermore exemptions from dismissal law only applied to companies with less than ten employees before in contrast to only five after the big changes (ibid.).

Counterfactual Questions in Economics

Counterfactual questions about the economy not only occupy economists but can also be heard from laymen in everyday life. The current financial crisis with the resulting abundance of press articles, TV programmes and discussions among the general public about the 'right' economic policy clearly illustrates the importance of counterfactual thinking. We can often hear questions like the following: What would have happened if there had been stricter regulation of financial products? Could the crisis have been avoided if rating agencies had more sophisticated methods for evaluating companies' financial performance? Would it have been possible to avoid the spread of the crisis if the US government had not decided to let a big financial-services firm go bankrupt?

In spite of the importance and topical interest in such questions related to the world of finance, the purpose of this article is to look more closely at the welfare state and its performance. Although in the last months, welfare state discussions did not receive as much public attention as the financial crisis, the topic is still relevant to researchers as well as the general public. Given the worries about unemployment, low real wages and the ageing of many Western economies, the right remedies against these phenomena are searched and widely debated. Taking into account the large amount of resources spent on the alleviation of risks like unemployment, illness and old-age, we first want to know if a specific programme has in fact been effective. Has it led to the intended effect? If a welfare programmed aimed at reducing child poverty by introducing in-kind benefits, like the provision of free school lunches has contributed to a lower overall percentage of poor children and a lower depth of poverty, then we may be safe to reject the hypothesis that this policy was not effective.

Nevertheless, effectiveness is not a sufficient condition for success. In a second step we want to investigate if the programme under scrutiny has been efficient. In order to answer this question the evaluator tries to find out if the benefit that is attributed to the programme is big enough in relation to the costs involved. Even if we are sure that the scheme has contributed to a reduction in child poverty, it may be the case that the financial outlays, i.e. government spending, have been very substantial, but the reduction in poverty only moderate. It could be the case that the same expenditure elsewhere would have had a more beneficial impact.

Cowan and Foray (2002) emphasise the importance of counterfactual economic history for the analysis of the evolution of economic phenomena. They claim that a large number of economists still regard explicit counterfactual reasoning with suspicion although nowadays there is an increasing amount of economic models which are based on *multiple equilibria*. Conventional models that had unique outcomes were more straightforward to anlayse concerning their predicted outcomes, whereas in situations with a large amount of possible equilibria questions about the 'best' outcome become more important. If only one equilibrium is considered the researchers slightly change some exogenous parameters and try to see which of the outcomes is the most preferred one. In a situation with more equilibria this is more difficult because not only small parameter changes have to be taken into account, but also decisions taken much earlier, which have led on a unique path to a certain equilibrium. In other words, the researcher has to consider questions of *path-dependency*. This means that the economist searches for points in the past when decision were taken which ex-post have led to a course of events that make a convergence to different outcome at later stages highly unlikely. Would it be more efficient if an alternative to the now common QWERTY-keyboard on typewriters had become the standard? Would this allow faster typing? (David 1985) This is in accordance with what Elster (1978) calls a *branching view* of history. Each point in history where a certain decision is taken is the beginning of a new branch, which in turn will split up into several boughs when new decisions are taken. This tree paradigm is in contrast to an alternative, less historical approach in which a world is modelled which is almost the same as the actual world. Lewis (1973a and 1973b) calls it a *possible world* that runs parallel to the actual world and teaches us something about reality. The crucial step is to model a world that is different from the real world (and therefore counterfactual), but "only to minimum extent" (Kluve 2004:93). This then facilitates the creation of an evaluation with appropriate counterfactuals.

