CAUSAL EFFECTS OF EUROPEAN ACTIVE LABOR MARKET POLICY – FOUNDATIONS AND EMPIRICS

Inaugural-Dissertation
zur Erlangung
der Würde eines Doktors der Wirtschaftswissenschaften
der Wirtschaftswissenschaftlichen Fakultät
der Ruprecht-Karls-Universität Heidelberg

vorgelegt von
Jochen Kluve
aus Eberbach

Heidelberg, Juli 2001

Contents

Chapter 1

Introduction and Overview

Over the last one or two decades most European countries have been subject to a strong and persistent increase in their unemployment rates. In fact, the combating of obdurately high unemployment has become one of the most urgent issues of economic policy. This phenomenon has disturbed European economies mainly since the late 1980s or even early 1990s – a fact that on the one hand results from a strong economic performance of virtually all Western European economies throughout the decades succeeding World War II. On the other hand, the countries of Central and Eastern Europe (CEE), too, have had to face large-scale *open* unemployment only since the beginning of the 1990s, the time when they embarked upon the transition process from a formerly socialist state to a market economy.

The pertinacious occurrence of high unemployment has far-reaching consequences on labor markets and demands adequate policy reaction. A straightforward implication is the need to adapt – if not restructure – social insurance and unemployment benefit systems. Such an undertaking follows from the mere increase in unemployment compensation claimants. It affects what is commonly referred to as the *passive* side of labor market policy – i.e. unemployment insurance regulations. Besides merely administering unemployment, however, most European countries have turned to attempt to *actively* combat unemployment by exercising measures of so-called *Active Labor Market Policy (ALMP).* In very general terms, ALMP measures epitomize the deliberate effort at acting against unemployment by trying to improve the chances of labor market success for the unemployed.

These active measures can be coarsely classified into three types of labor market programs. First, *training programs*, such as classroom training, on-the-job training, work

experience, or job search assistance. Second, *wage subsidies* to the private sector, i.e. subsidies to employers or financial incentives to workers. Third, *direct job creation* in the public sector. Clearly, ALMPs are aimed at combating structural unemployment. To give but the most general grounds for their introduction, the original policy objectives center around the classic efficiency and equity arguments: ALMPs are meant to (a) develop human resources and adjust manpower resources with a view to fostering economic growth, and (b) to enhance both employability of and opportunities for disadvantaged groups, thus contributing to social equity (OECD 1990).

One elucidating example of this newly emerged importance of employment issues within economic policy is the response of European Union (EU) member states to the alarmingly high unemployment rates across Europe. In June 1997, the EU member states agreed on the *Amsterdam Treaty.* The agreement included a new title on employment, which for the first time explicitly recognized the fact that employment issues have a status equal to that of other key aspects of EU economic policy. This marked the beginning of the European Employment Strategy, a strategy that was further elaborated during the subsequent Luxembourg Jobs summit in November 1997. Hence, this concerted strategic effort has become known as the *Luxembourg Process*. The Luxembourg Process aims at jointly assessing European employment policy issues by virtue of annual *National Action Plans (NAPs)* and *Joint Employment Reports (JERs).*

The principal idea of the Luxembourg Process is straightforward. EU member states want to engage in a joint employment policy. Therefore they declare an annual set of employment policy guidelines that each country has to translate into an appropriately formulated NAP. On the basis of both the NAPs and the actual economic development throughout the year, the annual JER reviews each country's performance and tries to give recommendations for the upcoming set of employment guidelines. The guidelines are then adapted and reformulated for the following year, resulting in adjusted NAPs, etc.

Nowadays measures of ALMP constitute a large proportion of each country's set of employment policies. ALMPs are the predominant means of fighting unemployment. Hence, they are included in the National Actions Plans and play a major role in the effort of the Luxembourg Process to implement a joint employment policy successfully addressing the unemployment problem. For achieving this, however, the Luxembourg Process (by means of the JERs) needs to identify effective policy instruments and examples of good practice across countries. But how can we know that a policy instrument is effective? This clearly implies the

necessity to evaluate any such program, in order to identify whether or not there is a positive causal effect of the labor market program on the desired outcome measure. But whereas the general objective of identifying "examples of good practice" itself along with the feedback structure of the NAPs and JER is desirable, even the 1998 JER has to admit that "[s]ystematic evaluation of employment and labour market policies is still not common practice in many Member States" (European Commission 1998). There is thus an apparent necessity for European economic policy to learn more about the impact of active policy interventions on the labor market.

This thesis assesses causal effects of European Active Labor Market Policy. It does so from a threefold perspective. First, I will give a thorough discussion of methodological issues that arise in the venture of causal inference. This foundational analysis includes an overview of the predominant procedures to model causation in the empirical sciences. It subsequently focuses on a particular statistical model – that is of prevalent use in evaluation research – and the causal queries that can be asked, and answered, within the model. Second, I place the undertaking of combating unemployment by means of ALMP in a European context. Following from disillusioning diagnostics on the state of European labor markets, I show that a large variety of active labor market programs has been implemented across countries in both Western and Eastern Europe. The analysis proceeds to investigate the findings of scientific evaluation research on the effectiveness of these measures, and concludes with a set of implications for economic policy derived from these results. Third, I will present two detailed country studies on ALMP effectiveness. These studies analyze the causal effects of three labor market programs in Poland using the statistical method of matching, a specific nonexperimental variant of the causal model laid out in the first part.

European governments employ ALMP to fight unemployment. Failure or success of this endeavor can only be judged by an evaluation of ALMP effectiveness. Determining effectiveness implies establishing a causal relation between the potential cause – the labor market program – and the presumed effect – on some appropriate response variable indicating labor market success. Thus, the enterprise of causal inference constitutes the core of the evaluation problem. Causal inference, however, has once been argued by Philip Dawid to be "one of the most important, most subtle, and most neglected of all the problems of Statistics" (Dawid 1979). Chapter 2 of this thesis reviews contemporary approaches at modeling causation in the empirical sciences. In fact, recent years have seen an increased discussion of causation and causal models in econometrics, statistics, sociology, computer science, and

epidemiology.

Looking at how causation is modeled in the empirical sciences, Chapter 2 finds that there are three major approaches: (a) Structural Equation Models, (b) Potential Outcome Models, and (c) Directed Acyclic Graphs. *Structural Equation Models (SEM)* for causal inference are mainly used in economics and the social sciences. SEM originated from path analysis developed by geneticists in the early 20th century (Wright 1921, 1934). Pioneering work in SEM was done above all by Haavelmo (1943, 1944) and Koopmans and Hood (1953). This work shaped the program of the Cowles Commission and had a decisive impact on the development of modern econometrics. In fact SEM has remained the paradigm of causal modeling in econometrics and the social and behavioral sciences (cf. Morgan 1990, Heckman 2000).

The *Potential Outcome Model (POM)* is the causal model predominant in statistics. In the POM, units are potentially exposed to a set of treatments, and have corresponding responses associated with each treatment. The causal relation of interest is the effect on the outcome variable of some particular treatment relative to some other particular treatment – frequently called the "control" treatment. Since in reality each unit can only be exposed to one treatment, the other treatment states and associated outcomes for the single unit represent potentialities and are expressed in terms of counterfactuals. In its essence the POM dates back to the work of Neyman (1923 [1990], 1935) and Fisher (1935). While Neyman was probably the first to suggest the notion of potential responses, Fisher is commonly credited for the invention of randomized experiments. Randomized experiments constitute one possible setting under which the POM produces valid causal inference. This notion of potential outcomes was extended to observational studies by Rubin (1974, 1977). Due to his contributions the model is often referred to as the "Rubin Model". Related work in economics are the earnings model of Roy (1951), and models for switching regressions (Quandt 1958, 1972).

Directed Acyclic Graphs (DAGs) represent a rather recent alternative approach at modeling causation. As DAGs are of subordinate importance for the subject of this thesis, I will touch upon them briefly. Let me merely mention what their proponents – cf. Pearl (1995, 2000a) and Spirtes, Glymour, and Scheines (2000) – think that DAGs can contribute to causal modeling: In their view, DAGs, by virtue of a certain graphical language, manage to make causal relations and assumptions and implications more explicit than other approaches such as SEM or POM.

Starting from these foundational findings, Chapter 2 picks out the POM as the main causal model of interest in evaluation research. Looking at the model I find that it is formulated in terms of counterfactuals. What, then, are counterfactuals? This leads me to analyze the semantic properties of counterfactuals, and the counterfactual approach to causation in philosophical logic. It turns out that the central element of this approach is the notion of 'possible worlds'. I proceed to connect the counterfactual-based POM with the possible world semantics for counterfactuals, reformulating the POM and its assumptions in terms of counterfactual statements. This procedure (i) connects statistical and philosophical understandings of counterfactuals and (ii) adds clarity to the counterfactual nature of the POM. The chapter then takes a closer look at this crucial notion of proximity of possible worlds, and finds that within the POM closest possible worlds are defined *a priori*, and merely differ with respect to elements of the treatment set T along with associated outcomes. Therefore, I give a detailed discussion of T using a simple set-theoretical framework. This analysis also elucidates which causally meaningful counterfactual questions can be asked, and answered.

Having thus established the underpinnings of causal inference, Chapter 3 turns to the situation on European labor markets. The desolate economic situation in terms of high and persistent unemployment rates led European governments to introduce or enforce efforts of combating unemployment by means of Active Labor Market Policy. But even though most countries spend a considerable share of their budget on these measures, a thorough evaluation of ALMP effectiveness has remained the exception. This fact is particularly obvious from a comparison with the US, where the conscientious evaluation of policy interventions has a long tradition due to the abiding interest of both policy makers and the American public. The US holds abundant examples of evaluations of labor market programs in which the evaluation effort accompanied the program from the very first step of its implementation. This proceeding has resulted in a large body of reliable evidence on program effectiveness. Indeed, the shape of new programs has often been determined on the basis of experiences with previous programs. The close connection between implementation of programs and their evaluation has also contributed to rapid advances in research methods, as these could be directly applied.

In Europe, however, evaluation research seems to have remained in its infancy, mainly due to a disconnection of the policy effort from scientific research. On the one hand we observe policy makers react to the bleak situation on labor markets and engage in an

increasing number of active policy measures. For the EU, this operation has even taken on the official shape of the Luxembourg Process. On the other hand, European scientists have closely followed US advances in methods and developed adequate tools to answer evaluation questions with confidence. Chapter 3 sheds further light on this juxtaposed yet separate progress on both the "policy side" and the "science side". A tighter connection of the two sides would imply a great leap forward for European evaluation research and ALMP effectiveness.

Chapter 3 highlights some further differences of ALMP evaluation between the US and Europe, such as the prevalence of nonexperimental data in Europe as opposed to a substantial number of results derived from randomized experiments in the US. I then proceed to investigate a selection of European country studies in detail. Across countries, this review entails a large variety of programs implemented, and various scientific evaluation methods applied to assess their impact. Quite a substantial number of studies utilize variants of the POM delineated in Chapter 2.

Various messages for the design of economic policy can be extracted from the presented evidence. In general, estimation results indicate that treatment effects are modest at best. Training seems to be the most promising program – if there is any –, and public sector programs fare substantially worse than private sector programs. Among the detailed empirical findings at least three are particularly noteworthy. First, measures of increased individual job search assistance – such as Counseling & Monitoring in the Netherlands, or the New Deal in the UK – seem to be promising, even though a careful targeting is imperative (van der Klaauw and van den Berg 2000). Second, an innovative ALMP measure in Switzerland called "temporary wage subsidy" (Gerfin and Lechner 2000) displays large positive effects. This program encourages job seekers to accept job offers that pay less than their unemployment benefit by compensating the difference with additional payments. In this respect it would be interesting to see whether other countries would make similarly positive experiences with this program. Third, it appears to be a major distorting factor for treatment effectiveness if program participation restores benefit receipt eligibility. As a substantial number of studies provides evidence for this hypothesis, this is one of the more robust results of current evaluation research in Europe. In fact it is surprising that such regulations are still common practice in many European countries, as too generous unemployment benefit systems have frequently been identified as one labor market feature in Europe associated with high unemployment (cf. for instance Nickell 1997).

One example among OECD countries exercising such regulations is Poland. In thus conducting an empirical evaluation of Polish labor market programs, Chapters 4 and 5 of this thesis further focus the analysis of causal effects of European ALMP on a specific country. Similar to other countries of Central and Eastern Europe, Poland has experienced substantial unemployment rates only since the beginning of the transition to a market economy. In order to combat unemployment and long-term unemployment the Polish government has applied a broad menu of Active Labor Market Policies. Within this set of ALMP three programs have been of particular importance: Training, Intervention Works, and Public Works.

Training programs are meant to solve skill mismatch in the Polish labor market. Workers with redundant or no skills are trained in those occupations that are presumably characterized by strong demand of entrepreneurs from expanding sectors of the economy. Training is thus clearly aimed at increasing participants' employment probabilities by enhancing individual human capital. In its essence, *Intervention Works* is a program that gives wage subsidies corresponding to the level of unemployment benefit payments. These wage subsidies are given to firms in the private or public sector if they hire an unemployed person. Subsidies are the larger the longer the individual is kept on in the firm. *Public Works* jobs are directly created by the government, in particular by the municipalities, and are targeted mainly but not exclusively at the long-term unemployed. Many of these jobs are in construction and cleaning of public buildings, parks etc., i.e. they have a low skills content. In principle, though, both Intervention Works and Public Works have been conceived to enhance or maintain the human capital of participants.

Chapters 4 and 5 provide microeconometric evidence on the effectiveness of these three ALMP measures in terms of their treatment effects on individual participants. In assessing causal effects these chapters apply a variant of the POM. As outlined in Chapter 2, randomized experiments are one setting under which the POM produces valid inference. Since the empirical application is set in a nonexperimental context, it is shown that in such an observational study matching estimators can serve as a substitute for randomization. Thus, in chapters 4 and 5 the method of matching is used to identify the desired counterfactual.

In terms of the POM, units, treatments, and outcomes are defined as follows: The units under consideration are unemployed individuals having either participated in an ALMP program – these are the treated units – or having not participated in any such program. The latter constitute the (potential) comparison units. The treatment set comprises the three ALMP programs Training, Intervention Works, and Public Works, and a non-participation state.

Causal effects are inferred pair-wise, in turn relating each of the ALMP programs and the non-participation state. The outcome variable of interest captures post-treatment labor market success in terms of employment and unemployment rates.

Using retrospective data from the 18th wave of the *Polish Labour Force Survey (PLFS)* as of August 1996, the studies focus on a supplementary questionnaire containing individual labor market histories. The data follow individuals for a period of 56 months (January 1992 to August 1996) entailing information on their respective labor force status for every single month. This rich information is condensed to a trinomial variable of labor market outcome (employed, unemployed, out-of-the-labor-force).

In Chapter 4, treatment and comparison groups are matched over individual observable characteristics and pre-treatment labor market histories. Matching proceeds using a dynamic 'moving window' feature accounting for changing macroeconomic environment: Each treated unit is assigned a comparison unit with identical pre-treatment history from an equal phase of the transition cycle. Furthermore, observations on controls are from the same regional labor market. The study uses the trinomial labor market outcome variable described above to analyze the effect of ALMP measures on employment and unemployment rates. Exploiting the history structure I take into account short-term (9 post-treatment months) and medium-term (18 post-treatment months) effects. The matching estimator implemented is a conditional difference-in-differences estimator of treatment effects.

Findings suggest that training has a positive effect on the employment probability for both men and women. This effect is slightly more pronounced for women. Therefore, this ALMP measure clearly seems to improve the efficiency of the Polish labor market. Regarding Intervention Works there is no overall treatment effect for participating women, while I report strong negative treatment effects on the employment rates of men who took part in either Intervention Works or Public Works (Public Works for women are not being analyzed due to small sample size). As participation in any of these two ALMP measures entitles the participant to a new round of unemployment benefits, Intervention Works and Public Works seem to be a common intermediate stage between two spells of unemployment benefit receipt, where the individual entered the program after having exhausted his benefit eligibility. Hence, while stigmatization might have some role to play, chapter 4 attributes most of the negative overall treatment effects of these programs to 'benefit churning'.

While chapter 5 focuses on a similar evaluation question – the impact of Polish ALMP on employment outcomes –, it changes perspective to a more detailed account of the matching

procedure and the discussion of results. In particular, chapter 5 discusses three stages of an appropriately designed matching procedure and demonstrates how the method succeeds in balancing relevant covariates. This procedure uses three matched samples from an exact matching within calipers algorithm imposing increasingly stronger requirements. The validity of this approach is illustrated using the estimated propensity score as a summary measure of balance.

Like the previous chapter, Chapter 5 also applies the conditional difference-in-differences estimator of treatment effects based on individual trinomial sequences of pre-treatment labor market status. In this case, however, I give a detailed discussion on the importance of considering these pre-treatment histories as determinants of program participation. In assessing program impact, the focus is on the short-term response only, but instead presents a more profound account of post-treatment outcomes. Again, general findings suggest that Training raises employment probability, while Intervention Works seems to lead to a negative treatment effect for men. The in-depth discussion of results finds that appropriate subdivision of the matched sample for conditional treatment effect estimation can add considerable insight to the interpretation of results. For instance, it turns out that the overall negative impact of Intervention Works is almost exclusively due to the dismal post-treatment labor market performance of male participants, and that the full sample effects are driven by those individuals whose pre-treatment labor force status history consists of a sequence of unemployment.

In short, the remainder of this thesis is organized as follows. Chapter 2 considers the foundations of causal inference in the empirical sciences. It sketches three different ways of modeling causation, and gives an in-depth account of the one causal model predominant in evaluation research, i.e. the Potential Outcome Model. This account shows what causally meaningful counterfactual questions can be asked, and answered. Chapter 3 considers Active Labor Market Policy in a European context. I delineate the status quo of European labor markets and the resulting urge of policy makers – exemplified by the Luxembourg Process – to introduce active labor market measures to combat high and persistent unemployment. Parallel to that, scientific evaluation research has come to answer evaluation questions with confidence. I present recent state-of-the-art evidence on ALMP effectiveness derived from academic research, and argue for an enforced inclusion of such research in the political process. Chapter 4 narrows the scope on a particular OECD country and transition economy and presents an evaluation of ALMP in Poland. Applying a special variant of the causal

model outlined in Chapter 2, the analysis implements a conditional difference-in-differences estimator of treatment effects. Considering as the outcome a trinomial variable of labor market status, the main findings suggest that Training has a positive effect on employment probability for both sexes, while Intervention Works display strong negative effects for men, and zero impact for women. This finding is underpinned by the complementary analysis of chapter 5. Here I discuss three stages of an exact matching within calipers approach, and how the method succeeds in balancing relevant covariates. Further attention is given to pre-treatment labor market histories as *the* main determinants of program participation. I argue that appropriate subdivision of the matched sample for conditional treatment effect estimation can add considerable insight to the interpretation of results. Chapter 6 concludes.

Chapter 2

On the Role of Counterfactuals in Inferring Causal Effects of Treatments

Abstract. Causal inference in the empirical sciences is based on counterfactuals. This chapter presents the counterfactual account of causation in terms of Lewis's possible-world semantics, and reformulates the statistical potential outcome framework and its underlying assumptions using counterfactual conditionals. I discuss varieties of causally meaningful counterfactuals for the case of a finite number of treatments, and illustrate these using a simple set-theoretical framework. The chapter proceeds to examine proximity relations between possible worlds, and discusses implications for empirical practice.

In spite of all the evidence that life is discontinuous, a valley of rifts, and that random chance plays a great part in our fates, we go on believing in the continuity of things, in causation and meaning.

– Salman Rushdie, 'the ground beneath her feet' –

2.1 Introduction

Recent years have seen an increased discussion of causation and varieties of causal models in the fields of econometrics, statistics, computer science, epidemiology, and sociology. Leaving for the moment the century-lasting discourse on accounts of causation in philosophy aside – I will get back to this at a later stage – this increased research on matters of causation in the above-mentioned fields has led to three major approaches to modeling causation currently dominating the debate on causal inference. These are (a) Structural Equation Models (SEM), (b) Potential Outcome Models (POM), and (c) Directed Acyclic Graphs (DAG). In this section, I will first give brief introductions to all three approaches, and then discuss somewhat further how they are perceived in the academic community. Subsequently I will focus on the scope of this chapter, its foundations and contributions.

2.1.1 Modeling Causation: Three Approaches

Structural Equation Models (SEM) as an approach to causation are mainly used in economics and the social sciences. SEM has its origin in path analysis developed by geneticists (Wright 1921, 1934). Founding work in SEM has been done by Haavelmo (1943, 1944) and Koopmans and Hood (1953), work that defined the program of the Cowles Commission and set the stage for modern econometrics (cf. Morgan 1990, Heckman 2000). In fact, SEM has remained the paradigm of causal modeling in contemporary econometrics and the social and behavioral sciences. A set of equations

$$Y = Xb + e$$

is meant to represent a stochastic model in which each equation represents a causal link (Goldberger 1972). All causal connection between Y and X is captured by β, and we infer the causal effect of variation of one element of X relative to its value before variation – holding

all other elements of X constant – on Y relative to its value before variation of X. The "all-other-elements-constant"-clause is well known in economics as the ceteris paribus condition, and in its core goes back to Alfred Marshall (1890 [1965]).

The Potential Outcome Model (POM) of causation predominant in statistics describes a setting in which units are potentially exposed to a set of treatments, and have corresponding outcomes or responses associated with each treatment. The causal connection of interest is the effect on the outcomes of some particular treatment relative to some other particular treatment (often called "control" treatment). Since in reality each unit can only be exposed to one treatment, the other treatment states and associated potential outcomes for the single unit are counterfactuals. In its essence the POM dates back to the work of Neyman (1923 [1990], 1935) and Fisher (1935). Fisher is commonly credited for the invention of randomized experiments, while Neyman was probably the first one to use a model for a treatment effect in which each unit has two responses. Major contributions to the development of the model include Cox (1958), Cochran (1965), and above all Rubin (1974, 1977), who was the first to apply the potential outcome framework to observational studies. See also Rosenbaum (1995a) for further discussion. Due to Rubin's contributions the model is frequently referred to as the "Rubin Model". Related work in economics are models for switching regressions (Quandt 1958, 1972) and the earnings model of Roy (1951). Due to the latter, economic applications occasionally call the POM the "Roy-Rubin-Model".

The use of Directed Acyclical Graphs (DAGs) to assess causal questions is a rather recent phenomenon. The main proponents of graphical approaches to causation are Spirtes, Glymour, and Scheines (2000, first edition 1993) and Pearl (1995, 1998, 2000a).[1] It is difficult to discuss the functioning and mechanisms of DAGs in just a few phrases – for an introduction see the mentioned papers and books. Rather, I want to describe what their advocates think DAGs are aimed at: They are aimed at making causal relations and assumptions and implications in causal models more explicit, in particular more explicit than – in the view of its proponents – other approaches. For instance, Pearl (2000a) claims that recent advances in DAGs have transformed causality from "a concept shrouded in mystery" into a mathematical object with well-defined semantics and well-founded logic. This is another aim of the graphical approach, namely to provide causal talk with a common language

[1] Robins (1986, 1987) offers a graphical approach within the framework of a general counterfactual causal model, related to the POM. See Robins (1995) for how this relates to Pearl's (1995) approach, and see, for instance, Greenland (2000), Robins and Greenland (2000), Pearl (2001) as starting points of the literature on causal inference in epidemiology and the health sciences.

helping researchers communicate (Pearl 1995, 1998), an aim that DAGs do not yet live up to in the view of everybody – see the discussion of Pearl (1995), in particular Imbens and Rubin (1995) and Rosenbaum (1995b). Pearl (2000a) strongly emphasizes the gain in clarity and explicitness gained from causal models based on DAGs in his view. For better or worse, his conclusion is that due to DAGs "causality has been mathematized" (Pearl 2000a).