Although at first sight it does not seem to be surprising that the performance of the welfare state is evaluated, the close collaboration between economists (as evaluators) and politicians (as the persons who decide on the introduction and specific format of welfare programmes) does not have a long tradition in Europe. In the United States such a pooling of efforts between science and

administration had already begun earlier. In Europe – not so long ago – large amounts of money were spent simply in the hope that the programmes must somehow make sense. The reasons why a change in the willingness to monitor welfare programmes has occurred are diverse. The steady increase in unemployment in (continental) Europe over time may have contributed to an increasing scepticism towards conventional active labour market policy. It may be the case that the situation was considered to be so alarming that scepticism towards evaluations was not tenable any more. Secondly high public debt and fiscal limits imposed by the European Monetary Union have probably contributed to a break-up of opposition to scientific investigations. Thirdly the growing importance may be explained by a wave from the United States with the common hand-in-hand evaluations of public policy. A reason may be that American politicians and economists have always been more sceptical towards state intervention into the economy and therefore a stronger eagerness to monitor welfare policy is the result. In the case of Germany, the so-called Hartz reforms were the first big attempt to systematically evaluate the performance of the welfare state. The introduction of the *Dritte Buch Sozialgesetzbuch* (SGB III) [Third code of social law] in 1998, which lays the foundations for active measures on the labour market, demands related policy to be systematically evaluated. Concerning the contents and the orientation of SGB III in comparison to its predecessor, the Arbeitsförderungsgesetz (AFG) [labour promotion law], the new code has a stronger focus on groups whose labour market status is especially difficult, like older people, long-term unemployed and lowly qualified persons (Fitzenberger and Hujer 2001). The Hartz reforms as the most intensive set of changes were then the first big application of the new monitoring requirements.

Before the introduction of Hartz there were in fact economists who evaluated the performance of elements of the welfare state, but as a consequence of the relatively widespread reluctance among policymakers towards such investigation there was not enough data for robust and strictly significant monitoring of the German welfare state. These studies hinted at possible problems and were able to push researchers' interests into certain directions, but in spite of the big efforts the lack of a crucial amount of reliable statistical material reduced the precision and explanatory power of the studies. The relatively recent abundance of statistical data has sparked researchers' interest in the evaluation of the German welfare state and has already contributed to a more precise body of knowledge. We will come to the results later and first address the difficulties of welfare state evaluations and the methods that economists apply in order to find answers.

Evaluation of Welfare State Performance: Difficulties and Solutions

Fertig and Kluve (2004) set up a framework that lays the theoretical foundations of the Hartz reforms' evaluation. Their work also elaborates on general problems of welfare state analyses, but also takes the particularities of the German situation into account. According to both scholars there are two conceptual problems associated with the comprehensive labour market and welfare state reforms. The first difficulty is grounded on the fact that changes in the administration of unemployment services provided by the BA and specific labour market policies occurred simultaneously. If caseworkers have more discretion in applying the appropriate policy mix, it is likely to affect the performance of a specific new or modified programme too. Therefore the scholars propose to include control variables that take the quality of the BA's services into account. The second difficulty has to do with the problem that some of the new measures are open to all members of the labour force equally (e.g the possibility to take up a Midi-Job). There is no appropriate counterfactual case because there are no people who are unaffected by the reform (ibid.).

Concerning the steps needed for a careful evaluation of changes like the ones caused by Hartz, it is important to analyse a programme's effectiveness, its efficiency and the reasons for these criteria. Effectiveness refers to the question whether the scheme under investigation has indeed led to the intended target. To give an example: 'Has the provision of placement vouchers contributed to a quicker matching of jobseekers with vacancies? How are in contrast to this the job finding results of those who did not get such a possibility?' are questions that aim at the effect of a scheme. The methods mentioned in the following subchapter take into account the effectiveness of welfare programmes, mainly on a micro level, which is less complex to evaluate than the overall macro effects. The efficiency of a programme under scrutiny addresses the question if the benefits (e.g. the placement success due to the voucher) are bigger than the costs involved. Most taxpayers would be reluctant to a scheme that requires substantial monetary outlays, but only yields minor benefits. Finding out the costs involved is a tricky endeavour, because there are not only direct costs involved, but also indirect pecuniary effects that researchers should – ideally – incorporate into their analyses (ibid.). Opportunity costs are an important keyword in this context. Money we spend on a welfare measure cannot be used for alternative uses any more. Even if a specific programme turns out to be cost efficient at first sight (direct benefit > direct costs) the money could have created an even more positive impact elsewhere in the economy (e.g. the stimulus caused by a tax cut). We will come back to such indirect effects when talking about macroeconomic analyses later in the following subchapter.