Naturally, different approaches to questions of causation are viewed differently by proponents of different approaches. For instance, perceptions of SEM as an adequate approach to causation diverge strongly. Pearl (1998) unfolds the idea that the original conceptual strength of SEM along with the clear conception of it among its founding fathers has been lost since, or at least become "obscured". In his perception, social and behavioral scientists – including economists – nowadays struggle for an understanding of either β, or the error term, or both (see Pearl 1998 for examples). In his belief, "the causal content of SEM has gradually escaped the consciousness of SEM practitioners" (Pearl 1998) for two reasons: (i) SEM practitioners have kept causal assumptions implicit in order to gain respectability for SEM, because statisticians, "the arbiters of respectability", abhor assumptions that are not directly testable, and (ii) SEM lacks the notational facility needed to make causal assumptions, as distinct from statistical assumptions, explicit. The latter point means that the SEM founding fathers thought of the equality sign as the asymmetrical relation "is determined by" rather than an algebraic equality, but did not invent a distinct sign for this relation. They were aware of this distinction in meaning, but now their descendants seem to have lost this clear conception – for more on this issue see Pearl (1998), who clearly develops this idea to contrast it with DAGs as a more coherent tool of causal language.

On the other hand, Heckman (2000) is a clear proponent of SEM and forcefully stresses the major role that econometric analysis played in the twentieth century analysis of causal parameters:

> A major contribution of twentieth century econometrics was the recognition that causality and causal parameters are most fruitfully defined within formal economic models and that comparative statics variations within these models formalize the intuition in Marshall's [notion of a ceteris paribus change] and most clearly define causal parameters.

This is how economists define causal effects.[2,3] Heckman (2000) argues that the statistical

[2] It is a bit puzzling, though, that Heckman (2000, p.56) correctly defines causal effects in SEMs as partial derivatives (or as finite differences of some factor holding other factors constant), but previously (p.53) speaks

POM is simply a version of the econometric causal model.[4] This is in line with his finding that the definition of a causal parameter does not require any statement about what is actually observed or what can be identified from data. A finding that also the SEM founding fathers would have subscribed to, as Pearl (1998) refers to Haavelmo (1943) who explicitly interprets each structural equation as a statement about a hypothetical controlled experiment. Here it becomes clear that the ideas of SEM and POM are not at all so far apart. In fact, a system of structural equations is a system of functions from inputs to potential outcomes (cf. also Greenland 2000). Fortunately, recent years have seen substantial convergence of methods from statistics and econometrics along with increased discourse between the two fields, even though some controversy on who deserves credit for what will most likely remain (cf. Heckman and Hotz 1989, Heckman 1996b, Heckman 2000, and Holland 1989, Rubin 1986, 1990, Angrist, Imbens and Rubin 1996).

In summary, one cannot but highly appreciate the vivid debate on causation in the various fields, the expanding amount of causal models suggested, and the many analogies, connections and distinctions that have been drawn between models from different fields. And while one might well wonder whether optimism such as the one recently expressed by Greenland (2000) – "[T]he near future may bring a unified methodology for causal analysis" – seems justified, we do appear to have overcome previous pessimism such as the one expressed by Pearl (1997): "Currently, SEM is used by many and understood by few, while potential-response models are understood by few and used by even fewer." At least, the number of applications has increased substantially; the "understanding" part, though, I would not feel confident to make any judgements about.

2.1.2 Counterfactuals and Causation

Causation has been a major field of philosophical discussion since at least the pathbreaking works of David Hume (1740a [1992], 1740b [1993], 1748 [1993]), disregarding for the moment early works on causes and effects by the Greek philosophers such as Aristotle and others. Indeed, history has seen an abundance of philosophical approaches to causation, and virtually all of them had repercussions on or counterparts in other scientific fields. The

---

of "marginal causal effects". The derivation given there defines the causal effect. There is no such thing as a marginal causal effect.

[3] A different concept of causation in econometrics is Granger causation, which I will not discuss in this chapter. See Granger (1969) for the original account, and, e.g., Holland (1986), Granger (1986), Sobel (1995) for discussion.

[4] The formal equivalence of POM and recursive SEMs has been established by Galles and Pearl (1998).

deterministic account of causation of David Hume demanded temporal priority, spatio-temporal contiguity, and constant conjunction as constituent components of a cause-effect relationship. Ideas that fit quite well with Newton-type mechanics. The other way around: Quantum mechanics did force philosophers to re-think possible theories of causation and consider incorporating probabilistic elements (cf. Skyrms 1984). Such a probabilistic account of causation also displays close intertwining between philosophy, statistics, and econometrics (cf. Salmon 1980, Skyrms 1988, Sobel 1995). These incidents of feedback in both directions are manifold between philosophy and other fields.

The volume edited by Sosa and Tooley (1993a) gives an excellent overview of different approaches to causation by various contemporary philosophers, including discussions of the problems inherent to each approach, and to a philosophically structuralist account of causality in general. One of the most remarkable approaches to causation has been the one suggested by David Lewis. Lewis (1973a) developed possible world semantics for counterfactual conditionals, and proceeded to ground his theory of causation on these counterfactuals (Lewis 1973b and 1986).

Causal inference in statistics is based on counterfactuals. In general this view is uncontroversial. A recent approach to causal inference without counterfactuals suggested by Dawid (2000) has met pronounced rejection – cf. the discussion of Dawid (2000), in particular the comments by Pearl (2000b) and Robins and Greenland (2000). The POM has remained the most prominent approach to causal inference in statistics. The previous subsection has already sketched the basic notions along with the historical evolution of the model. The fact that the POM is based on a counterfactual notion of causation has first been pointed out by Glymour (1986) in his discussion of the seminal paper on causal inference in statistics by Holland (1986). However, to my knowledge there has been no further effort to explicitly link the POM and the account of Lewis, even though the mere fact that there indeed is a link has been stated occasionally, and the notion of "closest possible worlds" is common in talk about causation (cf., e.g., Dawid 2000, Robins and Greenland 2000). The only explicit link I know of is made in Galles and Pearl (1998) who give an axiomatic characterization of causal counterfactuals in comparing logical properties of counterfactuals in structural equation models and Lewis's closest-world semantics.

Why is it that this link has been kept implicit and statistics has sought so little guidance in the philosophical account of counterfactual causation? Some might argue that the POM is a well-defined causal model that does not need further metaphysical or logical

underpinning. On the other hand, Pearl (1997) ascribed the – in his viewpoint – perceived failure of the POM to become standard language in statistical inference to it resting "on an esoteric and seemingly metaphysical vocabulary of counterfactual variables that bears no apparent connection to ordinary understanding of cause-effect processes." Given the pervasive number of recent applications, the reluctancy has disappeared. Or has it? In any case, a fundamental assessment of counterfactuals and their role in causal inference can clarify any researcher's thoughts on causation, the importance of which cannot be overstated (Sobel 1995).

Are counterfactuals "esoteric and metaphysical" entities? And does the idea that they are based on comparative similarity relations between worlds clarify much? Lewis (1973a) replies to the possible objection that these conceptions might be "unclear" with saying that "unclear" is unclear, and drawing the distinction between "ill-understood" and "vague". Counterfactuals and comparative similarity are not ill-understood concepts: they are vague – very vague indeed –, but in a well-understood way (Lewis 1973a).

This chapter concentrates on the counterfactual-based nature of the POM. There have been other connections of causes and counterfactuals (Simon and Rescher 1966) and logical theories of counterfactual conditionals (Stalnaker 1984), but – as mentioned above – the outstanding protagonist has been David Lewis with his account of counterfactual logic based on possible-world semantics (Lewis 1973a), a theory on which he then based his theory of causation (Lewis 1973b). Subsequent criticism on some details of his account led him to refine and supplement his original theory, including probabilistic elements, i.e. "chancy counterfactuals" (Lewis 1986).[5] From the perspective of those applying the POM there is need to further investigate its underlying counterfactual nature, and to clarify the counterfactual semantics that the model – implicitly – uses to infer causal relationships.

2.1.3 Chapter Outline

This chapter contributes to the literature on causation in explicitly linking the POM and Lewis's account, and in delineating which counterfactual causal questions can be asked, and answered, within the model. The procedure is as follows: I start with looking at how causation is modeled in the empirical sciences and find that there are three major approaches: SEM, POM, and DAG. I pick out the POM as the main causal model of interest in evaluation

research. Looking at the model I find that it is formulated in terms of counterfactuals. What, then, are counterfactuals? This leads me to analyze the semantic properties of counterfactuals, and the counterfactual approach to causation in philosophical logic. It turns out that the central element of this approach is the notion of 'possible worlds'. I proceed to connect the counterfactual-based POM with the possible world semantics for counterfactuals, reformulating the POM and its assumptions in terms of counterfactual statements. This procedure (i) connects statistical and philosophical understandings of counterfactuals and (ii) adds clarity to the counterfactual nature of the POM. The chapter then takes a closer look at this crucial notion of proximity of possible worlds, and finds that within the POM closest possible worlds are defined *a priori*, and merely differ with respect to elements of the treatment set T along with associated outcomes. Therefore, I give a detailed discussion of T using a simple set-theoretical framework. This analysis also elucidates which meaningful counterfactual questions can be asked, and answered.

The remainder of this chapter is organized as follows. The second section presents the counterfactual account of causation in terms of Lewis's possible-world semantics. It describes the most prominent features of the theory and includes a short assessment of potential metaphysical shortcomings – or, rather, an assessment of general obstacles for theories of causation, and how Lewis's account addresses these. This comprises a concise review of chancy counterfactuals. Section 2.3 unfolds the POM and its assumptions and reformulates it using the – de facto underlying – ideas of counterfactual conditionals presented in section 2.2. In this respect, sections 2.2 and 2.3 belong together in focussing on foundational aspects of the model.

Sections 2.4 and 2.5 slightly change perspective towards a more applied viewpoint. The fourth section delineates how possible counterfactual worlds within the POM differ only with regard to particular treatments and corresponding responses. In general the model allows for finite T, but both theory and practice have focused on only two elements within T, "treatment" and "control". This is intuitively appealing, as a causal effect can only be inferred for one treatment relative to some other treatment. However, as recent results in evaluation research, e.g., have made explicit extensions to multivalued treatment settings in observational studies possible (Imbens 2000, Lechner 2001a), it appears imperative to discuss various issues that arise for causal inference with finite T and relevant counterfactual queries.

[5] In fact Lewis recently suggested further refinement (Lewis 2000). For the discussion in this chapter, however, only the main results from his original theory are relevant. The ongoing discussion is more important from a

Section 2.4 thus presents a set of meaningful counterfactuals, and includes some examples for illustration. A third subsection proceeds to consider notions about proximity relations between possible worlds. I will show that the empirical procedure assumes or constructs closeness ensuring that only the factor we manipulate is different between worlds. If the assumption holds, then closeness is ensured, then the counterfactual conditions hold and the model produces valid inference. The fifth section gives more details on a few specific problems that could arise in practice – this section might be of interest above all for applied social scientists using matching methods. Section 2.6 concludes.

2.2 The Counterfactual Account of Causation

Many philosophical discussions of causation begin with – or entail at some stage – the probably most famous quote of David Hume, and present the puzzle which the quote comprises:

> […T]herefore, we may define a cause to be *an object, followed by another, and where all the objects, similar to the first, are followed by subjects similar to the second.* Or in other words, *where, if the first object had not been, the second never had existed.* [Hume 1748 [1993], his italics]

Of course it is not puzzling in the sense that Hume's work represents the foundation of the analysis of causation. His writings – including the major Hume 1740a [1992] – have shaped the examinations of the principles of causation until today. On the other hand, the cited passage is puzzling in the sense that Hume certainly was aware of the regularity-based nature of his first definition, but apparently not of the counterfactual nature of his alternative definition.

Counterfactuals therefore did not play any role in Hume's understanding of causation, and in fact it was not until the 1970s that the link between counterfactuals and causation became object of a thorough philosophical analyses. Before, philosophers had been concerned enough with such a vague concept like counterfactuals themselves – cf. Menzies 2001a for a concise review of early counterfactual theories.

The better understanding of a counterfactual approach to causation was basically due

---

strictly philosophical point of view (cf. Menzies 2001b).

to the development of possible world semantics for counterfactual conditionals by Robert Stalnaker (1984) and David Lewis (1973a, 1986). As the concept of possible worlds plays an important role in this chapter, I will review the basic ideas in some detail. The discussion is along the line of thought suggested in Lewis (1973a and 1973b) and further refined in Lewis (1986). "Along the line of thought" in this context means that there are many aspects in the philosophical assessment of causation that are of minor interest to econometricians, statisticians, and social scientists. Philosophical approaches to causation have always tried – or, rather, have always had to try – to give a metaphysical account fundamentally explaining how causation works in our world. Needless to say that there have always been possible objections and counterexamples to each 'structuralist' theory of causation, that the presented theory could not grasp (cf., e.g., Sosa and Tooley 1993b). The empirical sciences however, do not need a causal theory that explains how causal processes work under each and every circumstances in our world. A 'realist' or 'reductionist' approach is sufficient: Clearly, specific questions like whether it is the table top that causes the table feet to have exactly the length they have, or whether it is rather the table feet causing the table top to assume the spatio-temporal position it occupies, or maybe both, are of subordinate importance. Also, discussions about direction of time in a cause-effect relationship are of minor concern, as a situation in which the effect precedes its cause is extremely unlikely in the empirical sciences.[6]

This chapter therefore focuses on those aspects of a counterfactual theory of causation in terms of possible worlds that are of direct use to empirical social scientists. It is both impossible to include a full metaphysical account of this theory, or even to come anywhere near a fair metaphysical account, as well as unnecessary in the given context. I would hope that philosophers would nevertheless agree with (a) the main points I extract from Lewis's theory, and (b) the claim that thinking along these lines can considerably help to sharpen any researcher's thoughts on causal inference based on counterfactuals.

2.2.1 Possible World Semantics

Lewis's theory of causation employs possible world semantics for counterfactual conditionals, providing truth conditions for counterfactuals in terms of relations between possible worlds. Again, in this exposition we need not worry about the realism of these possible worlds,

[6] To take an example from economic evaluation of policy interventions: Even if an individual participates in a program because of her expecting the program to raise future earnings, it is not the (potential) future earnings that may have caused her to participate, but the thought (in the present) of the program raising the earnings. Expectation, though directed at the future, is very much a concept of the present.

whether they are "maximally consistent sets of propositions", or "theoretical entities having no independent reality", etc. (cf. Menzies 2001a). Regardless of metaphysical subtlety they provide us with a useful framework of causal thinking.

This section considers the deterministic perspective. Possible world semantics for counterfactuals are based on the main idea of *comparative similarity* between worlds. Given a set of worlds W, according to Lewis (1973b) one world $w_j \in W$ is *closer* to a given world $w_i \in W$ than another world $w_k \in W$ if w_j resembles w_i more than w_k resembles w_i. Naturally, this notion of closeness is based on the idea of w_i being the actual world, and defining $w_j, w_k \in W$ with respect to their proximity to actuality[7]. Lewis imposes two formal constraints on this similarity relation: (i) It produces a weak ordering of worlds such that any two worlds can be ordered with respect to their closeness to the actual world, where "weak" implies that ties are permitted, but any two worlds are comparable. (ii) The actual world is closest to actuality, resembling itself more than any other world does.

For any two propositions C and E, define the following counterfactual conditionals:

(2.1a) $C \Box\!\!\rightarrow E$ "If C were (had been) the case, then E would be (have been) the case."[8]

and

(2.1b) $\sim C \Box\!\!\rightarrow \sim E$ "If C were not (had not been) the case, then E would not be (not have been) the case."

Then the counterfactual conditional $C \Box\!\!\rightarrow E$ is characterized by the following truth condition in terms of the similarity relation:

(2.2) $C \Box\!\!\rightarrow E$ is true at a world $w_i \in W$ iff either (i) there are no possible C-worlds or (ii)

---

[7] "Actuality" means the "world of point of view", i.e. "actual" refers at any world w_i to that world w_i itself. Lewis (1973a): "'Actual' is indexical, like 'I' or 'here', or 'now': it depends for its reference on the circumstances of utterance, to wit the world where the utterance is located." The actual world is only one world among others, and we call our world "actual" because it is the one we inhabit, not because it differs in kind from all the rest (cf. Lewis 1973a).

[8] "$\Box$" represents a *necessity operator* or *'would'-counterfactual*. I have omitted the consideration of the *possibility operator* or *'might'-counterfactual* "$\Diamond$", from the discussion. In terms of the truth conditions for counterfactuals, "$\Box$" implies truth at all accessible (from the actual world) worlds, while "$\Diamond$" implies truth only at some accessible worlds (cf. Lewis 1973a). Causal inference in the social sciences is exclusively interested in the 'would'-counterfactual.

some C-world where E holds is closer to w_i than any C-world where E does not hold.

(i) is the trivial case and implies that the counterfactual is vacuously true. From the perspective of w_i being the actual world, the idea of (ii) is that C $\Box\!\!\rightarrow$ E is (nonvacuously) true in the actual world if it takes less of a departure from actuality to make the antecedent true along with its consequent, than it does to make the antecedent true without the consequent (Lewis 1973b). Under the assumption that there must always be one or more closest C-worlds this condition simplifies to C $\Box\!\!\rightarrow$ E being nonvacuously true iff E holds at all the closest C-worlds.

Example A. The classic illustration of Lewis (1973a): "If kangaroos had no tails, they would topple over."

Lewis (1973a) underscores this exemplification in explaining what he thinks that such a counterfactual sentence is supposed to mean: "In any possible state of affairs in which kangaroos have no tails, and which resembles our actual state of affairs as much as kangaroos having no tails permits it to, the kangaroos topple over". This statement entails most of what the analysis of counterfactual conditionals is about, namely that a counterfactual sentence corresponds to an actual state of affairs, and that the counterfactual is true if it deviates from actuality only to minimum extent.

Example B. Assuming that it characterizes minimum deviation from actuality, the counterfactual "If John participated in the computer course, he would find a job" is true corresponding to the actual state of affairs in which John does not participate in the computer course and does not find a job.

So far this principal idea considers propositions, not events. Lewis (1973b) extends this setting by pairing the two: To any possible event e there corresponds the proposition O(e) that holds at all and only those worlds where e occurs. Thus, O(e) is the proposition that e occurs, i.e. O(e) is a sentence describing the occurrence of the particular event e, and counterfactual dependence among events is simply counterfactual dependence among the corresponding propositions. We then have a definition of causal dependence:

(2.3) Let c and e be two distinct possible particular events. Then e causally depends on c iff O(c) $\Box\!\!\rightarrow$ O(e) and ~O(c) $\Box\!\!\rightarrow$ ~O(e).

This condition states that whether e occurs or not depends on whether c occurs or not. The dependence consists in the truth of the two counterfactuals O(c) $\Box\!\!\rightarrow$ O(e) and ~O(c) $\Box\!\!\rightarrow$ ~O(e). Consider two cases: first, if c and e do not actually occur, then the second counterfactual is automatically true because its antecedent and consequent are true. Thus, e depends causally on c iff the first counterfactual holds, i.e., iff e would have occurred if c had occurred. Second, if c and e are actual occurrent events, it follows from the second formal condition on the comparative similarity relation (cf. above) that the first counterfactual is automatically true, because the condition implies that a counterfactual with true antecedent and true consequent is itself true. Thus, e depends causally on c iff, if c had not been, e never had occurred. This is exactly Hume's second definition of causation.

To put it simply:

(2.3a) c causes e iff both c and e are actual occurrent events and if c had not occurred then e would not have occurred.

Or, using the possible world semantics for counterfactuals:

(2.3b) c causes e iff both O(c) and O(e) are true in the actual world and in the closest (to the actual world) possible world in which O(c) is not true, O(e) is not true.

2.2.2 Chancy Counterfactuals

What about this counterfactual theory if causation was probabilistic rather then deterministic? It appears natural to compare the counterfactual account of causation – be it from a strictly philosophical viewpoint (Lewis 1973b) or from a statistical perspective (cf. Holland 1986, 1988a, and see below) – with alternative approaches that identify causal dependence in terms of probabilistic relations. There is a rich philosophical literature addressing probabilistic causality in various forms, cf. Reichenbach (1956), Good (1961, 1962), and Suppes (1970) for three classic theories, but also Salmon (1980, 1998) and Suppes (1984), among others.

Probabilistic concepts of causality are used in innumerable contexts of everyday life and science.[9] They have the advantage that they can indeed accommodate many of our everyday experiences, that they appear to make it easy to understand how we can have

knowledge of causal relations, in particular in cases where we seem to observe causation, but no determination. As Glymour (1986) puts it: "Technical details aside, causal inference becomes a statistical estimation problem." Could it possibly sound any better to econometricians? But on the other hand, probabilistic concepts of causality have the disadvantage that they do not always coincide appropriately with our intuitive judgements about causal relations, and that causation in terms of percentages may be difficult to conceive.

Nevertheless, contemporary physics – here: quantum mechanics – tells us that our world is full of probabilistic processes that are of causal character (cf. Lewis 1986; or Skyrms 1984 for a discussion of the implications for causality of the Einstein-Podolsky-Rosen (1935) paradox). Thus, Lewis (1986) argues that a theory of causation must accommodate the conceptual possibility of chancy causation. He combines his counterfactual theory of causation with elements of probabilistic concepts, and defines a more general notion of causal dependence in terms of chancy counterfactuals.

(2.4) $C \boxrightarrow Pr(E)=x$ "If C were the case, then E would be the case with probability x"

The counterfactual is thus an ordinary world counterfactual that can be interpreted according to the semantics above. The *Pr* operator is a probability operator with narrow scope confined to the consequent of the counterfactual. In this context, the definition of causal dependence becomes:

(2.5) Let c and e be two distinct possible particular events. Then e causally depends on c iff, if c had not occurred, the probability of e's occurring would have been much less than it actually was (given that c occurred).

Obviously this definition comprises the deterministic causal relation in which the probability of the effect along with the cause is 1 and the probability of the effect without the cause is 0. Chancy counterfactuals are thus a straightforward extension to "normal" counterfactuals. For further discussion of probabilistic concepts of causation see, for instance, Hitchcock (1997), Skyrms (1988) and references therein, and Sobel (1995) for the connection to causal inference in the social sciences.