For Fertig and Kluve (2004) a third step in the evaluation of the Hartz reforms involves an investigation of the implementation and the process. The purpose of this task is to find the causes of the effectiveness and efficiency (or the lack of these first two criteria of success). This final step is less technical, relies mainly on qualitative data and considers the context of the reform. It may be the case that researchers survey BA staff or jobseekers in order to get an idea about the precise implementation of the policy (ibid.).

Microeconomic and Macroeconomic Analyses

A precise presentation of the methodologies that labour economists apply in their analysis of questions related to work and welfare would be too technical and beyond the scope of this article. Alternatively some of the methods used by professional evaluators will briefly be described in order to provide a general understanding about the logic behind such mechanisms. I will deal with economic experiments, matching and difference-in-differences estimations, methods which are according to Fertig and Kluve (2004) the most prominent ones for the evaluation of Hartz. Readers who are more interested in the precise empirical way of investigation applied in Labour Econonomics may refer to Angrist and Krueger (1999), Calmfors (1994) or Heckman, LaLonde and Smith (1999).

The microeconomic dimension of evaluations takes into account the impacts of a certain policy on those who are directly involved. When we analyse the impact of e.g. training vouchers on unemployment duration we only consider the effect on the unemployed. We are investigating partial equilibria, which means that we ignore effects on the overall economy, like the adverse effects of higher taxes that are needed due to the introduction of the programme. Microeconomic analyses are more common than those that deal with macro effects (Hujer and Caliendo 2000).

The big problem with such (microeconomic) evaluations is that they are inherently counterfactual. In the ideal case we would be able to analyse an individual's behaviour for the case that he or she is affected by the programme under investigation, and at the same time in the counterfactual scenario in which the person would be unaffected. For instance we would be able to observe an unemployed person's search behaviour, success in finding a job and the development of his or her earnings over time for the situation in which a certain policy – say the *Hartz reform* – has been introduced and for the counterfactual situation in which there had been no legislative and administrative changes. Similar problems arise for the related counterfactual question in which we assume the policy change to have occurred, but we alter the question about individual participation in a measure. Let us take the example of a long-term unemployed who due to the Hartz changes is eligible for obtaining a voucher that entitles him to placement services by a private company. Provided

that he or she obtains such a voucher and a private company tries to find an appropriate vacancy for this person, it is impossible to say what would have happened if he had not received such a voucher. This could be the case because the caseworker in charge did not consider this step necessary or was more inclined to take other measures.

These are examples of situations in which we are interested in potential outcomes that are important when trying to investigate the results of the welfare state. Even though they are difficult to find and intrinsically hypothetical they are the cornerstones of meaningful welfare state evaluations. This problem is referred to as the *potential outcome approach*, which is sometimes also called *Roy-Rubin model* (Hujer and Caliendo 2000).

The precise technical details of the model are beyond our scope here, but I will explain the general idea behind it. In this model there are individuals, choices (either being treated or not) and results. The choice of the appropriate outcome depends on the purpose of the evaluation. It may be the duration of further unemployment, the income in the year following the intervention, or any other result. An evaluation may even be positive for a certain programme when considering one outcome, but would be negative when looking at another one. Decreasing unemployment benefits may have a positive impact on the chances of finding a job, but a negative impact on earnings.

For every individual there is the outcome for the case that he or she is under treatment (Y^T) and the outcome for the case of not being treated, thus being in the control group (Y^C).

The treatment effect of a specific programme is therefore the difference between Y^T and Y^C. As mentioned before, the difficulty is that we cannot find out the counterfactual outcome for each individual. A solution to this problem is to concentrate on mean outcome differences between groups. In order to illustrate this point let us consider two groups of people. The members of the first group each receive a training voucher (treatment group) whereas the members of the second do not (control group). If then the average duration of time until the members of the treatment group find jobs after the training is 8 weeks in comparison to 13 weeks for members of the control group, we can attribute this difference to the participation in the training scheme. Then of course we have to assume that the members of the two groups do not differ in any way apart from the fact that some have participated in the education programme and others not (*ceteris paribus*). In reality this assumption does not always hold. It may be that the caseworker in charge of the decision whether to assign a training voucher only lets jobseekers with certain characteristics participate in the programme. If he or she thinks that especially the long term unemployed with low qualifications deserve it, and thus assigns the voucher only to them, this group is put at an advantage. Conversely if the caseworker believes that such people are lost causes anyway, he or she may privilege people with relatively promising CVs.

The consequence is that the evaluation results of the training programmes' effectiveness are strongly biased. For the first case it is likely that the impact is understated whereas in the latter case it is overstated.

This discretion, which leads to an unequal treatment and violation of the above assumption, can – theoretically – still be accounted for, e.g. by interviewing caseworkers about their criteria. These differences between individuals are observable, at least to the caseworker in charge. However many other characteristics are unobservable. How can we accurately measure effort, a factor that is definitely crucial to successful placement but unevenly distributed among jobseekers in treatment and in the control group?

The best solution to the above mentioned problems of unobserved differences would be a random assignment to the treatment and the control group. This is then a natural experiment in which it does not matter that there are also unobservable or unobserved differences between the groups. Due to the group assignment by chance these discrepancies will even out if there are a sufficiently large number of people involved in the experiment. In the ideal case with let us say 100,000 jobseekers which are assigned to treatment and control group by chance, it is safe to assume that unobserved differences will move into both groups to the same extent. As an example, there will be highly motivated individuals as well as lethargic people in both groups, and therefore the extent of potential bias due to unobserved characteristics is reduced in such experimental settings.

Many characteristics which are theoretically observable are in reality unobserved because the researcher does not take them into account – there are indefinitely many characteristics and the more we incorporate into an analysis the more difficult it becomes. Alternatively the evaluator does not account for them because they are unobservable, or at least difficult to observe. Economic experiments are so to say the 'gold standard' of evaluation due to their strengths in dealing with unobserved characteristics. The large number of people in the treatment group as well as in the control group makes it possible to attribute differences in the outcome variable (e.g. the duration of unemployment) to the treatment under investigation. The researcher can be relatively sure that that due to random assignment the effect of the policy measure becomes clear. We may say that the average differences in outcomes between the treatment and control group are an approximation of the unobservable treatment on the individual. This allows estimating the counterfactual scenario of what would have happened if an individual had participated in a measure (if he/she had not in reality) or vice versa. Even if the outcome of the experiment can only take average effects into consideration, this may be precise enough for a researcher who wants to investigate the aggregate effects of a policy while ignoring the question what would have happened to each individual. It is safe to say that such experimental investigation settings mirror the methodology of the natural sciences. Social experiments are especially popular in the United States,

whereas in Europe and especially Germany they are less common. Objections towards such social experiments may be on ethical grounds (Hujer and Caliendo 2000): Why does my neighbour obtain more unemployment benefit than I do? The fact of merely being in the control group with more or less support may not sound convincing to those who are affected. It is not clear if it is only due to ethical considerations that experiments are less pronounced in Europe than in the US, but due to the strong position of American researchers in the economic community social experiments will perhaps, with the passage of time, become increasingly popular and widespread in Europe (Heckman et al. 1999).

If natural experiments with random assignment are not possible, evaluators have to find alternative ways to answer their questions. A possibility that is often applied is matching. This means that the evaluator tries to find an unaffected 'twin' for each person who benefits from the programme under investigation. In order to give an example, let us refer to our job placement voucher described earlier. Assignment did not occur randomly but due to the caseworker's discretion. Now the researcher tries to find people who share exactly the same characteristics (e.g. age, level and kind of education, family status) apart from the fact that one 'twin' is in the group that was eligible for the voucher and the other 'twin' not. As both individuals share all observed characteristics apart from programme participation the difference in the outcome variable – say the time period until successful placement – can be attributed to the programme. Fertig and Kluve (2004) describe matching as a strategy to approximate the context of a natural experiment after the programme took place. However, also this strategy is not unproblematic. Consider the case of a smaller programme with an insufficiently large amount of participants. How is it possible to find a person that is almost identical? Similar problems arise in the case that we have detailed information about participants' and non-participants' biographies, which is also positive because thorough evaluations will try to find out lots of potential explanatory variables; but the more we know the harder it will be to find a 'twin' who shares all the characteristics (Fitzenberger and Hujer 2001). There are more sophisticated ways or matching that try to avoid the problems described above. The works by Rosenbaum and Rubin (1983) and Heckmann et al. (1999) have contributed to an advancement of a body of knowledge concerning the solution of problems related with matching (Fitzenberger and Hujer 2001). Such matching approaches can, like in the case of experiments, only try to find out average effects of a policy on the group of participants and non-participants and not say with precision what would have happened on the individual level in the counterfactual scenario. For most of the cases this should not be a problem, because the overall effect has a much bigger impact on the economy as a whole. This means that, although this question is relevant to the specific person, it is impossible to know what would have happened to single individual in the counterfactual case.