---

[9] Cf. Salmon (1980) both for an account of the three classic theories as well as a number of illustrative examples. To name but one: "We have strong evidence that exposure to even low levels of radiation can cause leukemia,

2.2.3 Applicability

In this subsection I will briefly review some of the problems that arise or need to be addressed in (philosophical) theories of causation. In particular, I will discuss event causation, spurious non-causal dependence, temporal asymmetry, as well as transitivity and preemption.

Above we adopted a definition of causal dependence relating two events. It is, however, not evident that it is events that are the fundamental relata of causal dependence. The philosophical literature clearly distinguishes between event causation and causal theories based on facts or state of affairs (cf. Bennett 1993). Moreover, it implies the necessity to define a certain notion of what is an event – cf. for instance Lewis (1986) for his construction of events as classes of possible spatiotemporal regions. But in general this problem does not arise in the applied social sciences: First, Menzies (2001a) states that even under a metaphysical perspective very different conceptions of events are compatible with the basic definition of causal dependence in terms of counterfactuals. Secondly, it appears straightforward to incorporate the events of interest in empirical research – such as, e.g., a training program, medical treatment, etc. – into the above framework of event causation.

Another element of the above definition requires the causally dependent events to be distinct from each other. This feature rules out what is called spurious non-causal dependence. Consider an example from Kim (1973): Writing the letter "r" twice in succession is a constituent event in the event of writing "Larry". Thus: "If I had not written 'r' twice in succession, I would not have written 'Larry'." The counterfactual is true, but there is no causal relation between the events. But since the events are not distinct from each other, the relation does not count as causal dependence.

> **Example C.** In econometric evaluations of employment programs one occasionally finds puzzling statements of the sort that the program "increased significantly the employment […] *during* the period of program participation" (Fraker and Maynard 1987, my italics). Clearly, it is impossible to disentangle causal dependence from spurious non-causal dependence if the potential cause (=being employed) is not distinct from the potential effect (=being employed).

What is the temporal structure of causation? As social scientists, both our intuition and the analyses that we deal with in practice suggest clearly that causes would typically precede their effects. But why do we commonly associate causal relations with the temporal direction from past to present or future? Lewis (1979) addresses this point and indeed argues that the

---

though only a small percentage of those who are so exposed actually develop leukemia."

direction of causation is the direction of causal dependence, and that it is typically true that events causally depend on earlier events, but not on later events. He notes, however, that the conceptual idea of time-reversed or backward causation cannot be ruled out a priori.

I do not intend to go deep into this analysis, cf. Lewis (1979) and Horwich (1993) to grasp the major issues, but I do want to note the two main points emerging from the discussion: (i) Lewis (1979) defines a determinant for an event as any set of conditions jointly sufficient – given the laws of nature – for the event's occurrence. Looking from the two directions of time, determinants can be causes or *traces* of an event. Any particular fact about a deterministic world is predetermined throughout its past and postdetermined throughout its future. Lewis (1979) observes it to be contingently true that events typically have very few earlier determinants but very many later determinants[10]. This is called asymmetry of overdetermination.

Lewis (1979) combines this de facto temporal asymmetry of causal dependence with (ii) his analysis of the comparative similarity relation between worlds. The comparative similarity analysis implies that the most similar worlds are those in which the actual laws of nature are never violated, and exact similarity regarding particular matters of fact in some spatiotemporal region is an important element of similarity if it can be ensured by a small, local miracle, rather than at the cost of big, global miracle. In connection with the asymmetry of overdetermination, this argument (cf. Lewis 1979, Menzies 2001a for details) implies that it is easier to reconcile a hypothetical change in the actual course of events by preserving the past and allowing for a divergence miracle than by shielding the future from change by virtue of a convergence miracle. The main result here is that – given the asymmetry of overdetermination – the present counterfactually depends on the past, but not on the future.

It has to be noted that – strictly speaking – Lewis (1973b) uses the definition given in (2.3a) and (2.3b) only as a definition of "causal dependence among actual events". His actual definition of causation is based on the notion of causal chains: Lewis (1973b) states that causal dependence between actual events is sufficient for causation, but not necessary. As it can happen that three actual events c, d, and e are of the form that d would not have occurred without c, and e would not have occurred without d, but e would still have occurred without c, causal dependence may not be transitive. Nonetheless, Lewis (1973b) insists that causation must always be transitive, i. He therefore extends causal dependence to a transitive relation,

[10] An example (Menzies 2001a): A spherical wave expanding outwards from a point source is a process where each sample of the wave postdetermines what happens at the point at which the wave is emitted.

where c, d, e, … is a finite sequence of actual particular events such that d causally depends on c, e on d, etc. This he calls causal chain. The definition of causation then becomes

(2.6) c is a cause of e iff ∃ a causal chain leading from c to e.

This definition ensures transitivity of causation, and it provides a solution to the problem of causal preemption. Causal preemption takes place when the cause of an event preempts something else from causing that event (cf. Horwich 1993 or Menzies 2001b for examples). Using definition (2.6) it is possible, however, to distinguish preempting actual causes from preempted potential causes.

2.3 The Potential Outcome Model for Causal Inference: A Reformulation

The statistical model called POM – based mainly on work by Neyman (1923 [1990], 1935), Fisher (1935), Cox (1958), Cochran (1965) Rubin (1974, 1977, 1978, 1980, 1986) – provides a solid ground for causal inference in experimental and observational studies. As it is implicitly cast into a counterfactual framework, it directly relates to – or: is grounded on – many of the aspects of counterfactual logic presented in the previous section. I will give a fairly detailed review of the basic model, and show how it is connected to the possible world semantics presented above. Much of the presentation of the original POM is based on the discussions in Holland (1986, 1988a), since these provide a very clear account of the theory.

2.3.1 The Causal Model

The logical elements of the POM are a quadruple of the form {U, T, D, Y}. These four elements constitute the primitives of the model. U is a population of N units u $[u_1,...,u_n]$, T is a set of M treatments[11] t $[t_1,...,t_m]$ to which each one of the units u may be exposed, D(u)=t

[11] I chose to stick to the formulation of treatment, rather than, e.g., calling it a "cause" (Holland 1988a) for two reasons: (i) The empirical context of the POM that we are interested in is exactly that of 'treatments' like medicaments in health sciences or policy interventions such as training courses in the social sciences. (ii) From an intuitive linguistic perspective a cause implies an effect. A priori we do not know whether an effect will be observed, the cause of which we desire to infer. So, from the point of view that we do not yet know about an

indicates that unit u is actually exposed to a particular treatment t out of T, and Y(u,t) equals the value of the outcome that would be observed if unit $u \in U$ were exposed to treatment $t \in T$. U and T are sets, D is a mapping of U to T, and Y(.) is in general a real-valued function of (u, d).

Note that the response variable Y depends on both the unit u and the treatment t to which the unit is exposed. If u were exposed to some $t_1 \in T$, we would observe the value of the outcome $Y(u,t_1)$, and if u were exposed to some $t_2 \in T$, the observed response value would be $Y(u,t_2)$. The meaning of Y to be a function of pairs (u,t) is that it represents the measurement of some characteristic of u after u has been exposed to $t \in T$. This requirement implies that it must be possible for any unit in U to be potentially exposed to any treatment t out of T. Holland (1988a) emphasizes the importance of this condition: It entails a certain notion of what is a cause, that is of fundamental importance in preventing us from interpreting associational relations as causal ones, like, e.g., associations between sex and income or between race and income. This condition of the POM and its relevance is discussed more extensively in Holland (1986, 1988a, 1988b) and Glymour (1986). The main point that we can derive at this stage is that this condition de facto states that causes must be events.

Call Y the outcome function and let $Y_t(u)=Y(u,t)$. The mapping D is called the assignment rule because it indicates to which treatment each unit is exposed. The observed outcome of each unit $u \in U$ is given by

$$Y_D(u)=Y(u,D(u)),$$

which is the value of Y that is actually observed for unit u. Therefore, the pair $(D(u),Y_D(u))$ – where D(u) indicates the treatment in T to which u is actually exposed – constitutes the observed data for each unit u. Note the distinction between $Y_D(u)$ and $Y_t(u)$: While the former is the outcome actually observed on unit u, the latter is a potential outcome being actually observed only if D(u)=t.

In the model, treatments are taken as undefined elements of the theory, and effects are defined in terms of these elements (Holland 1988a). The basic causal parameter of interest is

---

effect and that a zero effect is usually not called an effect, I think that we cannot call T a set of causes a priori. The formulation of treatment is unambiguous.

(2.7) *The unit-level treatment effect (UTE)*:

The unit-level causal effect of treatment $t \in T$ relative to treatment $c \in T$ (as measured by Y) is the difference $Y_t(u)-Y_c(u)=UTE_{tc}(u)$.[12]

There are three important things to note about this definition. First, the causal effect $UTE_{tc}(u)$ is defined at the individual-unit level. Second, $UTE_{tc}(u)$ is the increase in the potential value of $Y_t(u)$ over the potential value of $Y_c(u)$. Third, $UTE_{tc}(u)$ is defined as the causal effect of t relative to c. The following discussion will center around elements number two and three:

Consider $UTE_{tc}(u)$ being the increase in the potential value of $Y_t(u)$, which is what would be observed for the potential outcome if D(u)=t, over the value of $Y_c(u)$, which is what would be observed for the potential outcome if D(u)=c. Here it becomes clear that this is a definition based on causal dependence in counterfactual terms. Define the following set of events:

e_k: Unit $u \in U$ is exposed to treatment $t_k \in T$, i.e. $D(u)=t_k$, and

e^*_k: Unit $u \in U$ has the value $Y_{t_k}(u)$ for variable Y,

where k=1,…,m, so that the number of events for each individual unit is 2×M (as there are N units in U, the total number of events is 2×M×N). Then:

(2.8) The unit-level causal effect of treatment $t_i \in T$ relative to treatment $t_j \in T$ (as measured by Y) is defined by the difference $Y_{t_i}(u) - Y_{t_j}(u) = \mathrm{UTE}_{t_i t_j}(u)$ iff the counterfactual conditionals $O(e_i) \mathrel{\square\!\!\rightarrow} O(e^*_i)$, $\sim O(e_i) \mathrel{\square\!\!\rightarrow} \sim O(e^*_i)$, and $O(e_j) \mathrel{\square\!\!\rightarrow} O(e^*_j)$, $\sim O(e_j) \mathrel{\square\!\!\rightarrow} \sim O(e^*_j)$ are true.

To illustrate this reformulation, let us return to the treatment-versus-control case. Define the events:

e_1: Unit u is exposed to treatment t

[12] Note that the two treatments in this definition are denoted with t (like "treatment") and c (like "control"). This already hints at the discussion of randomized assignment of units into an experimental treatment or control group. It also gives a particular flavor to the definition of an effect of one treatment relative to a control

e_2: Unit u is exposed to treatment c

e^*_1: Unit u has the value $Y_t(u)$ for variable Y

e^*_2: Unit u has the value $Y_c(u)$ for variable Y

In this special case (2.8) simplifies to:

(2.8a) The unit-level causal effect of treatment $t \in T$ relative to treatment $c \in T$ (as measured by Y) is defined by the difference $Y_t(u)-Y_c(u)=UTE_{tc}(u)$ iff the counterfactual conditionals $O(e_1) \Box\!\!\rightarrow O(e^*_1)$, $\sim O(e_1) \Box\!\!\rightarrow \sim O(e^*_1)$, and $O(e_2) \Box\!\!\rightarrow O(e^*_2)$, $\sim O(e_2) \Box\!\!\rightarrow \sim O(e^*_2)$ are true.

Recall (2.3), and note that we have two underlying causal dependencies: e^*_1 causally depends on e_1, and e^*_2 causally depends on e_2. However, to infer either of the two causal dependencies – in this case: that between e_1 and e^*_1 – we need the other one. This is because causal inference can only be made relative to something (cf. below).

Furthermore note the formulation of "distinct possible particular events" in (2.3), because looking at (2.8a) and recalling the special case of (3a) we would seem to encounter a problem: Not all of the events e_1, e_2, e^*_1, and e^*_2 can be "actual occurrent events" for a specific unit u, because at the individual-unit level only either e_1 and e^*_1, or e_2 and e^*_2 can be "actually occurrent". Now consider the formulation using possible world semantics for counterfactuals in (2.3b): In our example, clearly e_1 causes e^*_1, because either both $O(e_1)$ and $O(e^*_1)$ are true in the actual world or, in the closest (to the actual world) possible world, both $O(e_1)$ and $O(e^*_1)$ are not true, because in that closest world $O(e_2)$ and $O(e^*_2)$ are true. It is easy to see that the same argument holds the other way around for e_2 and e^*_2. Note that in this particular case we only have two worlds, and we define the causal effect in one world (the e_1-e^*_1-world) relative to the second and trivially closest world (the e_2- e^*_2-world), whichever one may be the actual world. Thus, also the symmetry of the analysis appears obvious.

This causal analysis can be summarized in four steps: (i) The basic causal parameter of interest is the unit-level treatment effect UTE of some treatment t relative to another treatment c. (ii) UTE is defined iff the pairs of events e_1 and e^*_1, and e_2 and e^*_2 (as defined above) are both causally dependent. (iii) These pairs of events are causally dependent iff the

---

treatment, where the term "control" usually implies "no treatment". Moreover, note that the notation E_{tc} is meant to indicate the causal effect of "t relative to c".

counterfactual conditionals $O(e_1) \Box\!\!\rightarrow O(e^*_1)$, $\sim O(e_1) \Box\!\!\rightarrow \sim O(e^*_1)$, and $O(e_2) \Box\!\!\rightarrow O(e^*_2)$, $\sim O(e_2) \Box\!\!\rightarrow \sim O(e^*_2)$ are true. (iv) A counterfactual conditional of the type $O(a) \Box\!\!\rightarrow O(b)$ is true in the actual world iff any a-world along with b is closer to actuality than any a-world without b.

Return to (2.7) and the third aspect we noted, that $UTE_{tc}(u)$ is defined as the causal effect of t relative to c. Indeed, the effect of one treatment is always relative to the effect of another treatment. We can only draw inference on the cause of an effect by relating two effects of two distinct causes (or: potential causes, or treatments). This relativity condition is one of the central aspects of the POM: a causal relation between a treatment and an effect can be identified from "measuring" two alternative states of an outcome variable given some unit has been exposed to some treatment or another. As Glymour (1986) puts it: "Causation is a relation between two treatments and two possible variable states. The notion of t causing Y_t, without specification of any alternative treatment, or any alternative state of Y, is not defined." Glymour (1986) regards this as an improvement on the bare counterfactual account of causal relations, and he presents an example supporting his argument. I include this example here, because I think it is highly elucidating with respect to the relativity condition and the idea of possible worlds behind it [my italics]:

> **Example D.** My Uncle Schlomo smoked two packs of cigarettes a day, and I am firmly convinced that smoking two packs of cigarettes a day caused him to get lung cancer. But it may not be true that in the closest possible world in which Uncle Schlomo did not smoke two packs a day, he did not contract cancer. Reflecting on Schlomo's addictive personality, and his general weakness of will, it may well be that the closest possible world in which Schlomo did not smoke two packs of cigarettes a day is a world in which he smoked three packs a day. I can reconcile this reflection with the counterfactual analysis of causality by supposing […] that 'smoking two packs of cigarettes a day caused him to get lung cancer' is elliptical speech, and what is meant, but not said, is that smoking two packs of cigarettes a day, *rather than not smoking at all*, caused Schlomo to contract lung cancer.

This example highlights many of the inherent features of the model. I want to point out two more aspects that will require further discussion below. First, many of our casual causal thoughts are based on just that idea of inferring causal relations from saying "doing a relative to not-doing-a", where not-doing-a may equal doing-nothing, and for most analyses this is exactly the causal question of interest. Second, this example nicely raises the question of how can we identify the closest world to actuality (or, the world that we are interested in and use as a base category), or, for the very least, how can we derive any notion what the features of this

closest world are supposed to look like. I will examine these two issues in section 2.4.

Let us rephrase Example D using the framework introduced above. This will accentuate the way the model works, even though it does not exactly correspond to an "exposure to treatment" context and may thus sound a bit odd at first sight, and even though it disregards issues of timing, disturbing factors etc. Define the following events:

e_1: Schlomo ("Unit u") smokes 2 packs of cigarettes a day. ("Treatment t")
e_2: Schlomo does not smoke at all. ("Treatment c")
e^*_1: Schlomo contracts lung cancer. ("$Y_t(u)$")
e^*_2: Schlomo does not contract lung cancer. ("$Y_c(u)$")

The outcome variable Y can be regarded as "health status" or something similar. According to (2.3) and (2.3a,b) e_1 causes e^*_1 because both are actual occurrent events and if the former had not occurred then the latter would not have occurred. This alternative world is specified by e_2 and e^*_2. The two crucial things about this example are (i) that we have an explicit specification of the closest possible world to actuality, and (ii) that this closest world to actuality is defined by the fact that the actual occurrent events do not occur, i.e. e_2=~e_1, and e^*_2=~ e^*_1.

Feature (i) is far from unusual, because in fact the relativity condition in (2.7) *per definitionem* specifies the closest world – i.e. the "treatment-c-world" – to the actual world – i.e. the "treatment-t-world".[13] Causal inference about treatment t is based on the counterfactual relation to what would have happened under exposure to treatment c. In that sense the model does not depend on *searching* for the closest possible world, but rather on *justifying* the choice of what is claimed to be the closest possible world, or the relevant alternative world. Feature (ii) of our example says that in the relevant alternative world e_2=~e_1, and e^*_2=~ e^*_1, i.e. treatment c is merely the absence of treatment t. This is a special case in which UTE is defined under the simplified condition that only $O(e_1) \boxrightarrow O(e_3)$ and $\sim O(e_1) \boxrightarrow \sim O(e_3)$ need to be true.[14] In fact, this is how causal inference is usually made, and it is what Glymour (1986) means with "elliptical speech": We infer the causal effect of

[13] As pointed out in footnote 7 "actual world" means something like "world of departure", or "world of point of view", or "world of interest" due to the analysis being symmetric: If the c-world is closest to the t-world, then also the t-world is closest to the c-world, and the inferred causal effects are the same in magnitude, but in opposite direction or with opposite sign.

[14] Because $O(e_2) \boxrightarrow O(e^*_2) = O(\sim e_1) \boxrightarrow O(\sim e^*_1) = \sim O(e_1) \boxrightarrow \sim O(e^*_1)$, and $\sim O(e_2) \boxrightarrow \sim O(e^*_2) = \sim O(\sim e_1) \boxrightarrow \sim O(\sim e^*_1) = O(e_1) \boxrightarrow O(e^*_1)$.

something relative to not-that-something. The general definition (2.8) accommodates this simplified case, but above all it accommodates the case of "something relative to something else", i.e. in the example a valid causal relation between treatment t and its outcome (= the e_1-e^*_1-world) and some alternative treatment c and its outcome (= the e_2-e^*_2-world). It is important to note, however, that according to (2.8) this causal relation between the e_1-e^*_1-world and the e_2-e^*_1-world does entail some statements about the non-occurrence of either event, as both conditionals $\sim O(e_1) \boxrightarrow \sim O(e_3)$ and $\sim O(e_2) \boxrightarrow \sim O(e_4)$ need to be true, and therefore some consideration of the no-treatment-state is at least implicit.

2.3.2 Applicability

I have emphasized before that the definition of causal dependence is based on the notion of distinct possible particular events. And it is with respect to any two distinct treatments t and c that we face the fundamental problem of causal inference in practice: It is impossible to simultaneously observe $Y_t(u)$ and $Y_c(u)$, and therefore also the causal effect $UTE_{tc}(u)$ is never directly observable. This is why we need counterfactual statements about possible worlds. The counterfactual statement enters in the form illustrated in Examples A and B: We have an actual state of affairs – i.e. for instance we observe u being exposed to t and responding with $Y_t(u)$. We then infer the causal effect of t on Y(u) by relating this actual state of affairs to the counterfactual statement about how u would have responded – i.e. $Y_c(u)$ – if u had been exposed to c, where the counterfactual characterizes a possible world with minimum deviation from actuality.

Holland (1988a) stresses how the POM makes the unobservability of the causal effect explicit in separating the observed pair (D,Y_D) from the function Y. In fact, a model for causal inference can be interpreted as some specification of the values of Y. In Holland's (1988a) words, causal inference consists of combining (a) a causal model or causal theory, (b) assumptions about data collection, and (c) observed data to draw conclusions about causal parameters. The causal model has been laid out above – this section focuses on how and under what assumptions this model can be applied. I will first discuss one basic assumption and subsequently review the conditions under which we can identify the causal effect from data. Usually this implies imposing restrictions on either Y and/or U that make it possible to assess two potential outcomes for a single unit and therefore infer meaningful causal statements.

The *stable-unit-treatment-value-assumption* (SUTVA) is the pivotal assumption ensuring that the causal framework of the POM is adequate in practice. SUTVA is advocated

by Rubin (1980, 1986) to play a key role in deciding which questions are formulated well enough to have causal answers. It is the a priori assumption that the value of Y for unit u when exposed to treatment t is the same independent of (i) the mechanism that is used to assign t to u, and (ii) what treatments d the other units v≠u receive, and that this holds for all n units within U and m treatments within T. SUTVA is violated when, for instance, there is interference between units that leads to different outcomes depending on the treatment other units received – i.e. Y_{tu} depends on whether v≠u received t or some other d ∈ T – or there exist unrepresented versions of treatment or versions of treatments leading to "technical errors" (Neyman 1935)[15] – i.e. Y_{tu} depends on which (uninteded) version of treatment t unit u was exposed to.

In the counterfactual conditionals framework SUTVA can be represented as follows. Define the following set of events:

e_{ij}: Unit $u_i \in U$ is exposed to treatment $t_j \in T$, i.e. $D(u_i)=t_j$, and

e^*_{ij}: Unit $u_i \in U$ has the value $Y_{t_j}(u_i)$ for variable Y,

where i=1,…,n and j=1,…,m, so that the number of events is 2×N×M. Then SUTVA assumes that the counterfactual conditionals $O(e_{ij}) \Box\!\!\rightarrow O(e^*_{ij})$ and $\sim O(e_{ij}) \Box\!\!\rightarrow \sim O(e^*_{ij})$ are true $\forall$ e_{ij} and e^*_{ij} with i=1,…,n and j=1,…,m independent of (i) the mechanism leading events e_{ij} to occur, and (ii) the other occurring events e_{kl}, k=1,…,n, l=1,…m, i≠k, j≠l.[16]

---

[15] For further detail cf. Rubin (1980) and Rubin's (1990) discussion of Neyman (1923 [1990]). Many aspects of the POM for causal inference (in particular the notion of potential outcomes) are already present in the work of Neyman (cf. also Speed 1990), where they are based on the methodological discussion of agricultural experiments. In that context, possible violations of SUTVA are apparent: How should one avoid neighboring plots treated differently (by, e.g., different fertilizers) to "interfere" given nature's powers (wind, rain etc.), or how can one claim that each bag of fertilizer represents exactly the same treatment as any other bag of fertilizer (cf. Rubin 1986)? Moreover, as Rubin (1990) points out, interference between units can be a major issue when studying medical treatments for infectious diseases, or educational treatments given to children who interact with each other.