The third method we will highlight in this paper is the difference-in-differences (did) methodology. It is an extended version of a simple before-after analysis. The development of the outcome variable for a group of persons is compared to the fate of a second group that is expected to be unaffected by the policy change but still shares characteristics of the people from the group that is affected. The example described below can illustrate this point. It is an abridged and adapted form of did-estimation by Eissa and Liebman (1996) who investigated the impact of a tax reform act of 1986 in the USA (TRA-86) on labour force participation of single mothers. As we can see in Table 1, labour force participation of unmarried mothers with less than high school education (treatment group) increases by 1.8 percentage points (column 3, which is column 2 minus column 1) whereas for the first control group (same category of women who have no children) it fell by 2.3 percentage points. The policy change had the purpose to increase the number of lowly qualified single mothers who are working. Therefore the researches set their development in relation to the control group who is not especially targeted by the reform. The fall in labour force participation of lowly qualified childless women even increases the impact of the policy change: The did-estimator amounts to 4.1 percentage points. If we set the development of the target group in relation to more qualified unmarried women with children (control group 2) the impact of the reform is still positive, but the impact is smaller because also the second control group's labour market situation improved. The did-estimator for the comparison with the second control group only amounts to 1.8 percentage points (1.8 minus 0.9).

Having taken into account the microeconomic effects that researchers consider when carrying out evaluations this paper will shortly highlight some points that should be incorporated in macroeconomic evaluations. These relate not only to the relatively small group of people that the welfare programme under investigation seeks to target. They refer to the economy as a whole and must also investigate a large number of indirect effects. Fertig and Kluve (2004) as well as Fitzenberger and Hujer (2001) refer primarily to the leading contribution from Calmfors (1994) who makes a distinction between displacements effects, substitution effects, deadweight loss and tax effects. Displacement effects refer to the fact that some participants of a programme take up jobs that non-participants would otherwise have obtained. Substitution effects have to do with changes in relative wages induced by a programme which in turn reduces the demand for other types of non-participants. Deadweight losses have to do with results that would also have been achieved if there had been no specific programme. If, for example an employer would have hired a person even in the absence of a wage subsidy system, the subsidy turns out to be a redistribution of wealth from the state – or taxpayer – to the company that employs the person. Tax effects take into account that labour market policy has to be financed by taxation with adverse effects on the economy. A policy that

on the one hand increases the number of jobs by certain incentives to employers or workers, but on the other hand increases taxes and significantly reduces work incentives for all non beneficiaries is likely to be a failure on the macro level. Another crucial question in a thorough macroeconomic evaluation is the impact of reduced costs and higher tax revenues that are caused by a reduction in unemployment. As mentioned before, evaluations that take into account effects on the economy as a whole are more complex than microeconomic analyses which are, as we have seen above, already difficult to pursue.

As we have seen, the methods highlighted in this chapter try to find out the average effect of certain welfare state instruments on people who are affected and on people who are unaffected. These approximations are important to researchers and policymakers who want to find out whether a certain programme has been effective and efficient, because they are to the largest extent concerned with the overall effects and not so much about the counterfactual impact on an individual. If the target is to increase employment among single mothers, their overall employment ratio is of importance and not what happens to a specific mother. This may sound rude, but attempts to make causal claims on the individual level would be flawed and untrustworthy.