[16] I use individual statements of the form $O(e_{ij}) \; o\!\!\rightarrow O(e^*_{ij})$ to represent the POM using Lewis's semantics. Galles and Pearl (1998) use an equivalent representation that could be called a "covering counterfactual" and translate Lewis's statement $A \Box\!\!\rightarrow B$ as "If we force a set of variables to have the values A, a second set of variables will have the values B". If A stands for a set of values $x_1,\ldots,x_n$ of the variables $X_1,\ldots,X_n$ and B for a set of values $y_1,\ldots,y_m$ of $Y_1,\ldots,Y_m$ then

$$A \Box\!\!\rightarrow B \equiv \begin{array}{l} Y^1_{x_1\ldots x_n}(u) = y_1 \,\& \\ Y^2_{x_1\ldots x_n}(u) = y_2 \,\& \\ \ldots \\ Y^m_{x_1\ldots x_n}(u) = y_n \end{array}$$

The fundamental requirements of SUTVA therefore go hand in hand with the philosophical underpinnings I have presented above. To conclude with Rubin (1986, his italics):

> [T]he crucial point [...] is that we are not ready to estimate, test, or even logically discuss *causal effects* until units, treatments, and outcomes have been defined in such a way that SUTVA is plausible.

Unit homogeneity is a name given by Holland (1986a) to the assumption that the responses of all units to a particular treatment are the same, i.e. that units respond homogeneously to each treatment:

$$Y_t(u)=Y_t(v) \qquad \forall\, u,v \in U, \text{ and all } t \in T.$$

This is a partial specification of Y in that it restricts the values that Y can take on but does not specify them completely. The assumption is only likely to be justified if one can claim to be working with a homogeneous sample. Under the assumption of unit homogeneity, the causal effect of a treatment t relative to a treatment c is given by

$$UTE_{tc}(u)=Y_t(u) - Y_c(v)=Y_t(v)-Y_c(u)$$

for any two distinct units u and v in U. In this case, UTE_{tc} is a constant and does not depend on the unit under scrutiny. Evidently, unit homogeneity solves the fundamental problem of causal inference in that we only need to measure the two (observable) outcomes $Y_{D(u)=t}(u)$ and $Y_{D(v)=c}(v)$ for two units u and v to infer the causal effect of treatment t relative to treatment c on *any* unit within U. This assumption affects condition (2.8) simply by providing us with an easy answer with respect to what happens for any unit of U in the closest possible world to that very unit. In the closest possible world to the one in which a unit u is exposed to t and responds with $Y_t(u)$, u would be exposed to c and respond with $Y_c(u)$ by assumption, and under unit homogeneity the latter value of Y is given by the response of any other unit $v \neq u$ of U being (or having been) exposed to c.

Unless unit homogeneity holds, individual effects are impossible to observe. Therefore, one of the most important causal parameters of interest is the average causal effect

of a treatment, as it represents a useful summary of the unit-level treatment effects[17]. Let E(.) denote the average value of the argument.

(2.9) The *average treatment effect (ATE)* of treatment $t \in T$ relative to treatment $c \in T$ is the expected value of the unit-level difference $Y_t(u)-Y_c(u)$ over all $u \in U$, i.e. $ATE_{tc}=E(UTE_{tc})=E(Y_t-Y_c)=E(Y_t)-E(Y_c)$.

The ATE is an unobserved quantity, since expectations of Y for both t and c are taken over the full range of U. In practice it is only possible to observe D(u) and $Y_D(u)$ over U, and therefore only the joint distribution of D and Y_D rather than D and $\{Y_t: t \in T\}$. The average value of the observed outcome Y_D among all those units actually exposed to a particular treatment $t \in T$ can be written as $E(Y_D|D=t)$. For the two particular treatments t and c this becomes $E(Y_D|D=t)=E(Y_t|D=t)$ and $E(Y_D|D=c)=E(Y_c|D=c)$, respectively. These two quantities are always observed in the data, and we can therefore define:

(2.10) The *prima facie average treatment effect (FATE)*[18] of a treatment $t \in T$ relative to a treatment $c \in T$ is the difference in average responses between those units actually exposed to t and those units actually exposed to c, i.e. $FATE_{tc}=E(Y_t|D=t)-E(Y_c|D=c)$.

The distinction between FATE and ATE emphasizes the fact that the quantity that we can always compute from the data (FATE) does in general not equal the quantity about which we desire to draw inferences (ATE). This results from the difference between $E(Y_t)$ and $E(Y_c)$ on the one hand and $E(Y_t|D=t)$ and $E(Y_c|D=c)$ on the other hand. The former are averages of Y over all of U and constitute ATE, while the latter are averages of Y over only those units in U actually exposed to t and c, respectively, and constitute FATE.

---

[17] In practice, further questions arise as to whether it is e.g. the "average treatment effect on the treated", or the "average treatment effect on the population" that is the causal parameter of interest. Cf. Heckman (1992) and Heckman, LaLonde, and Smith (1999) for discussion, and Angrist, Imbens, and Rubin (1996a) for more on the POM and identification of the "local average treatment effect" (LATE).

[18] This follows Holland (1988a) who calls this parameter "prima facie average causal effect FACE". It is not to be confused with a "prima facie cause" as defined by Suppes (1970) in his probabilistic theory of causation (cf. Suppes (1970) for the original definition and e.g. Sobel (1995) or Salmon (1980) for a discussion): Given two time values t and t* with $t<t^*$, the event c_t is a prima facie cause of the event e_{t^*} if $Prob(e_{t^*}|c_t)>Prob(e_{t^*})$, i.e. c temporally precedes e and is positively relevant to it. As Holland (1986) points out, the association between cause and effect defining a prima facie cause is indeed a causal effect under "certain conditions that have wide use in science", while on the other hand FACE "is not always a causal effect". This is also why I prefer the labeling FATE to FACE.

The two quantities are only equal when *independence* holds. Suppose that the determination of which treatment a unit is exposed to is statistically independent of all other variables, in particular the response function. Following – as common practice – Dawid's (1979) notation of independence using the symbol "$\perp\!\!\!\perp$", this can be written as $D \perp\!\!\!\perp \{Y_t : t \in T\}$. Then $E(Y_t|D=t)=E(Y_t)$ for any $t \in T$, and we have:

(2.11) If $D \perp\!\!\!\perp \{Y_t : t \in T\}$, then the *prima facie average treatment effect* of a treatment $t \in T$ relative to a treatment $c \in T$ is equal to the *average treatment effect* of t relative to c, i.e. $FATE_{tc}=E(Y_t|D=t)-E(Y_c|D=c)=E(Y_t)-E(Y_c)=ATE_{tc}$.

The independence assumption is the key point to the applicability of the model, as it allows us to draw inferences on the unobserved causal parameter of interest, the ATE, directly from the FATE, which we can always compute or estimate from the data.

Under which conditions is independence likely to hold? The most probable case we have in practice is that of a randomized experiment, in which – coarsely speaking – units are randomly assigned to different treatments, so that the initial population and the subpopulations in the treatments do not differ from each other *on average*. This makes (2.11) likely to hold, thus yielding the ATE from the FATE. Holland (1988a) describes the relation between randomization and independence as follows: Independence is an assumption about the data collection process, i.e. about the relation of D and Y over the population U, while randomization is a physical process that gives plausibility to the independence assumption in many important cases. For instance, if U were infinite, then the law of large numbers together with randomization would imply that (almost) every realization of D would be independent of $\{Y_t\}$. However, randomization does not necessarily make independence plausible in each and every case, as randomization does not assure that each and every experiment is "adequately mixed", but only that "adequate mixing" is probable (Leamer 1983). To take the simplest example, imagine that U consisted only of very few units. Then the plain physical act of randomization would not render the independence assumption plausible.

What does it mean when we talk about populations that do not "differ" from each other, and "adequate mixing" in randomized experiments? This becomes clear when we introduce other variables into the model. So far Y was the only variable measured on the units u – apart from the treatment indicator D. Let us now add a variable X to the model, where X can be real-valued or vector-valued. In principle, X(u,t) is defined on $U \times T$ and depends on

both u and t. However, there is a special class of X-variables that are of specific interest, as defined in Holland (1988a):

(2.12) X is a *covariate* if X(u,t) does not depend on t for any $u \in U$.

Holland initially calls this class of variables "attributes" (in Holland 1986), but converts to (2.12) as the preferable definition because it corresponds to the usual experimental usage. If we consider specifically the values the X-variables take on for units *prior to treatment*, then the X-variables are always covariates.[19] For a post-treatment concomitant, however, the possibility that X(u,t) does depend on t cannot be excluded and "must be decided" (Holland 1988a), and if this the case, then X is not a covariate in the sense of (2.12).[20] Randomization *on average* guarantees balancing of covariates – observable and unobservable – across subpopulations in different treatments, which in turn makes the independence assumption plausible, so that (2.11) holds and we can infer the ATE from the FATE. In the words of Rosenbaum and Rubin (1983): With "properly collected data in a randomized trial", X is known to include *all* covariates that are both used to assign treatments and possibly related to the response $\{Y_t\}$.

The introduction of covariates into the model becomes even more important in cases in which we do not have randomization and therefore cannot arrange the values of D(u) to achieve independence. In such an *observational study* we are still interested in inferring causal effects of treatments, but now – differing from a randomized setting – D is not automatically independent of $\{Y_t\}$. Given a (n observable) covariate (vector of covariates) X one could check the distribution of X for subgroups in each treatment by comparing the values of Prob(X=x|D=t) across the values of $t \in T$ (Holland 1988a). If there is evidence that Prob(X=x|D=t) depends on t, then the independence assumption may not appear plausible in the observational study. Instead, in the nonexperimental setting one usually builds on a weaker conditional independence assumption which says that treatment assignment and the response are conditionally independent given a vector of covariates:

[19] Note that this does not exclude unobservables a priori. In fact, it is difficult to express this feature. Holland (1988a) speaks of "variables measured on units prior to […] treatment" always being covariates, but I find that misleading, in a sense that a priori the definition of a covariate says nothing about whether the variable can be observed or not, and "measurement" implies observation. The point is: A pre-treatment variable is always a covariate, be it observable or not.

[20] Cf. Rosenbaum (1984) for a discussion of adjustments for a concomitant variable that has been affected by treatment.

(2.13) [Rosenbaum and Rubin 1983:] Treatment assigment is *strongly ignorable* if the response $\{Y_t : t \in T\}$ is conditionally independent of treatment assignment D given the observed covariates X, i.e. $\{Y_t\} \perp\!\!\!\perp D|X$, and $0<Prob(D=t|X)<1$.

Rosenbaum and Rubin (1983) show that (2.13) also holds for a balancing score B(X) defined as a function of the observed covariates X such that the conditional distribution of X given B(X) is the same for the exposure groups (D=t), i.e. $X \perp\!\!\!\perp D|B(X)$.[21] Rosenbaum and Rubin identify all functions of X that are balancing scores; the most trivial one being B(X)=X, and the coarsest one being the *propensity score*: The propensity score $Prob(D=t|X)=P_t(X)$ is of particular interest in practice, as it reduces the potential problem of conditioning on a high-dimensional X – if X is vector-valued – to conditioning on a scalar, provided that $P_t(X)$ is known.

Strong ignorability is the basis for all causal inference on covariate-adjusted treatment effects in observational studies (Holland 1988a). Adjusting for covariates yields the *covariate-adjusted prima facie average treatment effect*, or C-FATE[22], based on conditional expectations:

(2.14) $C\text{-}FATE_{tc}=E\{E(Y_t|D=t, X) – E(Y_c|D=c, X)\}$.

Just like the FATE, the C-FATE does in general not equal the desired ATE. This only holds under conditional independence:

$$\begin{aligned} C\text{-}FATE_{tc} &= E\{E(Y_t|D=t, X) – E(Y_c|D=c, X)\} \\ &= E\{E(Y_t|X) – E(Y_c|X)\} \\ &= E(Y_t)– E(Y_c) \\ &= ATE_{tc} \end{aligned}$$

This finding concludes the discussion of the POM for causal inference, as we have now discussed all relevant features of the theory as well as the circumstances under which the model can be applied in randomized trials and observational studies. For further discussion cf.

[21] In fact Rosenbaum and Rubin (1983) prove this property for the two-treatment case D={0,1}. The extension of this result to multivalued treatments is shown in Imbens (2000), and Lechner (2001a). The main result, however, is that of Rosenbaum and Rubin.
[22] "C-FACE" in Holland (1988a).

Rubin (1974, 1977, 1986), Holland (1986, 1988a), Holland and Rubin (1983, 1988), Rosenbaum (1984, 1995a), Rosenbaum and Rubin (1983, 1984a, 1984b), and Angrist, Imbens, and Rubin (1996a).

2.4 Comparing Possible Worlds

The program of causal inference is clear from the previous sections: to draw inference about the effect of some treatment t on some response variable Y. It is therefore necessary to establish a counterfactual state of the world – i.e. some other possible world – characterized by an alternative treatment c (with respective associated outcome) to which we can causally relate the treatment-t-world. We have seen that c could be either simply not-t or any other – distinct possible particular – treatment. In this section I will show what such possible worlds look like in the POM, and give some guidelines on the choice of appropriate alternative worlds for inferring and interpreting causal relations.

From the discussion above, in particular the definition of causation based on possible world relations, one could infer that something like "the quest for the closest possible world" is at the heart of the problem of causal inference in statistics. But this is not the case – at least not in a sense that we would have to compare multitudes of worlds and judge degrees of proximity between them. In fact, the POM simply defines closest possible worlds. Section 2.4.1 considers these ex ante defined worlds and examines relations between them with respect to meaningful causal interpretations. Subsequently, section 2.4.2 gives an account of proximity relations between worlds and discusses why we might not always be interested in the "closest" possible world, and under what circumstances this can be problematic.

2.4.1 Varieties of Counterfactuals

By definition the closest world to actuality is the one to which we relate the causal comparison: If we want to infer the causal effect of treatment t relative to treatment c, then we set the c-world as the closest world to the t-world. Were we instead interested in the effect of t relative to some other treatment c*, then we would establish the c*-world as closest to the t-world. This is clear from (2.8) and the explanation I have given in section 2.3. The idea that we can consider these worlds as being "close" to each other becomes clear in the experimental

context: The t- and the c-world are based on the same background – ideally represented by a large number of concurrent covariates – and they merely differ in the fact that in the t-world units are exposed to treatment t, while in the c-world units are exposed to c.

Let us slightly refine the discussion. First, call T_1 the world solely defined by treatment t_1 *and* associated outcome Y_{t_1}, T_2 the world defined by t_2 and Y_{t_2} etc., so that the M treatments $t_1,\ldots,t_m$ with associated outcomes $Y_{t_1},...,Y_{t_m}$ constitute the M subsets $T_1,\ldots,T_m$ within the set of treatment-worlds T. Second, define the causal effect of some treatment t_i relative to another treatment t_j as

$$\Delta_{t_i t_j} = Y_{t_i} - Y_{t_j} \ ,$$

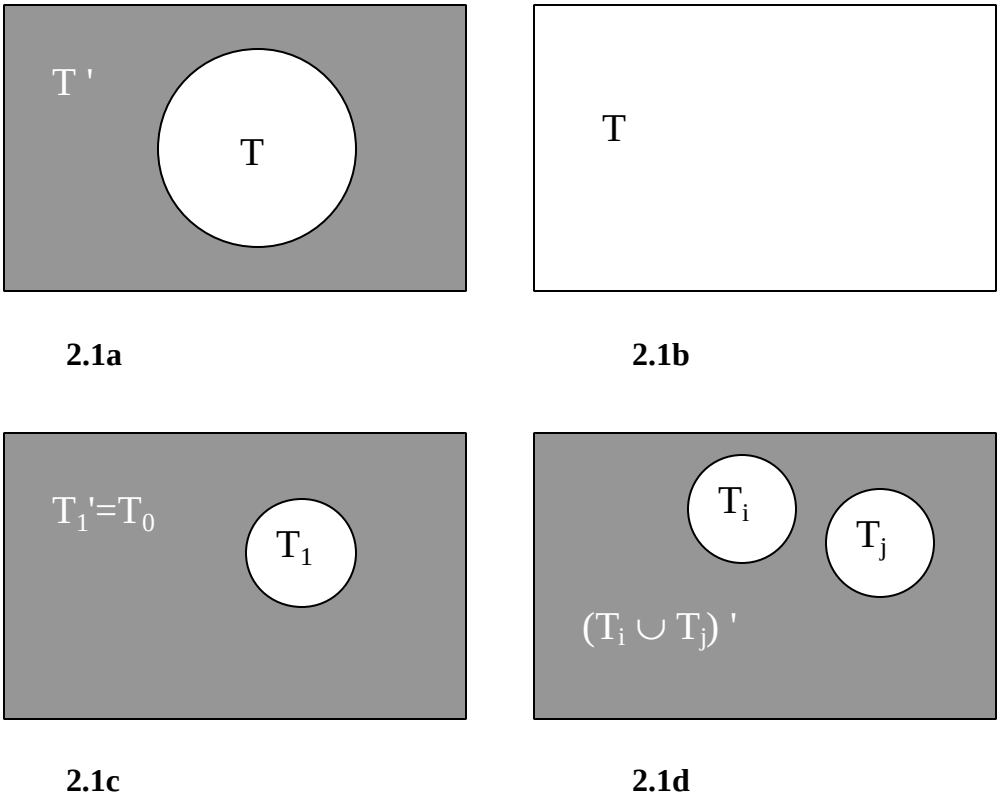
disregarding whether we are looking at unit-level or average treatment effects.

Let W denote the universal set comprising all possible worlds that differ only with respect to the characteristic "treatment and associated outcome", so that $T \subseteq W$.[23] In general, T does not need to equal W, if we regard T as comprising just those worlds where we can either control the types of treatment t_i or at least observe them, i.e. T is meant to comprise those worlds with well-defined types of treatment. The complement T' of T is then given from $W=T \cup T'$ and contains treatment-worlds we can neither control nor observe. For both groups, however, it is in principle possible to construct valid comparisons, and thus infer causal relations, as T' can always be defined recursively as "everything that is not T". The relationship is depicted in the Venn diagram in Figure 2.1a, where W is represented by the rectangle. In Figure 2.1a, $T=T_1 \cup T_2 \cup \ldots \cup T_m$ and $T_i \cap T_j = \varnothing$ for all $T_i, T_j \in T$, i.e. T is meant to consist of exactly M mutually exclusive treatment-worlds. Clearly, once more this captures the idea of distinct possible particular treatments – Each T_i is well-defined, there is no interference between the T_i worlds, and no unrepresented versions of treatment exist. The special case for $T'=\varnothing$, and thus W=T, is displayed in Figure 2.1b.

Let us consider the complement T' and what is meant by "everything that is not T". One could argue that "not T" is a well-defined treatment and should be included as one subset in T, yielding T=W. The distinction, however, illustrates the difference between what could be called a *controlled control treatment* and an *uncontrolled comparison treatment*.

Consider the case in which treatment can only take on two values, M=2, the classic treatment-t-versus-control-c setting, where the causal effect of interest is that of t relative to c. In a randomized medical trial, where t is the medicament under study and c is a placebo, c represents a controlled control treatment. It is (a) controlled by the experimenter, and (b) a distinct alternative treatment in its own right, which is not merely characterized by the absence of t, i.e. $T_c \neq T_t$'. Therefore, in this case, W={T,T'} with T={T_t,T_c} and T'=($T_t \cup T_c$)', where T' is some unspecified treatment outside T characterized by not given the medicament *and* not given the placebo. Of course, T' might not be of interest in the study, or we might not even be able to obtain any information about it, but nonetheless it is an implicit part of the study.

Figure 2.1 Possible treatment worlds in the POM.



2.1a **2.1b**

2.1c **2.1d**

On the other hand, consider the case of an observational study in labor economics, for instance, aiming at evaluating some government training program (=treatment t). In this case,

[23] This account is still in the framework of the POM of section 2.3, and thus considers a finite number of treatments. For an extension of the model to the case where the set of treatments is not finite see Pratt and

the "alternative treatment" c is characterized retrospectively by the absence of training, so that c represents an uncontrolled comparison treatment. It is (a) not under control of the researcher, and (b) not defined on its own, but just by the absence of t, i.e. $T_c=T_t'$. Therefore, W=T with $T=\{T_t,T_c\}$ or, equivalently, $W=\{T,T'\}$ with $T=T_t$ and $T'=T_c$.

Note that the distinction between "controlled control treatment" and "uncontrolled comparison treatment" is about the distinction itself, and does not imply that the one can be used for valid causal inference, and the other cannot. But it is important to note the difference. Clearly placebos are used to learn something about "not given the medication", and in that respect they may perform even better than "actually not given the medication", because with placebos the control units cannot be influenced by knowing that they are not given the medication. Use of placebos ensures that the response is to the treatment itself, not the idea of treatment. Hence, the controlled control treatment gives a well-defined alternative to t, while the uncontrolled comparison treatment necessarily remains more vague.[24] However, we will see that this need not be a disadvantage in interpreting results. It has to be noted, though, that if the control treatment is not well-specified, and the treatment shows no effect relative to the control treatment, then it might well be that the control treatment is or contains a pre-empted potential cause (cf. section 2.2.3) of the same effect, i.e. both treatment and control have the same causal effect on the response variable, which the causal comparison between the two cannot reveal.

In the next step, let us adopt the notion of treatment c meaning the absence of any treatment, be it controlled or uncontrolled, defined uniquely or recursively. Thus, denote $T_c=T_0$ and let $T=T_0,T_1,\ldots T_{m-1}$ with M-1 "real" treatments and the "null" treatment, W=T. Figure 2.1c depicts the case for M=2 and $T_t=T_1$, $T_0=T_1'$. As T_1 is the world with treatment t_1, and T_0 the world with treatment t_0:

(2.15) $$\Delta_{t_1t_1'} = Y_{t_1} - Y_{t_1'} = Y_{t_1} - Y_{t_0} = \Delta_{t_1t_0}$$

This is the basic case which almost all causal inference studies are based on. We have just two

---

Schlaifer (1988).

[24] Experimental settings do not necessarily imply a well-defined null treatment. While this is possible in medical experimental studies of the type described above, it is far more difficult in experimental studies in labor economics, e.g., due to the length of treatment (several months of participation in a training program) and the difficulty of defining a proper alternative (cf. later this section). One example is the experimental evaluation of the National Supported Work Demonstration (NSW) in the US: "Those assigned to the treatment group received

treatment-worlds differing only by the treatment under study, where the alternative world is characterized by the null treatment which equals the absence of treatment. This setting is intuitively appealing: The closest-world relation is obvious, and the interpretation of results is straightforward.

Let us extend this to the case where M>2, i.e. we have at least two "real" treatments besides the null treatment. Figure 2.1d illustrates the simplest case for M=3 under consideration and $T_0=(T_i \cup T_j)'$. The case of multivalued treatment has several important implications for interpretation. First, consider particular treatments t_i, t_j, t_k and the following decomposition:

$$
\begin{aligned}
\Delta_{t_i t_j} &= Y_{t_i} - Y_{t_j} \\
&= Y_{t_i} - Y_{t_k} - Y_{t_j} + Y_{t_k} \\
&= (Y_{t_i} - Y_{t_k}) - (Y_{t_j} - Y_{t_k}) \\
&= \Delta_{t_i t_k} - \Delta_{t_j t_k}
\end{aligned} \tag{2.16}
$$

Of particular interest is the special case where $t_k=t_0$.

$$\Delta_{t_i t_j} = \Delta_{t_i t_0} - \Delta_{t_j t_0} \tag{2.16a}$$

(Of course, if $t_k \neq t_0$ in (2.16) we can only use the decomposition if M>3). Expressions (2.16) and (2.16a) nicely show that any causal comparison between two treatments is implicitly always related to any other baseline-treatment within T. The case in which the null treatment is the baseline (2.16a) is of particular interest, since we have seen above that we are usually used to inferring causal effects relative to the null treatment. This relating of causal comparisons between two treatments – neither of which is the null – to the null treatment is also necessary to identify the level of effects.

For the M=2 case property (2.15) has shown that the causal comparison of some treatment t_i relative to the absence of t_i equals the comparison of t_i to the null. Unfortunately, this convenient feature does not hold for a causal comparison of t_i relative to t_i' in the case of M>2. There are two aspects to the t_i-versus-t_i' relation in this context. First, we have the basic

---

all the benefits of the NSW program, while those assigned to the control group *were left to fend for themselves*." (LaLonde 1986, my italics)

result that

(2.17) $\Delta_{t_i t_i'} \neq \Delta_{t_i t_0}$

because $t_i' \neq t_0$ and $Y_{t_i'} \neq Y_{t_0}$. This can be seen when we consider what the effect of t_i relative to t_i' actually is:

$$\begin{aligned} \Delta_{t_i t_i'} &= Y_{t_i} - Y_{t_i'} \\ &= Y_{t_i} - F(Y_{t_0}, Y_{t_1}, ..., Y_{t_{i-1}}, Y_{t_{i+1}}, ..., Y_{t_{m-1}}) \\ &= Y_{t_i} - \sum_{\substack{k=0 \\ k \neq i}}^{M-1} w_k Y_{t_k} \end{aligned} \tag{2.18}$$

where $\sum w_k = 1$ and, for instance,

(2.18a) $w_k = \overline{w} = \frac{1}{M-1}$ or (2.18b) $w_k = \frac{P(t = t_k)}{\sum_{r=0, r \neq i}^{M-1} P(t = t_r)} = \frac{P(t = t_k)}{1 - P(t = t_i)}$.

The causal effect of treatment t_i relative to t_i' as given in (2.18) is therefore the difference in outcomes under t_i and t_i' (first line), which equals the difference between the outcome under t_i and some function of the outcomes under all other treatments except t_i (second line), which could in an empirical application equal the difference between the outcome under t_i and the weighted sum of all other outcomes (third line). I will refer to the function of the outcomes under all other treatments in T_i' as the *absolute counterfactual* to treatment t_i, as it is a summary expression of all counterfactual possible worlds. Examples of weight functions for empirical work are given as (2.18a) equal weights, and (2.18b) the probability of exposure to a particular program (that is not t_i) relative to the sum of probabilities of exposure to any program that is not t_i.[25]

The second aspect to the t_i-versus-t_i' relation is that the complements to particular treatments cannot be used as a common baseline, i.e.

[25] This expression has been used, e.g., in Lechner (2001b). See also section 2.5.

(2.19) $\Delta_{t_i t_j} \neq \Delta_{t_i t_i'} - \Delta_{t_j t_j'}$

because clearly

$$\begin{aligned}\Delta_{t_i t_i'} - \Delta_{t_j t_j'} &= (Y_{t_i} - \sum_{\substack{k=0 \\ k\neq i}}^{M-1} w_k Y_{t_k}) - (Y_{t_j} - \sum_{\substack{l=0 \\ l\neq j}}^{M-1} v_l Y_{t_l}) \\ &= Y_{t_i} - Y_{t_j} - \sum_{\substack{k=0 \\ k\neq i}}^{M-1} w_k Y_{t_k} + \sum_{\substack{l=0 \\ l\neq j}}^{M-1} v_l Y_{t_l} \\ &= \Delta_{t_i t_j} - (\sum_{\substack{k=0 \\ k\neq i}}^{M-1} w_k Y_{t_k} - \sum_{\substack{l=0 \\ l\neq j}}^{M-1} v_l Y_{t_l})\end{aligned}$$

Table 2.1 presents an overview of different causal queries and the corresponding counterfactuals.

In the M=2 case, there are two possibilities, either (a) $t_0=t_1'$ or (b) $t_0\neq t_1'$. The first case (a) is the usual one, and applies for observational studies. The second case (b) comprises two possibilities depending on a relevance criterion. On the one hand, if $t_0\neq t_1'$, so that there exists a T' world besides $T=\{T_0,T_1\}$, and t_1' is considered *irrelevant* for some reason, such as t_0 being explicitly specified – like in an experimental study –, then this implies that T' is irrelevant. On the other hand, if one has reason to believe that $t_0\neq t_1'$ and if T' is *relevant*, then there are two further possibilities: Either (i) one has some usable information about T', then this converts to the M>2 case, or (ii) one does not have such information, which hints at a violation of SUTVA because there exist unrepresented versions of treatment. Usually (as in the agricultural setting of Neyman 1923 [1990]) one thinks of unrepresented versions of treatment as unrepresented versions of the "actual" treatment – in this case, however, T' comprises unrepresented versions of the null treatment.

For the M>2 case, as we have a variety of well-defined treatments, it makes sense to assume that we have a specific t_0 (even if it is defined via the absence of all other treatments) and thus W=T. Table 2.1 depicts some possible counterfactual comparisons. First, the causal effect of a particular treatment could be inferred relative to the null treatment. As in the M=2

Table 2.1 Varieties of Counterfactuals

Number of treatments in T	Treatment of interest	Counterfactual treatment	Causal effect	Interpretation / Notes
M=2	t_1	t_0	$\Delta_{t_1 t_0} = Y_{t_1} - Y_{t_0}$	The null treatment, in most cases the counterfactual of interest. Usually equals t_1', differs only if explicitly specified (as in experimental studies), or if SUTVA is violated.
		t_1'	$\Delta_{t_1 t_1'} = Y_{t_1} - Y_{t_1'}$	*Anything* that is not t_1. Usually applies in observational studies, where it equals t_0.
M>2	t_i	t_0	$\Delta_{t_i t_0} = Y_{t_i} - Y_{t_0}$	The null treatment, again the counterfactual of interest in most cases. Relevant as baseline.
		$t_j \neq t_i$	$\Delta_{t_i t_j} = Y_{t_i} - Y_{t_j} = \Delta_{t_i t_k} - \Delta_{t_j t_k} = \Delta_{t_i t_0} - \Delta_{t_j t_0}$	Any other particular treatment can be used as counterfactual, for interpretation important to note that the baseline (usually the null) is implicit.
		t_i'	$\Delta_{t_i t_i'} = Y_{t_i} - Y_{t_i'} = Y_{t_i} - F(Y_{t_0}, Y_{t_1}, ..., Y_{t_{i-1}}, Y_{t_{i+1}}, ..., Y_{t_{m-1}})$	*Everything* that is not t_i – the *absolute counterfactual*, the outcome of which is given as a function of the outcomes of all treatments except t_i

For M>2, as in the discussion in the text, assume that W=T.

case, this would be the causal question of interest in most cases. Second, one could construct the causal comparison of a particular treatment relative to any other treatment within T. In interpreting the effect it should then (a) be pointed out why this is considered to be a causal question of interest, and (b) be noted that any other treatment (besides the two we relate) can be used as baseline. The most relevant baseline is the null treatment, and in fact it should be considered in any case in order to identify the level of the inferred effect.[26] The third possible counterfactual for M>2 in Table 2.1 relates a specific treatment t_i to a function of the outcomes of all other treatments except t_i. This I labeled absolute counterfactual. It infers the causal effect of some treatment relative to (an appropriate combination of) all other alternative treatments. This could be a weighted average as given in (2.18). In a sense this is similar to the *t_1'=anything-that-is-not-t_1*-case for M=2, with the decisive difference that now it is "everything", not "anything", expressing the fact that all alternative treatments are well-defined – and that the corresponding outcomes can therefore be appropriately weighted in an empirical study. With respect to the absolute counterfactual, it can be of particular interest to compare the null to the summary over all other treatments to infer whether the introduction of the overall set of treatments yielded any positive response.

Finally, it should be noted that one could of course construct many more counterfactuals. For instance, one could use causal relations between treatments as a baseline for causal relations between other treatments, or construct the comparison between a particular treatment and a weighted combination of some, but not all of the alternative treatments, etc. That, however, is pure mechanics, and I suppose it will be difficult – though not impossible – to unfold the exact causal interpretations of such counterfactuals.

2.4.2 Illustration

This short subsection entails a few examples that further illustrate some of the ideas unfolded in the previous section, and shows why we need a clear conception of the T worlds for causal inference.

> **Example E.** In the M=2 case, why can it can be insightful to distinguish a known or well-defined no-treatment state ($t_0 \neq t_1'$) from a no-treatment state defined merely by the absence of treatment ($t_0 = t_1'$)? Imagine some researcher planning to evaluate some

[26] If the effect of t_i is positive relative to the null, and the effect of t_j is negative relative to the null, then the effect of t_i is strongly positive relative to t_j. Looking only at the last effect does not reveal the negative effect of t_j relative to the null. Similarly, the effect of t_i relative to t_j could be positive, but still the effects of both of them could be negative relative to the null.

government training program for the economically disadvantaged in a nonexperimental setting. She constructs a retrospective comparison group defined by not having participated in the program. However, training usually takes time. Assume an average of 2 months in this example. What did comparison group units do during that time? Remain unemployed, continue job search, do nothing, take private training course, etc.? Maybe some of that, maybe all of that, maybe none of that. In most cases, the data doesn't tell. Thus, as it is impossible to open this black box, one needs to make some assumption about the comparison treatment. It is then fairly convenient to define the no-treatment state as just that, the absence of the treatment under study. The causal effect is that of the training program relative to any other possible (but unobserved) alternative action the program participants would have engaged in had they not participated. Clearly, this is quite different from the explicit specification of the no-treatment state in an experimental medical study (t_0=placebo).

Example F. Consider the problem of compliance. For instance, in a long-term medical study one could in principle distinguish four groups: those assigned to treatment who are good compliers, those assigned to treatment who are poor compliers, those assigned placebos who are good compliers, and those assigned placebos who are poor compliers. In principle this defines four different treatments, and only the randomized comparison gives the correct inference. Cf. Rosenbaum (1995b) for a discussion, and Angrist, Imbens and Rubin (1996a) for more on compliance.

Example G. An observational study by Larsson (2000) evaluates labour market programs in Sweden. In the study M=3, and the treatments are Youth Practice (YP), Labor Market Training (LMT), Non-participation (=Null). In personal communication with the author the interpretation was given that the null treatment comprises a state of job search rather than non-participation. This finding has several implications: (a) If one has usable information to distinguish job searchers from non-participants, this converts to a case of M=4 with treatments YP, LMT, job search, non-participation (=Null). (b) If in fact all individuals in the null treatment are in job search, this changes the counterfactual question, and the causal inference is on the effect of YP (or LMT) relative to job search, and not relative to non-participation. (c) If the null treatment comprises both individuals in job search and non-participants, this hints at a violation of SUTVA.

2.4.3 Comparative Similarity

Much of what has been said in the previous sections referred to notions of possible worlds, to entities that exist parallel to – or in addition, or besides – something we referred to as actuality, and we viewed these possible worlds in terms of similarity or comparability. Although I am convinced that the intuitive conceivability of this concept of actuality and "surrounding" possible worlds is straightforward, I will discuss a few inherent aspects in this section. More about the foundations of this concept can be found in Lewis (1973a, Ch4), in which, for instance, he replies to a fictitious questioner asking him what sort of thing possible worlds are [Lewis' italics, my underscoring]:

> I can only ask him [the questioner] to admit that he knows what sort of thing our actual world is, and then explain that other worlds are more things of *that* sort, differing not in kind but only in what goes on at them. Our actual world is only one world among others. We call it alone actual not because it differs in kind from all the rest but because it is the world we inhabit.[27]

The POM picks up this idea in that the treatment worlds $T_i \in T$ do certainly not differ in kind from each other, but only in the treatment by which they are defined. In fact, the treatment worlds T_i are defined to differ from each other in exactly two aspects: the treatment t_i that "goes on" at each (distinct possible particular) treatment world T_i, and the outcome Y_{t_i} associated with t_i on world T_i. Even though the treatment worlds T_i coexist, they only represent potentialities: Recall that treatments are defined on units u, so that we actually have the two differing features $t_i(u)$ and $Y_{t_i}(u)$ on each $T_i(u)$. However, for each u only one particular $T_i(u)$ is realized. This realized $T_i(u)$ represents actuality. From the point of view of actuality, the other treatment worlds are entities that might be called "ways things could have been" (Lewis 1973a). These he calls *possible worlds*.

In applying the POM we do not search for possible worlds. The treatment worlds T_i are assumed to be, and defined to be the possible worlds. And the treatment world $T_j \neq T_i$ to which we relate treatment world T_i is defined to be the closest possible world to infer the causal effect of treatment t_i relative to treatment t_j. Let us examine this a bit further and return to the example with Clark Glymour's uncle Schlomo from section 2.3.

> **Example D [ctd].** In the actual world Schlomo smokes 2 packs of cigarettes a day. It has been argued that the closest possible world to that actual world might be the one in which Schlomo smokes 3 packs day. Nonetheless, we choose to define the world in which he does not smoke at all as the closest world. Thus we infer a causal comparison between actuality with Schlomo smoking 2 packs a day – associated with the outcome "contracting lung cancer" – relative to the closest possible world in which he does not smoke at all – associated with the outcome "not contracting lung cancer". At first sight this appears to be an easy solution. But note that in this causally relating an actual world to a closest world by definition it is implicitly assumed that *no other element* than the one we either manipulate (in an experiment) or study (in an observational study) and the outcome associated with this element differs. This is just the *ceteris paribus* clause of economics. In the example, the element that differs is the reduction in packs of cigarettes a day from two to zero. But in the zero-world it may be that (a) Schlomo takes the healthy road and stops drinking and starts working out a lot, or that he takes on even worse compensatory vices instead, and e.g. starts taking

[27] Cf. also section 2.2.1 and footnote 7.

> cocaine. Moreover, underlying changes in the zero-world may be that (b) the quitting makes Schlomo lazy, silent, unmotivated, or maybe vivid, lively, energetic instead. Elements (a) refer to observable differences, and elements (b) to unobservable, and the examples show that both (a) and (b) could point into either a positive or negative direction in health terms. For a causal comparison of the actual world relative to the defined closest possible world it is necessary to either control for these potential changes or to ensure that the assumption that these differences equal zero appears credible. Clearly, in the example with Schlomo neither seems very likely.[28]

This is where the proximity relation enters: The 3-pack-world may be closer to the 2-pack-world in a sense that it is more likely that all other factors are the same. But still it may not be the alternative world that we are interested in for inferring a causal relation. Therefore, the causal effect (on some outcome) of some treatment t_i relative to some alternative treatment t_j is based on T_j being (i) the counterfactual world of interest relative to T_i and (ii) the closest possible world by definition. Closeness has to be ensured either by assuming proximity, i.e. on plausibility grounds, or by controlling for background factors establishing that they are the same in both T_i and T_j, in particular those that could potentially influence the outcome.

Finally, one could proceed to discuss distance measures between possible worlds. Thinking of closest possible worlds, this discussion comes up naturally: If M>2, then there is the actual world and at least two alternative worlds. Now which one of the alternative worlds is closer to actuality? This line of thought is a bit misleading, because actually – as shown above – we would condition on background factors (or plausibility grounds, if these can be conditioned on) to ensure that the possible worlds are equidistant. If T_i and T_j differ only in elements t_i/t_j and Y_{t_i} / Y_{t_j}, and T_i and T_k differ only in elements t_i/t_k and Y_{t_i} / Y_{t_k}, then this is also true for T_j and T_k, and the three worlds are pairwise equidistant. In practice, this argument would hold in an experiment with randomized controlled exposure to a set of treatments. In an observational study, however, differences between groups of units across treatment worlds do arise. Therefore, distances between treatment worlds can in principle be measured by appropriately weighting the background factors, or calculating weight functions such as (2.18b) based on propensity scores for each treatment world (cf. also section 2.5). I will not pursue this discussion any further here, but conclude with Lewis' observation on the constructability of such proximity measures (Lewis 1973a, his italics):

[28] The predominant difficulty in this example is the duration of treatment (=smoking a certain number of packs of cigarettes a day), which goes on for decades. How could you possibly control for background factors, observable or not, over such a long time period? – In any case, I think that this example is easier to reconcile with chancy counterfactuals: Given that Schlomo had not smoked 2 packs of cigarettes a day, the probability of his contracting lung cancer would have been (much) lower.

> We could, however, define exact distance measures […] for […] constructions of ersatz worlds. At worst, we might need a few numerical parameters. For instance, we might define one similarity measure for distribution of matter and another for distribution of fields, and we would then need to choose a weighting parameter to tell us how to combine these in arriving at the overall similarity of two worlds. All this would be easy work for those who like that sort of thing, and would yield an exact measure of *something* – something that we might be tempted to regard as the similarity 'distance' between worlds.

Clearly, this is about distance measures between any two worlds. In the POM we depart from actuality and look for that alternative treatment world that sets all these differences to zero, except for the treatment and its associated outcome.

2.5 Practical Considerations

The following section briefly discusses some specific problems that might arise in empirical work, in particular in observational studies. For causal contrasts, in the POM it is often assumed that the N units are exposed to the M treatments at equal shares. In practice this is unlikely to hold, even in a randomized experiment. And while in a randomized experiment this does not necessarily influence causal comparisons between two treatments – because subsamples are still balanced –, it would indeed affect the absolute counterfactual: Participation probabilities are no longer the same, and therefore the assumption of equal weights is unrealistic. Still, participation probabilities in a randomized experiment would be known: This problem is more severe in an observational study when the null treatment group as a comparison group is unknown and has to be constructed. This usually implies having to estimate the participation probabilities.

The absolute counterfactual of (2.18) as a summary measure for the effect of some treatment relative to all other treatments can also be represented using a weighted aggregate of the pairwise causal comparisons between the particular treatment and all other treatments:

$$(2.20)\quad \Delta_{t_i t_{i'}} = Y_{t_i} - \sum_{\substack{k=0 \\ k \neq i}}^{M-1} w_k Y_{t_k} = \sum_{\substack{k=0 \\ k \neq i}}^{M-1} w_k (Y_{t_i} - Y_{t_k}) = \sum_{\substack{k=0 \\ k \neq i}}^{M-1} w_k \Delta_{t_i t_k}$$

This expression retains the causal interpretation of the effect of treatment t_i relative to the hypothetical state of random exposure to any other program that is not t_i. Lechner (2001b) uses this expression and calls it the *composite treatment effect*. Furthermore, Lechner (2001b) shows that in an applied observational study it does indeed make a difference whether one assesses t_i' as $T'=T_0$ using a binary probability model or t_i' as $T'=\{T_0 \cup T_1 \cup ... \cup T_{i-1} \cup T_{i+1} \cup ... \cup T_{m-1}\}$ using a multinomial probability model (and then equations (2.18) or (2.20)). The first results in an insufficient specification of the alternative state by aggregating groups into one alternative group without taking into account the different composition of subgroups, while the second appears to correctly disentangle the desired absolute counterfactual. This finding emphasizes the importance attributed to the definition of T' in section 2.4.1.

Note, though, that this is a problem arising *in practice*. In theory – or in an ideal randomized experiment – the calculation of the absolute counterfactual in the multinomial case would equal the T versus $T'=T_0$ in the binomial case, as it captures all relevant alternatives to a particular T. This holds even if participation probabilities differ across subgroups. In an applied observational study, however, this does not hold, because group compositions do differ, because binomial and multinomial probability models would yield different participation probability estimates, and because the multinomial case compares treatments pairwise (and the overall equivalence above would require a common support of covariates over all subsamples). This is unlikely to be achieved in an observational study, also because in practice heterogeneous programs are aimed at heterogeneous groups.

Finally it has to be noted that even though these two causal comparisons should be the same in theory but differ in practice, they can nonetheless be calculated. However, a meaningful causal interpretation may be difficult to derive (cf. also Lechner 2001b). This again points to the fact that it may not be a problem to mechanically produce various causal comparisons in mechanically ensuring proximity between worlds, but that it may well be a problem to give these comparisons a clear-cut causal interpretation.

2.6 Conclusion

This chapter has tried to add clarity to the understanding, applying and interpreting of the potential outcome model for causal inference commonly used in statistics and econometrics.

At the outset we have found that there are three predominant approaches at modeling causation in the empirical sciences: SEM, POM, and DAG. Being the model of particular interest in evaluation research, the chapter has focused on the POM and its inherent counterfactual nature. In order to clarify what is actually meant by counterfactual statements, this chapter has presented the main elements of the counterfactual account of causation in terms of Lewis's possible-world semantics. This included the basic notions of counterfactual logic, and some of the problems associated with philosophical approaches to causation in general, and the counterfactual approach in particular. I have pointed out that the pivotal notion of Lewis's account is that of "closest possible worlds".

The chapter has then proceeded to explicitly reformulate the potential outcome model for causal inference using counterfactual conditionals. The main ingredients of the POM – such as SUTVA – have a straightforward and elucidating representation in terms of counterfactual events and their truth conditions. I have discussed various causally meaningful counterfactuals that arise in applications of the potential outcome model with a finite number of treatments, and illustrated these using a simple set-theoretical framework. The main result in this respect was that the notion of closeness and proximity between possible worlds is an inherent part of the statistical model, yet one that is implicitly used and taken care of. Causal comparisons in the POM *a priori* assume that possible worlds differ only with respect to the particular treatment and the associated response. However, one has to be aware of the fact that mechanical productions of proximity do not necessarily generate clear-cut causal statements.

Chapter 3

Can Training and Financial Incentives Combat European Unemployment?

Together with Christoph M. Schmidt

Abstract. Training programs and the creation of financial incentives are prevalent instruments of Active Labor Market Policy throughout the European Union and its neighboring economies. Yet, by contrast to the tradition of program evaluation in the US, in Europe the thorough evaluation of their impact is still in its infancy. This chapter shows how the desolate state of European labor markets in the 1990s led EU policy makers to initiate a European Employment Strategy that has become known as the Luxembourg Process. Part of this strategy is to combat unemployment by means of active labor market programs. We demonstrate that – separate from the Luxembourg Process – academic evaluation research has developed and implemented adequate tools to assess program effectiveness with confidence. This finding is illustrated by a review of recent state-of-the-art evaluation studies across Europe. Various messages for the design of economic policy can be extracted from the available evidence. Training seems to be the most promising program – if there is any – and public sector programs fare substantially worse than private sector programs. Major distortions in program effectiveness emerge if benefit receipt eligibility is renewed after ALMP participation. However, in terms of the European Employment Strategy it is surprising that the Luxembourg Process seems to be largely disconnected from academic evaluation knowledge. We strongly recommend policy makers to include independent researchers in the effort of evaluating labor market policy interventions.