Table 1: TRA-86 Labour Force participation (in %)

Group	Before (1)	After (2)	Difference (3) (2)-(1)	Difference-in differences (4)
Treatment: Unmarried women, <12 years of schooling, with children	47.9	49.7	1.8	
Control 1: Unmarried women <12 years of schooling, without children	78.4	76.1	-2.3	4.1
Control 2: Unmarried women, >12 years of schooling, with children	91.1	92.0	0.9	0.9

Source: Eissa and Liebman (1996); Simplified and abridged version

Evaluation Results of the Hartz Reforms

A thorough presentation of the impacts caused by the set of comprehensive labour market and welfare state reorientation would be beyond the scope of this paper. Nevertheless we will have a brief look at some findings related to the Hartz reforms. However we should keep in mind that there may be long-term consequences that cannot be foreseen now. Furthermore every evaluation must be clear about the counterfactual scenario to which the real world is compared. In the case of the Hartz reforms a natural candidate would be the situation in which legislation did remain the same as before the reforms. This allows investigating the specific changes brought by Hartz and thus attributes them to the new rules. But of course there may also be other counterfactuals that could, at least under different circumstances, make perfect sense.

Jacobi and Kluve (2006) summarise a large amount of evaluations that have been carried out. For the introduction of placement vouchers that jobseekers can use for being placed by a private firm, an example mentioned before in this paper, the results seem to be poor. Many jobseekers did simply not use these vouchers. Additionally those who did and found a job were employed for a shorter time horizon than jobseekers who did not use placement vouchers (ibid.). Results for training programmes are not unambiguous, but it seems that due to shorter duration of programmes, assignment to training after shorter durations of joblessness and more competition among training providers, post-Hartz training has become a bit more cost-efficient, although the net results are still negative. Additionally locking-in effects seem to be less dramatic than before.

Jacobi and Kluve (2006) consider wage subsidies to employers as well as start-up subsidies as positive elements of the reform because the first increase the chances of employment and the latter have reduced the risk of unemployment. Minijobs have contributed to the creation of a large amount of jobs whereas Midijobs have only modestly been taken up. In the case of Midijobs displacement effects cannot be rules out.

Temporary work deregulation has also increased the number of employees in such jobs. However, we should take into account that temporary workers are at the same time those who lose their jobs first in times of recession, a phenomenon that has become apparent since the autumn of 2008. The decision to make fixed-term contacts renewable more easily for older employees does not seem to have had a significant impact (ibid.).

Concluding Remarks

We have seen that trying to evaluate the performance of welfare states is an important but difficult task. Therefore efforts to analyse social policy are becoming increasingly numerous. This is especially the case for Germany where

thorough scientific investigation about the effectiveness and efficiency of labour market policy has become imperative since the introduction of SGB III and the Hartz laws. The latter laws had the purpose to improve placement and training services and to activate the unemployed, which means that they are under a stronger financial pressure to take up jobs. Furthermore some deregulation of the labour market occurred. Conducting economic experiments would in many cased be the ideal method of evaluation, but there are ways to approximate the conditions of social experiments.

Nevertheless we may not be able to find out the effects of a certain policy on a single individual because the counterfactual situation is not observable. Instead we consider average effects on a larger group of comparable people in order to make strong predictions. This may give us the feeling of what might have happened on the individual level, but this question remains highly speculative and is beyond the scope of serious analyses.

For thorough evaluation efforts it is important to keep in mind effects on the micro as well as on the macro level, which is more difficult. Keeping in mind all the difficulties and problems of such analyses we should interpret all results with care. This of course also holds for the impact analyses about the Hartz reforms. Even if an evaluation seems to have a positive effect on the outcome variable under investigation, deciding to focus on another outcome criterion may lead to different results. Even if a policy increases employment this does not rule out negative effects on earnings or outcome variables that are not taken into account by economists. It remains to be seen what future evaluations will find out, which methods will emerge and which outcome variables they will take into account.