3.1 Introduction

During the last century, the US have acquired an extensive experience with public sector programs aiming at improving the economic situation of disadvantaged individuals. Since labor market programs tend to bind substantial economic resources, a careful scientific evaluation of these endeavors has been a long-standing concern of US policy makers and the American public. By contrast, such measures of *Active Labor Market Policy (ALMP)* as instruments to combat unemployment and poverty, and their respective evaluation, are a rather recent phenomenon in Europe.

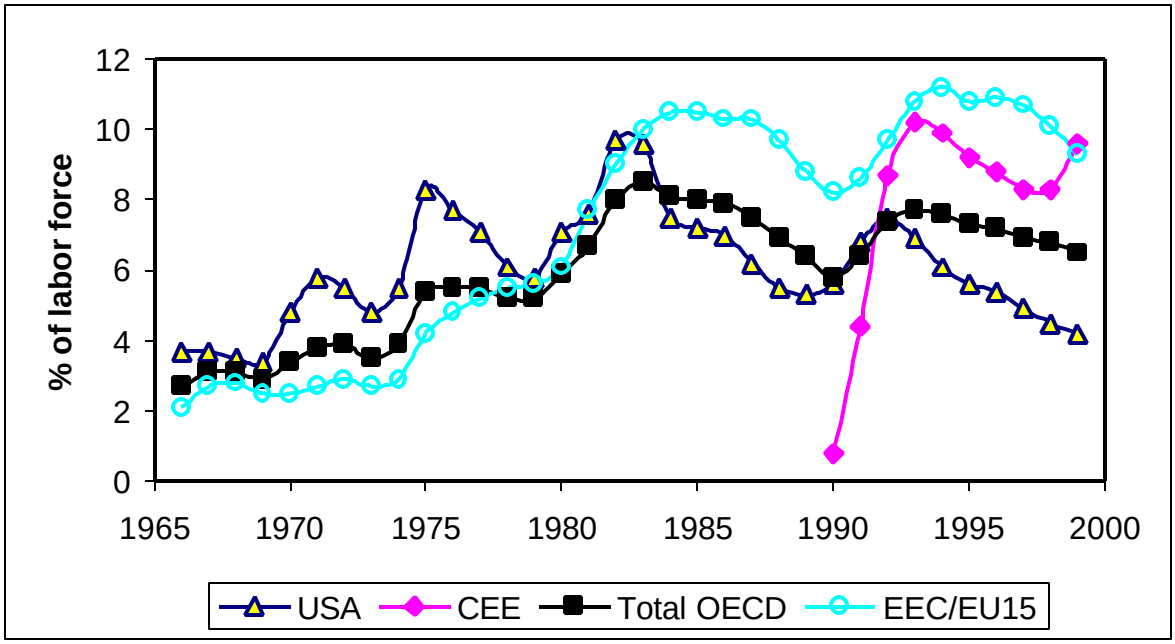
For Western European countries this mainly reflects the strong economic performance of the post-WWII era – they have experienced high and persistent unemployment rates only since the 1980s or 1990s. Similarly, the formerly socialist countries of Central and Eastern Europe – now being in the transition to modern market economies, many of whom are knocking at the door to the *European Union* – did not have to address large-scale *open* unemployment before the demise of the socialist regimes around 1990. Therefore, it has been only over the last one or two decades that combating high unemployment, and in many countries the particularly severe youth unemployment, has without doubt become one of the most urging policy issues across Europe.

Figure 3.1 depicts the development of unemployment rates since 1966 for countries of the European Union (EU, formerly EEC), transition countries of Central and Eastern Europe (CEE), the US, and the total of OECD countries. While EU unemployment had been slightly below or around the OECD average until 1981, since then it has persistently exceeded the OECD average, and since 1984 this difference has always exceeded two percentage points. US unemployment displays an opposite development: Whereas the unemployment rate was above OECD average until the early 1980s, it has been decreasing substantially since then, lying below OECD total most of the years, and always since 1993. As described above, European transition countries enter the picture only in the early 1990s, when unemployment rates skyrocketed to levels clearly above OECD average.

Looking at this bleak situation, European policy makers felt the pressure to react. In 1997 the *EU Commission* started what has become known as the *Luxembourg Process* – the formal recognition of employment issues as one of the key aspects of EU economic policy in the *Amsterdam Treaty*, and thus a matter of common concern pointing to a joint *European Employment Strategy*. Central elements of this strategy are the annual declaration of

employment policy guidelines by the EU Commission that are then followed by the formulation of *National Action Plans* according to these guidelines.

Figure 3.1 Unemployment Rates 1966-1999



Source: OECD (2000c), OECD (1998), OECD (1991).
OECD total=unweighted average, CEE=Hungary, Poland, Czech Republic, for 1990 and 1991 Czech Rep. only.

As one measure to combat high unemployment, most European countries entertain programs of Active Labor Market Policy. ALMP can be broadly classified into *training programs*, such as classroom training, on-the-job training, work experience, or job search assistance, *wage subsidies* to the private sector, i.e. subsidies to employers or financial incentives to workers, or *provision of jobs* in the public sector. Many Western European countries spend a considerable share of their budget on these measures. The transition countries copied much of the design of their benefit systems and ALMP regulations from Western Europe – even though the effectiveness of the programs is still at question.

Thus, there is a high demand for a better understanding of European labor markets and of the impact of public sector programs on participants and labor markets at large. In addition, due to substantial heterogeneity across European labor markets, it remains unclear what any

one country can eventually learn from experiences made in any other country – in an economically integrated Europe it is therefore imperative to collect evidence throughout all its constituent economies. The coarse but crucial distinction between the US on one side and Europe on the other should not be confused with a uniformity of program effects across European economies. While many of the particular evaluation questions asked are of a distinct European character – in particular in their emphasis on youth programs, or the focus on employment instead of earnings –, the programs having been implemented across Europe and their respective effects are very heterogeneous.

It is certainly correct that ever since the first introduction of labor policy measures into European economies there has been ongoing research on their evaluation. However, mainly due to a lacking communication between the scientific community and those financing, designing, and implementing public sector programs – even after introducing the Luxembourg Process – we find this research to be – figuratively speaking – still in its infancy. Many methodological issues remain unresolved, and the data often do not live up to the sophisticated econometric techniques imposed upon them. Most importantly, social experiments – which were able to generate a considerable body of valuable evidence on North American labor market programs – are still the exception. In consequence, nonexperimental data are highly prevalent in European evaluation studies, with the unsurprising result that the conclusions are often far from being clear-cut.

This chapter elucidates the conclusions that *can* be drawn regarding ALMP effectiveness in Europe on the basis of the current state-of-the-art of evaluation research. We embed this analysis in the context of the Luxembourg Process, which – as we will show – could use a stronger connection to scientific evaluation practice if it wants to properly pursue its goals. On the one hand, we provide some diagnostics on the desolate state European labor markets have dwelled in over the last decade, and show how this led the *policy side* to initiate a European Employment Strategy and the Luxembourg Process. On the other hand, we sketch recent advances in evaluation research and show how the *science side* has developed adequate tools to answer evaluation questions with confidence. It will be clear from our account that a closer connection of the two sides would be a great leap forward.

In much of what follows we do not further distinguish between the EU, OECD Europe, the "Euro Zone" etc., as we believe that in most cases such a distinction is unnecessary and would make the discussion overly complicated. Therefore – unless mentioned otherwise – we use the EU as a synonym for the general idea of "Western Europe"

(thus including Norway and Switzerland) for sake of the argument and contrast with the US, and within Europe merely contrast it with the transition countries of Central and Eastern Europe. We are aware of the fact that it has been argued that "while it is sometimes convenient to lump all the countries of western Europe together in order to provide a suitable contrast to North America, most of the time it is a rather silly thing to do" (Nickell 1997). Not only are we conscious about substantial intra-European differences, but we also do believe that with respect to the evaluation of ALMP – ultimately the theme of this chapter – a contrasting of the US and Western Europe is not merely a convenient but clearly the correct thing to do.

Section 3.2 delineates the status quo of European labor markets. We begin with some stylized facts on the economic environment, and proceed to show how this led the EU to introducing the Luxembourg Process. We discuss this Process in some detail and connect it with the current role of ALMP in Europe. Section 3.3 offers some intuitive reasoning on why labor market programs can in principle be an effective policy tool in combating unemployment. Section 3.4 describes and contrasts the received empirical evidence on ALMP effectiveness for the US and Europe. Here we also discuss some methodological issues that emerge in the undertaking of program evaluation. Section 3.5 presents a detailed yet selective review of recent state-of-the-art European evaluation studies organized along coarse regional groupings. Specifically, we will discuss singled out examples from (i) the Nordic countries, (ii) the UK and Benelux, (iii) Central Europe, and (iv) the CEE transition countries. This review contains a quite intriguing variety of program types implemented and estimation methods applied. Section 3.6 gives an overview of the empirical results. Section 3.7 concludes with a discussion of the general lessons for economic policy arising from the available evidence.

3.2 The Status Quo of European Labor Markets

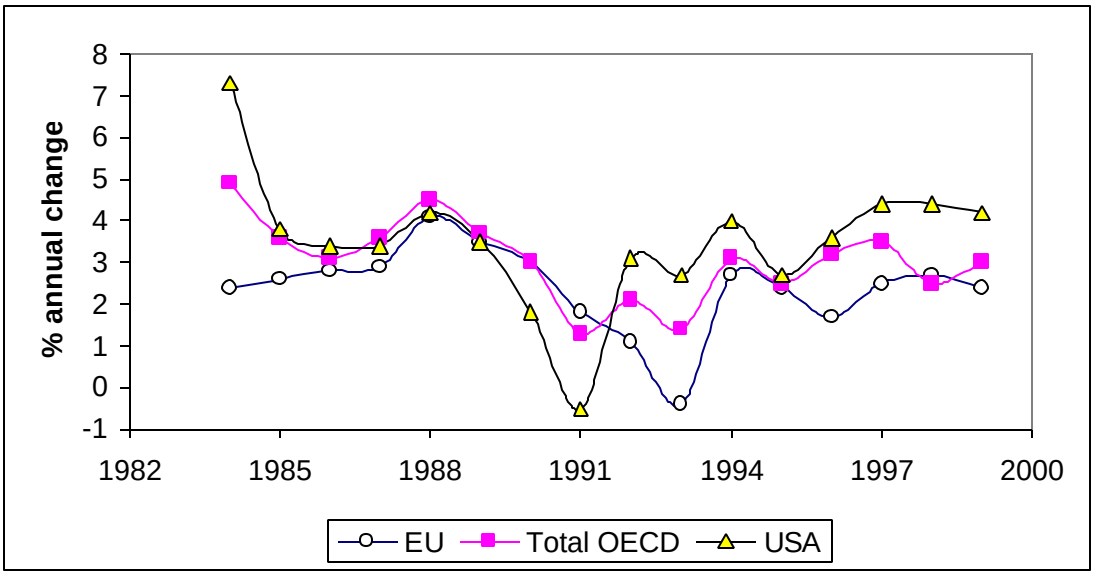
While section 3.1 presented a snapshot at the current situation, in this section we take a somewhat closer look at various aspects of European labor markets that are of interest for our study. First, we look at the more general European employment situation, and then try to identify what the Luxembourg Process intends to change or improve about this situation. As measures of ALMP to fight unemployment are of predominant interest also in the

Luxembourg Process, we then proceed to characterize their current role in European employment policy.

3.2.1 Diagnostics

What is the status quo of European labor markets? For a glance at the economic environment, Figure 3.2 displays annual growth rates of real GDP since 1984 for the US, the EU, and the total of OECD countries.

Figure 3.2 Real GDP Growth 1984-1999

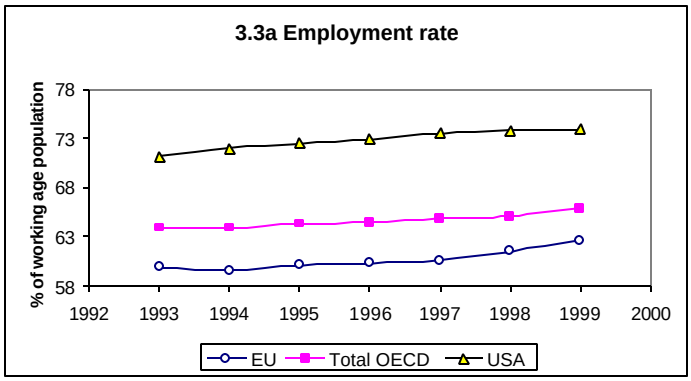


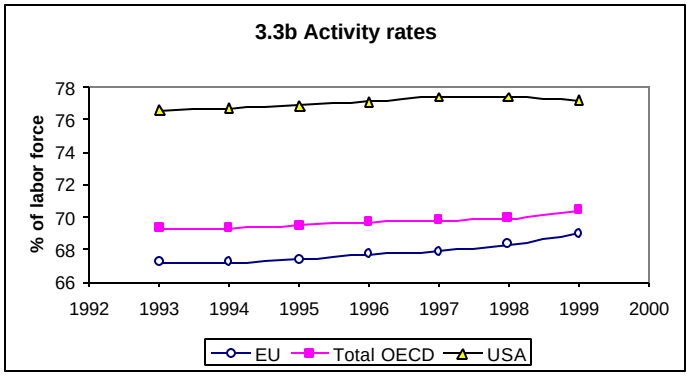
Source: OECD (2000d) Annex Table 1.

We observe a similar declining trend for all three series from the mid-1980s to the early 1990s, with respective dips down to negative growth rates for the US in 1991 and the EU in 1993. Since then growth rates seem to have recovered. However, in line with our observations regarding the unemployment rate series (Figure 3.1), GDP growth since 1995 has been above OECD average for the US, and below average – with the sole exception of 1998 – for the EU. OECD predictions for the years 2000-2002 suggest that this trend will continue, although both series are predicted to come closer to the OECD average, the US from above, and the EU from below (OECD 2000d, Annex Table 1).

Looking at some aggregate labor market indicators, we also find different sides to the same story. Panels a and b of Figure 3.3 show the development of employment rates and activity rates, respectively, for the US, the EU, and the OECD overall, from 1993 to 1999. Again, the US are clearly above OECD average in terms of employment rates and labor force participation, while the EU countries lie below the OECD total. Furthermore, there appears to be little change, at least not in this short time series. The only aspect worth noting is a slight but apparent upward trend in both employment and activity rates for EU countries in recent years. According to OECD predictions this trend is likely to continue (OECD 2000d, Annex Tables 19 and 20).

Figure 3.3 Labour market indicators

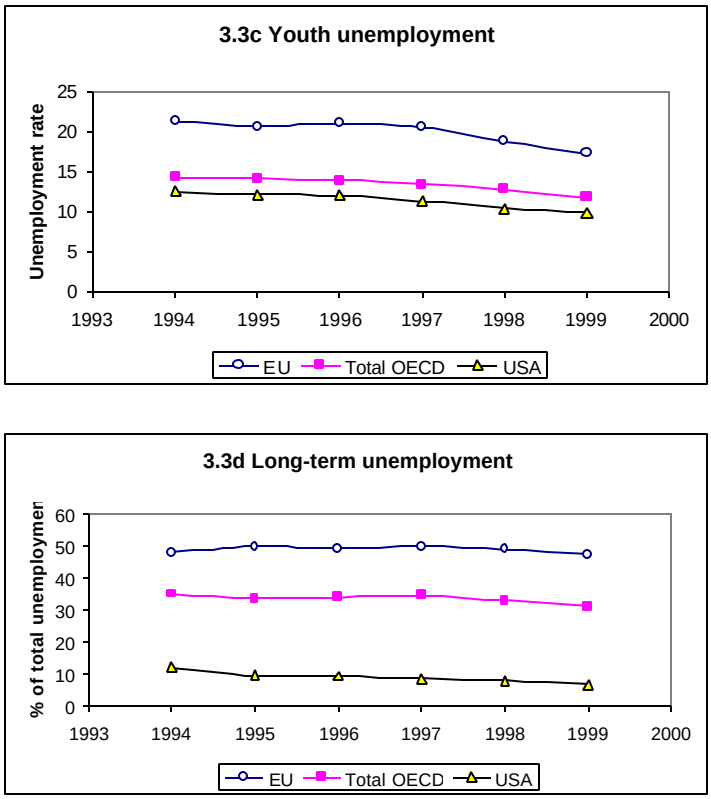




A corresponding picture emerges from plotting unemployment rates for the young (15-24 years old) and the long-term jobless (here: more than 12 months) among the unemployed. Panels c and d of Figure 3.3 underscore the difference between US and EU labor markets in illustrating that youth and above all long-term unemployment is far less of a problem in the

United States than in the European Union. The high incidence of long-term unemployment is widely regarded as one of the distinct and most serious problems of European labor markets – cf. Machin and Manning (1999) for an analysis of the causes of long-term unemployment and the correlation between high unemployment and high long-term unemployment. For EU countries in Figures 3.3c and 3.3d, one might expect a downward trend in youth and long-term unemployment corresponding to the upward trend in employment and participation rates of Figures 3.3a and 3.3b. Such a trend is – if at all there – less pronounced.

Figure 3.3 [ctd]



Source: OECD (2000b) Tables B,C,G, OECD (1998) Tables B,C,G.
3b: OECD (2000d) Annex Table 18 reports different activity rates for US (and thus for total OECD) due to different base (>15 years rather than 15-64)

As a diagnostic result, we find that over the last decade European economies display stagnant growth accompanied by low employment rates and low activity rates as well as high unemployment, in particular long-term unemployment. “Low” and “high”, that is relative to

the OECD average and above all the US. Unsurprisingly, unemployment has become the most feared and most severe problem in European economies. It has been argued that structural unemployment arises from the gap between the pressure on economies to adapt to change and their ability to do so (OECD 1994). OECD economies are found to be inadequately equipped and thus unable to cope with ongoing restructuring from manufacturing to service industry, with adoption of new information technologies, and with a rapidly changing international economy. This is also the basic conclusion of Ljungqvist and Sargent (1998) who analyze the "European unemployment dilemma" in the framework of a general equilibrium search model and who strongly argue for reforming of benefit systems, i.e. the design and interplay of active and passive measures of labor market policy. In their model, a welfare state with a very generous entitlement program is a virtual "time bomb" (p.546) waiting to explode.

3.2.2 The Luxembourg Process

Facing the pressure of labor markets in decline, the *Amsterdam Treaty* – which EU member states agreed on in June 1997 – introduced a new title on employment. This new *Employment Title* for the first time gave explicit recognition to the fact that employment issues have a status equal to that of other key aspects of EU economic policy. This is regarded as a crucial step in what has been called the *European Employment Strategy*. While the Amsterdam Treaty recognizes that the primary responsibility for design and implementation of employment policies resides at member state level, it emphasizes that "member states [...] shall regard promoting employment as a matter of common concern and shall co-ordinate their actions" (Article 2). Furthermore, in the Treaty the Union commits itself to a high level of employment as an explicit goal: "The objective of a high level of employment shall be taken into consideration in the formulation and implementation of Community policies and activities" (Article 3).

The term *Luxembourg Process* results from the fact that it was at the Luxembourg Jobs summit in November 1997 when member states decided that this European Employment Strategy should be built on thematic grounds, grouped in four *pillars* and described in *Employment Guidelines*. The four pillars of the strategy are (I) Employability, (II) Entrepreneurship, (III) Adaptability, and (IV) Equal Opportunities. The procedure is as follows: Each year, the Commission and the Council formulate employment guidelines within each pillar. These guidelines then are translated into *National Action Plans* (NAPs) for employment by the member states. In turn, these NAPs along with labor market developments

in each country over the year are then analyzed by the Commission and the Council and result in an annual *Joint Employment Report*. In the next step, the findings of the Joint Employment Report constitute the basis for reshaping the guidelines and country-specific recommendations for member states' employment policies for the following year.

Box 3.1 The Luxembourg Process – Employment Guidelines for 1998

The following overview of central guidelines is taken from European Commission (1998). These guidelines call member states to "undertake concrete action to attain the following objectives":

Pillar I – Employability
- Implement a preventive approach so as to reduce significantly the inflow of young and adult unemployed persons into long-term unemployment [Guidelines 1,2]
- Shift people from welfare dependency to work and training, namely through a more active labour market policy [3]
- Develop partnership as a framework for the provision of training and lifelong learning [4,5]
- Facilitate the transition from school to work [6,7]

Pillar II – Entrepreneurship
- Promote job creation in the social economy and at local level [10]

Pillar III – Adaptability
- Encourage the development of in-house training and investment in human resources [15]

Pillar IV – Equal Opportunities
- Tackle gender gaps in employment and unemployment [16]
- Facilitate reintegration into the labour market [18]

To illustrate the Luxembourg Process initiative, Box 3.1 gives an overview of those initial 1998 objectives – based on employment guidelines – that are of major interest for the purposes of our chapter. In particular, the first three guidelines focus on the tackling of youth and long-term unemployment as well as the restructuring of the benefit system in moving from passive to active measures. It is interesting to note that these three are the only ones in the set of guidelines that are formulated using concrete figures as objectives, rather than only subjective measures and/or unspecified declarations of intent. To quote from the original 1998

guidelines (European Commission 1997b [their boldface, our underscores]):

> [W]ithin a period to be determined by each Member State which may not exceed five years and which may be longer in Member States with particularly high unemployment, Member States will ensure that:
> (1) every unemployed **young person** is offered a new start before reaching six months of unemployment, in the form of training, retraining, work practice, a job or other employability measure;
> (2) **unemployed adults** are also offered a fresh start before reaching twelve months of unemployment by one of the aforementioned means or, more generally, by accompanying individual vocational guidance.
> [...] Benefit and training systems [...] must be reviewed and adapted to ensure that they actively support employability and provide real incentives for the unemployed to seek and take up work or training opportunities. Each Member State:
> (3) will endeavour to increase significantly the number of persons benefiting from active measures [...]. In order to increase the numbers of unemployed who are offered training or any similar measure, it will in particular fix a target, in the light of its starting situation, of gradually achieving the average of the three most successful Member States, and at least 20%.

Clearly, this does leave room for interpretation, but still these guidelines are surprisingly concrete. Even more concrete answers to the aims behind the guidelines can be found in the FAQ website to the 1997 Luxembourg Jobs summit (European Commission 1997a). There it is indicated that the goal of the European Employment Strategy is to increase employment by some 12 million jobs in 5 years, i.e. raising the employment rate from 60% to 65%. This would imply bringing unemployment down to about 7%.

Even though this 5 year time period is not over yet, figures 3.1 and 3.3a along with the aforementioned OECD predictions indicate that it will be difficult to achieve this objective by the end of 2002. Thus, also the latest Council Decision on the 2001 employment guidelines has adjusted the goals accordingly. However, rather than decreasing the 65% target in the light of the actual recent development, the time frame has been extended and the current objectives target an overall employment rate of 70% to be reached by 2010 (cf. Council of the European Union 2001). We do not intend to give a deeper analysis of the current employment guidelines, as these have – unsurprisingly, we are still within the initial 5 year horizon – remained more or less the same. While there has not been any further specification of concrete targets (small exception: all EU schools should have internet access by the end of 2001), it seems very much as if the initial guideline formulations had been softened even further in order to accommodate individual member states' desire (or necessity) to possibly deviate from the objectives.