References

Angrist, Joshua D. and Krueger, Alan B. 1999. "Empirical Strategies in Labor Economics", in: in: *Handbook of Labor Economics* Vol III, ed. By Ashenfelter, O. and Card, D. Amsterdam: Elsevier.

Caliendo, Marco and Steiner, Viktor. 2005. "Aktive Arbeitsmarktpolitik in Deutschland: Bestandsaufnahme und Bewertung der mikroökonomischen Evaluationsergebnisse." Zeitschrift für ArbeitsmarktForschung – Journal of Labour Market Research 38, 2/3, 386-418.

Calmfors, Lars. 1994. "Active Labour Market Policy and unemployment – a framework for the analysis of crucial design features". OECD Economic studies No. 22, Spring 1994. http://www.oecd.org/document/52/0,3343,en_2649_-34117_33840116_1_1_1_1,00.html (accessed 10 November 2008).

Cowan, Robin and Foray, Dominique. 2002. "Evolutionary Economics and the counterfactual threat: on the nature and role of counterfactual history as an empirical tool in economics" In *Journal of Evolutionary Economics*, 12: 539-562.

David, Paul A. 1985. "Clio and the Economics of QWERTY" In *The American Economic Review*, 75: 332-337.

Elster, Jon. 1978. “Logic and society: contradictions and possible worlds.” Toronto: Wiley.

Engels, Dietrich. 2001. “Abstand zwischen Sozialhilfe und anderen Arbeitseinkommen: Neue Ergebnisse zu einer alten Kontroverse”. Sozialer Fortschritt 3/2001.

Fertig, Michael and Kluve, Jochen. 2004. “A conceptual framework for the evaluation of comprehensive labor policy reforms in Germany”. Bonn: Institute for the Study of Labor. Discussion Paper No. 1099. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=527123 (accessed 27 October 2008).

Eissa, Nada and Liebman, Jeffrey B. 1996. “Labour Supply Response to the Earned Income Tax Credit”. In *Quarterly Journal of Economics*, 111(2):605-37.

Fitzenberger, Bernd and Hujer, Reinhard. 2001. “Stand und Perspektiven der Evaluation der aktiven Arbeitsmarktpolitik in Deutschland”. Centre for European Economic Research. Discussion Paper No. 02-13. ftp://ftp.zew.de/pub/zew-docs/dp/dp0213.pdf (accessed 15 October 2008).

Hagen, Tobias und Steiner, Viktor. 2000. Von der Finanzierung der Arbeitslosigkeit zur Förderung von Arbeit – Analysen und Empfehlungen zur Arbeitsmarktpolitik in Deutschland. Baden-Baden: Nomos Verlagsgesellschaft.

Heckman, James J., Lalonde, Robert J. and Smith, Jeffrey A. 1999. “The Economics and Econometrics of active labour market programmes,” in: *Handbook of Labor Economics* Vol III, ed. By Ashenfelter, O. and Card, D. Amsterdam: Elsevier.

Hujer, Reinhard and Caliendo, Marco. 2000. “Evaluation of Active Labour Market Policy: Methodological Concepts and Empirical Estimates”. Bonn: Institute for the Study of Labor. Discussion Paper No. 2100. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=264805 (accessed 20 October 2008).

Jacobi, Lena and Kluve, Jochen. 2006.”Before and After the Hartz Reforms: The Performance of Active Labour Market Policy in Germany”. Bonn: Institute for the Study of Labor. Discussion Paper No. 2100. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=900374 (accessed 27 October 2008).

Kluve, Jochen. 2004. “On the use of Counterfactuals in Inferring Causal Effects”. In *Foundations of Science* 9:65-101.

Lewis, David. 1973a. “Causation”. In *Journal of Philosophy* 70: 556-567.

Lewis, David. 1973b. “Counterfactuals”. Cambridge, MA: Havard University Press.

Rosenbaum, Paul R. And Rubin, Donald B. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects”. Biometrica 70: 41-55.

Wunsch, Conny. 2005. “Labour Market Policy in Germany: Institutions, Instruments and Reforms since Unification”. Discussion Paper 2005-206. Universität St. Gallen. http://ideas.repec.org/p/usg/dp2005/2005-06.html (accessed 25 October 2008).