But even if many objectives remain vague, the basic feedback set-up of guidelines that lead to NAPs, NAPs and actual developments that lead to annual Joint Employment Reports, which in turn lead to revised guidelines for the next year etc. does seem promising. In particular the Joint Employment Reports give a clear – if not always concise – account of problems that have been addressed and problems that need to be addressed (cf. for instance European Commission 1998, 2000). The Reports' "Identification of Good Practice" might indeed help identify policy measures across countries that work. However, the main caveat and major problem in this feedback set-up remains: "Systematic evaluation of employment and labor market policies is still not common practice in many Member States" (European Commission 1998). And how else would you be able to judge the performance of any employment policy? Indeed it remains strikingly odd that only few member states turn to independent scientific evaluation in order to achieve the desired monitoring of progress and learn about the desired effectiveness of policy mix (cf. Council of the European Union 2001).

Ideally, this study would review evidence on ALMP programs implemented after the beginning of the Luxembourg process. This, however, is not yet possible, and therefore most or almost all available results refer to programs either finished or at least initiated before the introduction of EU Employment guidelines. Still, we are convinced that these results can show how existing policies should be continued, adjusted, or abolished, and hence show which of them should play a further role in the Luxembourg process.

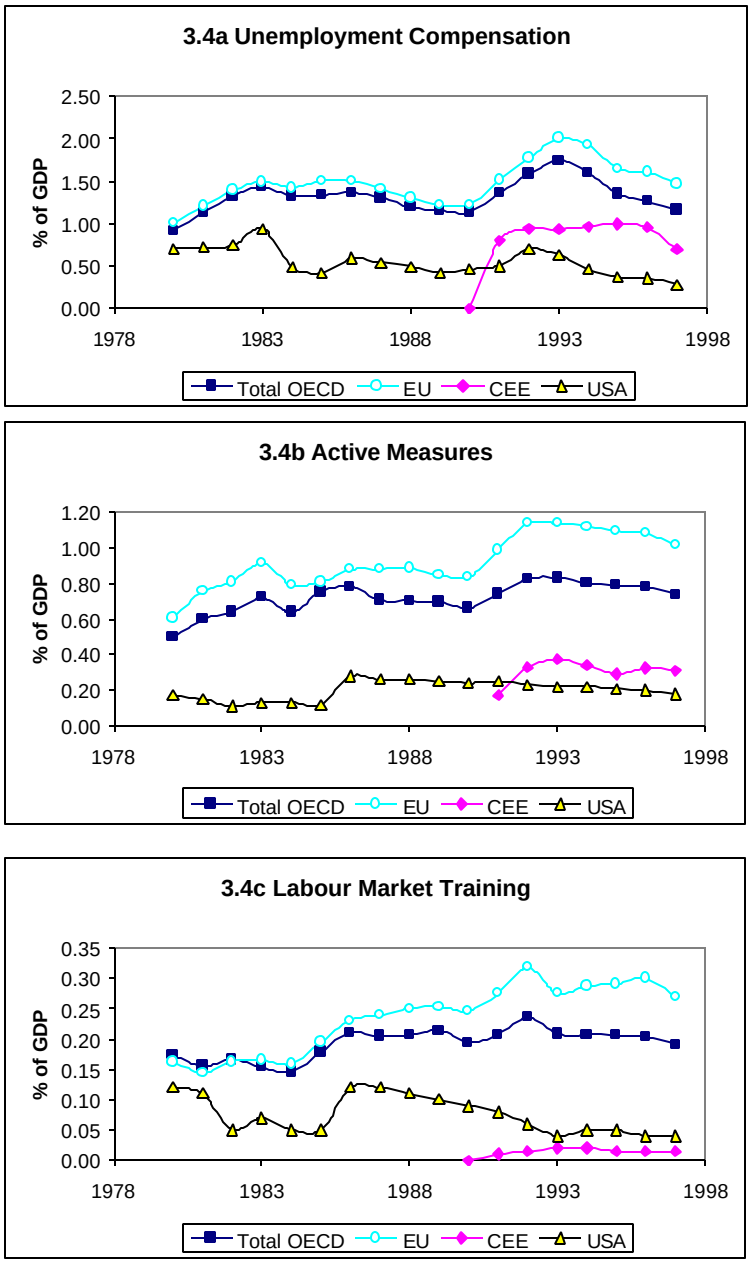
3.2.3 Active Labor Market Policies in Europe

Panels a to c of Figure 3.4 depict time series of the share of their GDP that the US, the EU, the CEE countries, and the total OECD, respectively, allocated to (a) unemployment compensation, (b) active measures in general, (c) labor market training in particular. For the unemployment benefit allocation in the EU, Figure 3.4a shows a more or less stagnant series until 1990, then a steep incline until 1993, followed by a substantial decline until 1997 almost back down to mid-1980s level. This shape is in line with the development of the unemployment rate shown in Figure 3.1. While the curve for the OECD average is very close to the EU series at a lower level, the US series varies less, although it displays a similar behavior for the 1990s – this, however, at a substantially lower (relative) level.

In terms of active measures delineated in Figure 3.4b, the EU has seen a more or less steady increase since 1980, also including a more pronounced increase between 1990 and 1992, followed by a slow decrease since. The OECD average shows a similar though less

distinctive development. The US has spend a substantially lower GDP share on active measures than the OECD total, decreasing very little but steadily ever since a short increase between 1985 and 1986. The series for labor market training in Figure 3.4c display similar shapes, although the EU high of 1992 is followed by a short steep decline, a steady increase until 1996 and another dip in 1997. For the US, the 1986 high is also more pronounced, as is the steady and substantial decrease up to a stagnant series from 1993 on.

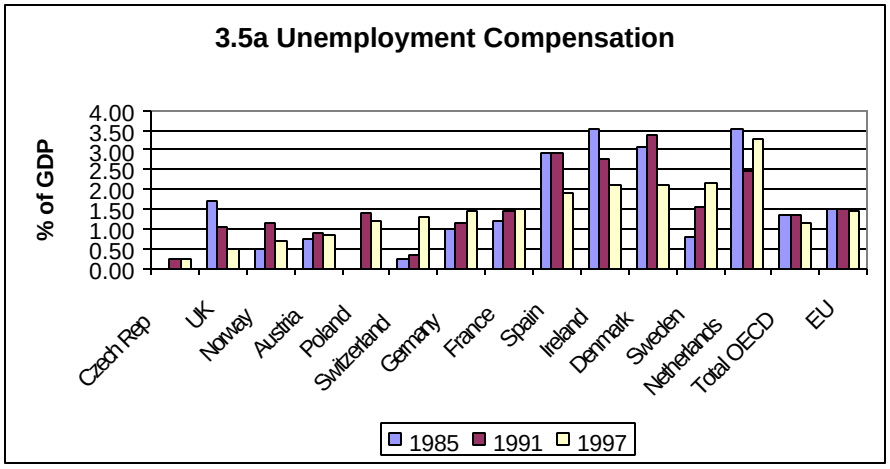
Figure 3.4 Public Spending 1980-1997



Source: OECD (2000a).

Figure 3.5 illustrates the development of public spending on (a) unemployment compensation, (b) active measures, and (c) labor market training for the years 1985, 1991, 1997 for selected European countries (cf. Martin 2000 for the composition of total spending on the various measures). Countries are ordered by 1997 expenditure from left to right, including bars for the EU average and OECD total to the far right. Unsurprisingly, Sweden and Denmark can be found to the right in all panels, with Sweden leading expenditure on active measures, and Denmark on labor market training. Both countries are similar in terms of their 1997 spending on unemployment compensation. However, whereas in Sweden this implies a strong increase, the panels underscore the fact that Denmark has been found the prime example among European countries regarding the transition from a benefit system of passive measures to one of active measures.

Figure 3.5 Public Spending by Countries



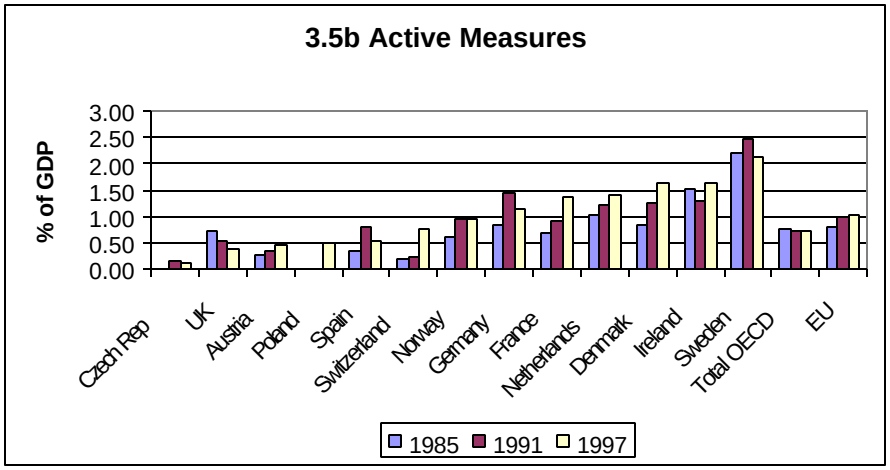
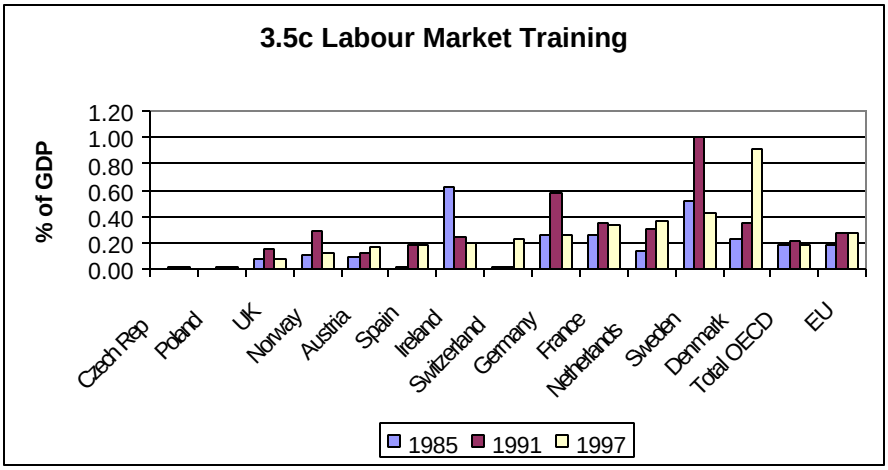


Figure 3.5 [ctd]



Source: OECD (2000a).

Other findings include the UK having reduced public spending on active measures (3.5b), and having strongly reduced spending on benefits (3.5a), which results in a position to the far left. Furthermore, a general look at Panel a reveals that the contemporary 1994 peak in benefit expenditure in the EU seen in Figure 3.4a is concealed by looking at pre-peak (1991) and post-peak (1997) numbers only. Countries like Germany and France display relatively little change, noteworthy being France increasing its spending on active measures (3.5b) and Germany temporarily devoting large budget shares to labor market training in 1991. The latter is due to the offering of labor market training to large numbers of newly unemployed from the Eastern Länder succeeding reunification (cf. the study by Lechner (2000) presented in section 3.5). Ireland managed to reduce unemployment compensation payments (3.5a), and largely reduced spending on labor market training (3.5c). The Netherlands allocate a large share of GDP to benefits (3.5a). Public spending of substantial amount on both passive and active measures is a very recent phenomenon in Switzerland – a finding in line with significant unemployment having only occurred in recent years. OECD and EU averages, respectively, display relatively little change.

3.3 Why can Measures of ALMP be Useful Policy Tools?

From the viewpoint of the individual worker – which is the perspective most microeconometric evaluations take – participation in an ALMP program might increase earnings and/or employment probabilities via increasing human capital. Whereas this argument is intuitively appealing on the individual level, in fighting structural unemployment training schemes, subsidized employment and similar policies can in principle only be useful policy instruments if the problem they address results from some kind of market failure. In theory, four basic functions have been attributed to measures of ALMP (cf. Calmfors 1995) – for further discussion we refer to the prototypical paper by Calmfors (1994).

The first and most general possible function is to raise a society's welfare and output by either letting the unemployed invest in human capital or putting them to work. This can be argued to go hand in hand with a secondary goal of increasing the welfare of the unemployed by providing meaningful activities for them. A second function of ALMP can be to maintain the size of the effective labor force, an argument that holds above all in the framework developed by Layard, Jackman, and Nickell (1991). The main idea here is that participation in labor market programs can maintain the search effectiveness and skill level of the unemployed, and thus keep up competition for the available jobs. Moreover, program participation might prevent discouraged workers who do not find jobs from leaving the labor force.

Thirdly, it has been proposed that measures of ALMP can help alleviate the moral hazard problem of unemployment insurance, i.e. to counteract misuse of unemployment benefits. As, in a deep recession, it may be impossible to test benefit claimants' willingness to work by offering regular jobs, program participation offers may be a suitable substitute. Clearly, such a policy would not increase employment, but rather reduce the number of benefit claimants by "harassing" the unemployed (Calmfors 1995).

A fourth possible function of ALMP might be to induce re-allocation of labor between different sub-markets. This argument has been one of three main policy directions identified by the Manpower and Social Affairs Committee of the OECD, which was created in 1961. In the original account, this is a Phillips curve-based argument, where ALMPs are meant to improve the trade-off between inflation and unemployment by stabilizing employment during the cyclical downswing and by removing labor-market bottlenecks during the upswing (OECD 1990). The other two original policy objectives cover the classical efficiency and

equity arguments: ALMPs are meant to (i) develop human resources and adjust manpower resources to structural changes with a view to fostering economic growth, and (ii) to improve both employability of and opportunities for disadvantaged groups, and thus contribute to social equity (OECD 1990).

Conceivable theoretical drawbacks to ALMP are also manifold. In this connection the main worries center around job creation schemes possibly giving rise to deadweight losses, i.e. the subsidized jobs would have been created anyway, or substitution effects, i.e. the subsidy leads firms to employ workers qualifying for the subsidy instead of unsubsidized workers, or displacement effects, i.e. subsidized firms can expand employment at the expense of employment in unsubsidized firms of the same sector. Another possible adverse side effect is the lock-in effect of training and job creation programs: Even though programs might have positive effects after completion, job search activity is likely to be low during participation.

The OECD (1993) and Calmfors (1994) employ "macroeconomic" models trying to incorporate the various beneficial and adverse effects of ALMP. However, whereas this type of analysis does serve to illustrate that labor market programs work through several channels, it becomes very difficult to empirically disentangle the effect of policy on the labor market (OECD 1993). It is therefore no surprise that the evaluation literature has almost exclusively focused on program effects on the average individual participant, leaving general equilibrium effects aside.

3.4 Received Wisdom: US versus Europe

What have we learned so far from empirical research on active employment polices? This section reviews some received evidence by contrasting evaluation research on both sides of the Atlantic. In a first step, this mainly contrasts the facts that (a) US studies to a great extent are based on experimental evaluations, while (b) European evaluation studies predominantly (or almost exclusively) make use of nonexperimental data. These observations also reflect different cultures of program evaluation: To have labor market programs evaluated by independent researchers or (non-)profit agencies unrelated to the government is an almost self-evident fact in the US. The nowadays widespread use of experimental techniques results from previous experience with nonexperimental evaluations in the 1970s and 1980s, where researchers found these results to differ too strongly to deliver conclusive results. In Europe,

on the other hand, policy makers in many countries are still reluctant to introduce social experiments. Moreover, in many respects there does not seem to be a strong co-operation and communication between policy makers and researchers.

3.4.1 The US Experience

A thorough scientific evaluation of the impact of Active Labor Market Policy on individual employment prospects has been a long-standing concern of US policy makers and the American public. The US have thus acquired a substantial amount of empirical evidence over the last decades, as programs have been accompanied by scientific evaluation and the design of subsequent programs has often been adjusted according to scientific advice based on previous experience. In particular, this procedure led to the predominant use of experimental evaluation methods, since these have been believed to generate the most robust and unbiased treatment effect estimates. Clearly, it is beyond the scope of this chapter to give a full account of the US experience with evaluating labor market policies. This section merely intends to sketch both the relevant methodological topics and the empirical lessons.

Methodological Considerations. With respect to the scientific approach to evaluation in the US we find two central aspects. First, the distinction between studies set in an experimental or a nonexperimental context, the predominant use of the former, and the discussion about which approach can produce reliable estimates under what circumstances. Second, the focus on earnings rather than employment as the outcome variable of interest.

It is the fundamental problem of any evaluation study that it is impossible to observe individuals in different states of nature at the same time and place[29]. Specifically, while the post-program employment performance of a participant in a training program can be observed, we will never be able to observe the employment performance the same individual would have experienced if he or she had not participated. Yet, a comparison between these two states in order to establish the causal effect of treatment is at the heart of any sensible study about the impact of a program. Thus, finding a credible estimate for the *counterfactual* state of nature is the principal task of researchers engaged in evaluation studies.

This in fact is the main underlying problem of any empirical study trying to assess labor market program treatment effects. The two principal evaluation designs are those of an *experimental* study vs. a *nonexperimental* study. In a social experiment, potential would-be-

participants are randomly selected into a *treatment* and a *control group*, thus ensuring that both groups do not differ from each other systematically in neither observable nor unobservable characteristics. This procedure rules out any potential problems of bias due to selection-into-treatment and makes post-treatment outcomes of both groups directly comparable. In a nonexperimental or *observational* study, however, an appropriate *comparison group* (and thus a credible counterfactual) has to be constructed retrospectively.

Social experiments offer thus a very convincing study design for generating a credible estimate of the counterfactual situation – since participants are chosen into the program by a random mechanism, it is straightforward to find comparable individuals that happen to have been randomly excluded in this assignment process. The construction of the desired counterfactual does not require extensive statistical or econometric techniques. Many evaluations of US labor market programs follow this reasoning and perform a social experiment, like the *National Supported Work Demonstration NSW* in the 1970s, or the *Job Training Partnership Act JTPA* in the 1980s. The most stable and widely accepted empirical results on North American training and incentive programs stem from experimental evaluations.

Often an experimental evaluation study is not a technically feasible option or meets the resistance of program administrators. Many researchers who had to rely on nonexperimental data in their analysis have emphasized the benefits of using longitudinal data (cf., e.g., Ashenfelter and Card 1985). This argument has shaped many nonexperimental analyses, most significantly in the discussion of the major US training program of the 1970s, training under the *Comprehensive Employment and Training Act* (CETA). A prominent nonexperimental alternative to such a longitudinal analysis takes a cross-sectional approach, using so-called *control functions* to model the implied sample selection bias. This is a strategy dominating the European literature on program evaluation.

We have seen above that, whereas an experimental study design partitions individuals into treatment and control group and thus supplies the desired counterfactual directly, the comparison group does not come naturally to an observational study and has to be constructed. In many cases the researcher can at least rely on the same dataset to establish a comparison group, e.g. via matching methods. Occasionally, however, comparison groups have to be constructed from data sets other than those from which the treatment group originated, where that data set may contain, e.g., different variables, or the same variables

[29] A comprehensive review of this problem is given by Heckman, LaLonde and Smith (1999).

measured differently. The construction of a comparison group is thus not a trivial exercise. The researcher – in her effort to establish the counterfactual – has to certify that his nonexperimental study meets the requirement formulated by Heckman, Ishimura and Todd (1997) as "to compare the comparable". In this sense, especially matching methods – by conditioning on pre-treatment covariates – seem to go a long way towards establishing an appropriate comparison group. This recent development has led to a renaissance of nonexperimental evaluation methods, the applicability of which had been strongly questioned since the influential paper by LaLonde (1986).

It is with regard to the nature of the *outcome variable of interest*, where we find one fundamental difference between North American and European evaluation studies. The US literature primarily focuses on the impact of employment programs on post-treatment incomes, rather than dealing with the participation effect on subsequent employment histories. The latter is far more common in European evaluations. This distinct emphasis is in line with the fact that US labor market programs aim at reducing wage inequalities in raising individual human capital of the economically disadvantaged (and therefore raising their wages), while European labor policy measures focus on reducing unemployment in raising individual employment probabilities. In the latter case, employment performance is usually being measured by either employment (unemployment) rates or employment (unemployment) durations, i.e. hazard rates. From a methodological viewpoint this obviously implies the use of evaluation methods designed for dealing with discrete variables such as labor market states rather than continuous variables such as incomes[30].

Empirical Evidence. This subsection presents a short summary of the quantitative evidence available from the evaluation of US labor market programs along the lines of Stanley, Katz and Krueger (1999). First, we look at basic findings for the various populations examined, before turning to conclusions on the effects of various types of programs.

A first finding on the program impacts by target population is that disadvantaged youth are in general difficult to assist, although some programs have succeeded in doing so. Programs for disadvantaged youth under 21 – in particular high school dropouts – have been less successful than programs for other populations, unless training was highly intensive or quite well implemented. Programs offering government jobs appear successful in improving

[30] Cf. Ham and LaLonde (1996), Magnac (2000), and van den Berg (2000) for a comprehensive account of duration methods for estimating treatment effects.

employment rates during the subsidy period, but there is little or no evidence on long-term positive effects. The target population of poor adults – especially single parents seeking to leave welfare – appears to respond well to training programs. Above all those programs with a subsidized employment focus seem promising. Furthermore, programs aimed at encouraging additional job search assistance for dislocated workers appear to have positive effects.

Looking at program types job search assistance in general seems to be a measure that reliably speeds the return to work and saves government money. While earnings impacts tend to be moderate at best, Stanley et al. (1999) find the record for these efforts to be "consistent and clear", as on average they do lead to a faster return to work. Among other program types, earnings supplements, hiring subsidies, and subsidized employment lead to employment gains for the disadvantaged. However, the increase in employment rates mainly occurs during the period when the actual subsidy is offered. On the other hand, if subsidized employment is combined with on-the-job training, adult trainees – but unfortunately not the disadvantaged youth – in such programs show higher employment rates and earnings well after the period of subsidized employment is over.

3.4.2 The European Experience

Quite a few of the distinct features of European evaluation research have already been pointed out in the previous section in order to appropriately contrast them with the US experience. We thus keep the discussion rather short in this section. In a way this structure of discussing the US first and Europe subsequently also reflects the fact that European evaluation research appears to be (at least) one step behind US evaluation research.

The fact that European evaluation research is lagging behind the US can certainly neither be attributed to the fact that unemployment is a comparatively "recent" phenomenon in Europe (cf. section 3.1), nor even to a lack of understanding or a lack of willingness of European researchers to follow American researchers in their development of modern evaluation practice. Quite the contrary, one has to confess that in recent years evaluation research – above all in economics – has been one of the "hot" academic topics in Europe, resulting in a large number of studies across countries. Many European economists have specialized in the field of evaluating labor market policies. European researchers clearly do understand the methodological problems of an experimental vs. a nonexperimental design and the inherent difficulty to establish a credible counterfactual etc., and European researchers can handle the analytical tools that try to solve these problems.

In our assessment, what makes Europe "lag behind" is the prevalence of nonexperimental data, or – from the opposite perspective – the too few cases where an evaluation is set in an experimental context. This in turn may partly be due to a lack of communication between policy makers and the scientific community – policy makers in many European countries still seem to be reluctant to introduce social experiments, even though most academic experts are calling for their introduction (cf., e.g., Schmidt 1999). Furthermore, even nonexperimental studies rarely go hand in hand with the program itself – rather, programs are often implemented without any thought of whether they will ever be evaluated. High quality data remains the exception. On the other hand academics often conduct their studies without real connection to the program and possibly without ever being able to communicate their results to policy makers[31]. Therefore, unsurprisingly, this European "evaluation culture" has a lot to learn from US practice, and unfortunately we still have to assert European evaluation research to be in its infancy[32] and results to be far from clear-cut, even though we do observe an increasing number of evaluation studies, and even though these studies increasingly display methodological rigor. The resulting evidence will be presented in the next section.

3.5 Recent European Evaluation Studies

The following selective review of recent European evaluation studies is organized along coarse regional groupings. Specifically, we will discuss selected studies from (i) the Nordic countries (ii) the UK and Benelux, (iii) Central Europe, and (iv) transition countries of CEE. Rather than claiming to be comprehensive our review concentrates on very recent and partially ongoing work that has been conducted in the field. For complementing information on previous work see, for instance, Heckman, LaLonde and Smith (1999).

One aspect of particular interest in this section is the broad range: Even though we present only a limited number of studies, and even though evaluation research in Europe still has to catch up with the US, we find a substantial variety of studies. With respect to elements of study design, the presented analyses vary e.g. in their evaluation design

[31] There is, however, the occasional counterexample where policy makers and academics do work together, cf. the study by Gerfin and Lechner (2000) presented in section 3.5.3.

(experimental/nonexperimental), in the target group being analyzed (youths/low-skilled unemployed etc.), in the type of program being focused on (training, wage subsidies etc.), in the outcome variable measuring program success (hazard rate, employment rate, earnings), and in the econometric estimation technique being applied (duration models, matching estimators, selection models). We thus believe that this selective review does go a long way in giving an insight into current state-of-the-art European evaluation research and its policy implications.

3.5.1 The Nordic Countries

Sweden. The Nordic countries - in particular Sweden - were among the first to introduce and evaluate measures of Active Labor Market Policy. The paper by Larsson (2000) evaluates and compares the treatment effects of two Swedish labor market programs directed at the young unemployed: *Youth Practice* and *Labour Market Training*. Youth Practice was a large-scale youth program targeting unemployed aged 18-24. It was a subsidized program placing participants in both the public and the private sector. In order to minimize potential displacement effects, participants were supposed to perform tasks that otherwise would not have been done. The program also included job seeking activities. Youth Practice was subject to some entry requirements trying to ensure that it was only used as a "last resort" after all other alternatives had been tried.

The second labor market program evaluated in the study is the traditional Labour Market Training, a program that already had existed for a "very long period". Its aim was to improve the skills of unemployed job seekers in such a way that they are better matched to labor demand. Regulations regarding allowance as well as job search during participation were the same for both programs, while the main purpose differed to some extent: Labour Market Training was targeted mainly at low-skilled individuals in the field in which they were searching for jobs, whereas Youth Practice aimed at increasing young people's working experience.

As Swedish Active Labor Market Policy is explicitly meant to enhance the employability of the unemployed, Larsson (2000) chooses both employment probability and earnings as outcomes to measure the programs' success. She complements these two by also estimating "probability of transition from unemployment to education", in order to see

[32] This has been a common assertion in European evaluation research, cf. – among others – Zweimüller and Winter-Ebmer (1996) or Magnac (2000).

whether for some people regular education might be a way of avoiding a vicious cycle of temporary unskilled jobs, unemployment, and programs. To estimate treatment effects the study employs a multivalued treatment setting, assuming an unemployed individual has three treatment options: Youth Practice, Labour Market Training, or job search remaining openly unemployed. The estimation strategy then follows a propensity score matching set-up based on the conditional independence assumption.

The paper finds the short-term effects on both earnings and employment to have been significantly negative throughout for both programs. However, two years after program start they began to become more positive and the only significantly negative effects were those of Labour Market Training on earnings. Youth Practice does not seem to have had any effect on the probability of studies, while participation in Labour Market Training significantly decreased the study probability of participants. A direct comparison of the two programs finds Youth Practice to have been better (or, rather, less harmful) for participants than Labour Market Training in terms of all outcome measures. Larsson (2000) tests the robustness of these results by comparing them to estimates obtained from standard OLS regression and probit analyses, where she notes them to be almost identical, except for Youth Practice employment effects being zero in the short run and slightly positive in the long run.

What may be reasons behind the negative treatment effect estimates? With respect to Youth Practice, participants apparently put less or no effort into finding a job during the program, even though program regulations demanded continued active job search. Furthermore, the negative outcomes might be due to insufficient planning and follow-up, as well as low-qualified tasks that did not provide any human capital accumulation. Labour Market Training, on the other hand, has been there for decades, so that start-up problems cannot provide an explanation for its bad performance. A potential explanation might be that courses do not fit the employers' requirements for labor, and that training thus displays both professional and regional "lock-in" effects. Such lock-in effects also might be present in the sense presented in section 3.3, namely that program participation substantially decreases job search relative to the comparison group: In fact it turns out that in this study the "non-participation" state is rather characterized by an alternative "program", i.e. ongoing job search assistance.[33]

The recent study by Sianesi (2001) comes to a similar conclusion with regard to non-participation stating that "in Sweden nobody is really left 'untreated'". Therefore it is

[33] This interpretation was given in personal communication with the author.

important to note that the effectiveness of an ALMP program will be judged against the alternative of continued receipt of employment offices' services. In her analysis of Swedish ALMP Sianesi (2001) assesses the whole set of programs – such as labor market retraining, public sector employment, subsidized jobs etc. – condensed into one artificial "treatment" state. This treatment is evaluated using propensity score matching. The findings indicate that this treatment – trying to capture Swedish ALMP overall – at best displays zero effects on employment probabilities. This is the result when cycling behavior is ruled out.

In general, however, findings point to negative treatment effects due to work disincentives provided by the system: Participation in a labor market program for five months counts as employment and renews eligibility for another spell of unemployment compensation. This interaction between the active and the passive part of Swedish labor market policy appears to be the most important and most distorting factor in evaluating the effectiveness of Swedish ALMP. This conclusion receives strong support from the study by Regnér (2001) who analyses employment training programs. Applying various specifications of a selection model – linear control function, fixed effect, and random-growth – Regnér (2001) finds that training had no or significantly negative effects on earnings. He also attributes this to cycling behavior of program participants.

Norway. Using the same three types of selection models as Regnér (2001) – complemented by a modified random growth model with unrestricted growth component – the study by Raaum and Torp (2001) evaluates the effect of Norwegian labor market training on earnings. The specification tests conducted by Raaum and Torp (2001) reject all models but the linear control model, which reports a positive training impact on earnings. However, the authors attribute this to favorable unobservable characteristics among the participants. An interesting feature of this study is the availability of an internal comparison group, i.e. a group of unemployed who applied for participation in the program but were rejected. This group displays higher "post-training" earnings than non-applicants, and thus appears to provide a more suitable nonexperimental comparison group.

Denmark. Two closely related Danish studies, Jensen (1999) and Jensen, Svarer Nielsen and Rosholm (2000), focus on the rather newly introduced *Youth Unemployment Program* (YUP) as of 1996. In strong contrast to other OECD countries, Denmark has experienced a dramatic decline in its youth unemployment rate in recent years. The papers investigate whether this

unique effect has been due to the YUP. The YUP appears to be a thoughtfully targeted and carefully implemented labor market program, which works through a "combination of benefits, incentives and sanctions" (Jensen (1999)). The unemployed, low-educated youth constitute the target group, with the purpose of motivating them to undertake an education or to find a job.

Jensen et al. (2000) investigate the effect of the program on the duration of unemployment spells and on the transition rates from unemployment to schooling and employment. Based on nonexperimental data, they estimate duration models for grouped duration data, allowing for the presence of unobservables which may be correlated across destination states in a competing risks model. The main finding of this analysis is a significant increase in the transition rate from unemployment to schooling due to the YUP. Jensen et al. (2000) attribute this effect mainly to a direct program effect (the effect experienced by individuals who leave unemployment to participate in an educational program), and to a smaller extent to a sanction effect (the effect from the removal of unemployment benefits after 6 months if an individual does not accept the offer of a YUP slot).

They find these effects after correcting for seasonality in the transition rate from unemployment to schooling. Furthermore, Jensen et al. (2000) report somewhat weaker effects on the transition rate from unemployment to employment. They do not find an announcement effect (individuals at risk of being affected by the YUP behave differently to those not at risk). Thus, the two studies conclude that the YUP has been successful in lowering youth unemployment in Denmark, at least in the short run.

In another Danish study Rosholm (1999) evaluates the Danish employment subsidy program. In particular he studies the individual effects of having completed a temporary employment subsidy on the hazard rates out of employment and unemployment. He accounts for between-program selection bias by explicitly modeling the selection process. Rosholm (1999) estimates two variants of a competing risks duration model on a non-random sample of Danish workers followed during the period 1983-1990. This sample consists of treatment participants and a nonexperimental comparison group consisting of participants in other types of programs and of future participants, i.e. individuals who participated after 1990. Rosholm (1999) reports the following findings on the treatment effect of the Danish ATB ("ArbejdsTilBud" = job offer) program:

The effect on the unemployment to employment hazard of a private sector ATB is generally positive, while the effect of a public sector ATB is mostly negative. The only

exception to the latter is the largest subgroup – medium-aged females – for which it is significantly positive. Rosholm (1999) states the difference between private and public sector ATB to be "real", since it persists even after including the selection process into the model. A large average fraction of participants (approximately 50%) remain in their subsidized workplace, with highest fractions for private sector ATBs, and for women. The effect on the hazard rate from employment is strongly negative. Rosholm (1999) concludes that employment subsidies appear to improve the employment chances of long-term unemployed individuals. He tends to attribute the bad performance of public sector employment subsidies to (perhaps unjustified) stigmatization.

3.5.2 The UK and Benelux

The United Kingdom. A recent paper by Bell, Blundell and van Reenen (1999) examines alternative approaches to the evaluation of the impact of a temporary wage subsidy and a training program. It does this in the context of a recent active labor market reform for the young unemployed in Britain: Labeled the "New Deal for the Unemployed Youth", this active labor market program was introduced in 1997 by the newly elected Labour Party government. It was part of a general package of welfare-to-work reforms directed toward the low wage labor market, with the "New Deal" specifically aimed at enhancing the employability of the young long-term unemployed.

One of these unemployed individuals first enters a "Gateway" period, a period lasting up to four months during which the individual receives extensive help in job search. According to Bell et al. (1999) a substantial proportion of the unemployed are moved off the register during this Gateway period. Those remaining are being offered four options, comprising either work in a subsidized job in the private sector, or work in a subsidized job in the public sector (two distinct but similar options), or participation in a training program. All of the wage subsidy programs also contain a one-day-per-week training element. These options, however, are semi-mandatory, as failure to comply without good cause may result in benefit sanctions being applied.

The long-term unemployed in the UK, including those below 25 years of age, are disproportionately male, even though "it is clear that long-term unemployment is a far greater problem amongst the old than among the young". Moreover, Bell et al. (1999) find the key characteristic of those eligible for the New Deal to be their low level of skill and consequent low productivity. The authors argue that a key rationale of the scheme is to enhance the

participants' employability by making them more productive. Potentially, productivity might increase through an experience or tenure effect and training opportunities associated with having a job – a dynamic effect that could have a permanent effect on unemployment.

In their evaluation approach Bell et al. (1999) emphasize the need to use intertemporal or dynamic methods in order to understand whether the program will have any long-run effects on the employment probabilities of the target group, i.e. will participants be able to hold on to a job once the subsidy runs out. The paper suggests using a "trend adjusted difference-in-differences" approach as an empirical strategy for the ex-post evaluation, an approach that potentially is able to deal with many econometric problems associated with evaluation. However, as the authors point out, there are some difficulties arising from the fact that there is no obvious control group which the treatment group could be compared to. A complementing method proposed by Bell et al. (1999) is the construction of an ex-ante general equilibrium model of the labor market, using existing information to calibrate the parameters of the model.

In the empirical analysis they estimate the effect of job duration on productivity (in terms of wages) for the target group using micro data from the *British Labour Force Survey*. The paper concludes that the productivity effects appear to be relatively modest compared to the size of subsidy deemed necessary to get the group into jobs. Thus, it seems likely that the policy effects of the New Deal will be far more modest than its proponents have hoped for.

In this respect it is interesting to note that the 1998 Joint Employment Report (European Commission 1998) already identified the New Deal as an “Example of good practice suggested by Member States” in preventing youth and long-term unemployment (Guidelines 1 and 2). This finding basically seems to be derived from the fact that the New Deal “underlines the trend towards the activation approach” (European Commission 1998, p.114). Moreover, in trying to assess the effective scope of the New Deal and other measures in the UK the Report simply concludes that “it is impossible to determine the exact proportion of long-term unemployed who benefit from such measures” (p.115).

The Netherlands. Using data from a social experiment, the paper by van den Berg and van der Klaauw (2000) evaluates the effect of *Counseling and Monitoring* (C&M) on the transition rate from unemployment to work. In the Netherlands, the Active Labor Market Policy C&M is an activity provided by the local unemployment insurance (UI) agencies. It is provided to UI recipients with relatively good labor market prospects, and it consists of

monthly meetings with a local UI agency employee for a period of 6 months. "Good labor market prospects" refers to a Type I unemployed in a four-type categorization set up by UI agencies in order to better tune its services to the needs of the unemployed. During the monthly meetings past job search activities are evaluated and plans for future job search activities are made. Hence, the main goal of C&M is the reduction of the duration of unemployment and consequently the total amount of UI benefits paid. Van den Berg and van der Klaauw (2000) emphasize the need for an evaluation study of C&M to focus on the duration until exit to work and to take place on the individual level.

The data are administrative data coming from a social experiment, so that sample selection bias from nonrandom participation or reliance on instrumental variables or functional form assumptions are not an issue. The experiment was conducted at Rotterdam and Eindhoven, the 2nd and 5th largest cities of the Netherlands, respectively. Van den Berg and van der Klaauw (2000) report the local UI offices in these cities to be relatively large (large inflow into UI) and to provide C&M of high quality. The participants in the experiment were randomly selected into the treatment group receiving C&M and the control group not receiving C&M. None of them knew in advance that the experiment was going on. Usually, all of them would have received C&M. However, none of the individuals in the control group complained about not receiving C&M. This set-up ensures that the data do not suffer from initial nonrandom non-participation in the experiment, and participants cannot leave the experiment for any reason other than stopping collecting UI benefits.

Van den Berg and van der Klaauw (2000) offer a theoretical and empirical investigation of C&M of unemployed workers. In the theoretical part, they investigate the exit rate to work using a job search model with multiple search channels and endogenous search effort. The search channels include formal and informal job search. In the empirical analysis, the authors estimate the effect of C&M on exit to work with non-parametric and parametric methods, with duration models and with limited-dependent variable models. The duration models concern common reduced-form hazard rate models, including e.g. a mixed proportional hazard specification where the transition rate from unemployment to employment is allowed to depend on observed individual characteristics, on the elapsed unemployment duration, on unobserved determinants and on a variable indicating whether the unemployed receives C&M or not. The estimation of the model is based on a flow sample of UI recipients.

The empirical results show that providing C&M does not have a significant effect on

the individual transition rate from unemployment to employment. From their theoretical model and comparisons to other studies the authors find two reasons why the analyzed C&M does not seem to be a successful labor market policy. First, the population of unemployed individuals who receive C&M consists of UI recipients with good labor market prospects, and second, the program does not provide sufficient assistance. But although the estimated effect of C&M is small, van den Berg and van der Klaauw (2000) conduct a cost-benefit analysis showing that the Dutch C&M can be considered as cost effective. They attribute this finding to the low costs of providing C&M. Looking at the results, the main question concerns the choice of target group of the program: Obviously the Type I unemployed workers with relatively good labor market prospects do not seem to be the ones who profit most from receiving C&M, as they would be the ones who would be thought best in doing it "on their own". On the other hand, however, they are the ones who already have sufficient human capital and mainly need job search assistance, while for Type II to IV unemployed C&M may not be too helpful, as they need to acquire human capital first.

3.5.3 Central Europe

France. Our review of ALMP in France focuses on the well-received study by Bonnal, Fougère, and Sérandon (1997). A complementing overview of recent microeconometric results on the evaluation of the effects of ALMP on youth employment in France can be found in Fougère, Kramarz, and Magnac (2000). These authors report training programs for unemployed young workers in general to have no effect on post-treatment wages or employment probabilities, except if they have a large training content.

In their paper Bonnal et al. (1997) deal with the evaluation of public employment policies set up in France during the 1980s to improve the labor market prospects of the most disadvantaged and unskilled young workers. The evaluation is restricted to the short-term impact of youth employment schemes on subsequent unemployment and employment durations of recipients. Bonnal et al. (1997) estimate a reduced-form multi-state multi-spell transition model that includes participation in the program as an additional state. Given their framework, participation in a program is allowed to affect the transition rates out of the state following the program, and distinct types of programs (i.e. public sector programs vs. private sector programs) are allowed to have different effects. Their model also allows for possibly related unobserved heterogeneity in the specifications of all transition states, capturing the potentially selective nature of program enrolment.

In the empirical analysis they use nonexperimental micro data from administrative records collected in the period from 1986 to 1988. The data provide information on the dates of program entry and on durations of subsequent spells of employment and unemployment. Bonnal et al. (1997) distinguish between two types of programs. First, the alternating work/ training program provided by private firms, including apprenticeship, qualification and adaptation contracts, and "courses for the preparation to working life". Second, the "workfare" program provided by the state and the public sector, including community jobs and "courses for the 16-to-25-year-old". In the second type the amount of vocational and specific training is generally lower than in the first type. In this respect the paper addresses the main question: can we also differentiate these two types of program when we consider their impacts on durations and outcomes of subsequent unemployment and employment spells?

Bonnal et al. (1997) use individual labor market transition data distinguishing between six labor market states: unemployment, permanent job, temporary job, public employment policy job, out-of-the-labor-force, and attrition. Their statistical model is of the mixed proportional hazard type, with piecewise constant duration dependence. Their estimates show the following: (i) French youth employment programs have differing effects on the recipients' trajectories. For instance, participation in the private sector program of the first type increases the transition rate from the following unemployment spell to regular employment for young low-skilled males, while it has no effect on the same transition for young men with better education. At the same time, the experience of a public sector "workfare" program has no effect on the intensity of transition from unemployment to regular jobs for the least educated young people, while this transition rate decreases significantly for young men with a vocational diploma. This subgroup may even end up stigmatized with having a low employment performance.

(ii) Participation in programs is highly selective. It depends firstly on the state currently occupied (better educated youths have higher transitions into a program from unemployment than from temporary employment). Secondly, it depends on the educational level (the least educated move less intensely from unemployment to programs). Finally, it depends on past occurrences of program participation, but also on unobserved individual heterogeneity. (iii) The duration of the period of entitlement to unemployment insurance does not increase the expected duration of unemployment spells. While they are still qualified for UI, the least educated young workers enter programs more intensely.

In a more recent paper Brodaty, Crépon, and Fougère (2001) re-examine the

evaluation of French youth employment programs. They use the same data as the above study by Bonnal et al. (1997) and re-estimate the impact of these programs on the subsequent employment status by implementing matching estimators. In their specification they focus on the propensity score as matching criterion. As the sample is extracted from the stock of unemployed people at a given date (August 1986), a natural specification of the participation probability may supposedly be derived from a competing risks duration model.

The results obtained by Brodaty et al. (2001) highlight the variability of program effects, both between programs and among recipients of the same program. They emphasize that this may be a particular problem as regards policy implications: due to the fact that their results are pairwise comparisons, different improvements may be sometimes proposed to the same person, or vice versa. In general on-the-job training programs in the private sector (associated with higher amounts of vocational and specific training) give better results than public sector programs. This result is in line with the Bonnal et al. (1997) study using a very different approach. In order to assess differing program effects for participants with varying conditional participation probabilities, Brodaty et al. (2001) study the relative effects of different programs along subintervals of the common support, i.e. for specific values of the propensity score. In general, positive effects on the whole common support are associated with significant positive effects on the highest part of the support and no significant effects on the lower part, while negative effects on the overall common support are associated with significant effects on the lower part and no effect on the highest part.

Germany. While high and persistent unemployment has plagued the German economy for a considerable time, it is the painful aftermath of the integration of East Germany which generated most concern among economists. Correspondingly, there are several studies of the impact of ALMP measures in Eastern Germany, one of which is reviewed here.

The paper by Lechner (2000) analyses the effect of public sector sponsored continuous vocational training and retraining in East Germany directly after unification. This training program was part of an extensive (above all in monetary terms) introduction of ALMP measures in Eastern Germany, aimed at avoiding high unemployment in a yet destabilized and slowly adjusting market. Using nonexperimental data from the *German Socio-Economic Panel* (GSOEP) for the period 1990-1994 the author presents estimates of the average individual gains from training participation in terms of earnings, employment probabilities and career prospects after the completion of the training program.

The group being analyzed consists of workers of the former GDR having participated in such a program between July 1990 and December 1992. Found to be a "highly informative data set" in this study, the GSOEP comprises a random sample from the East German population, thus containing both trainees and non-participants. Apart from many socio-economic variables being included, the data also make it possible to track individual employment histories on a monthly basis back to one year preceding unification.

Based on a nonparametric matching approach the findings suggest that in the short run public sector sponsored training has a negative impact. Lechner (2000) attributes this to the fact that participation in training reduces the job search efforts during treatment compared to a comparable spell of unemployment. Several months succeeding training, though, these effects diminish and no statistically significant differences between trainees and controls can be observed. Thus, the general finding is rather to attribute no positive earnings and employment effects to public sector sponsored continuous vocational training. While the risk of being unemployed seems to increase directly after treatment ends, this negative effect disappears during the first year after training. Further analysis of long-term effects was not possible using that very data set. It remains an open question whether the lack of a positive effect is due to either participants being stigmatized for future employers or insufficient quality of the program.

In finding the program to apparently be "very much a waste of resources", Lechner (2000) draws a rather pessimistic conclusion. He somewhat qualifies his view by saying that (i) at this early stage the East German training structure had just been built from scratch, that (ii) a significant reduction in the official calculated unemployment rate was indeed achieved, being one of the political aims of the program, even though individual labor market prospects were apparently not enhanced, and that (iii) positive training effects may materialize only after a longer time horizon.

Switzerland. A recent paper by Lalive, Zweimüller and van Ours (2000) presents some initial evidence on the impact of ALMP and benefit entitlement rules on unemployment duration in Switzerland. Switzerland exercises a reward-or-punish system of benefit regulation, where after 7 months of unemployment duration unemployment benefits are conditional upon program attendance. In this respect the Swiss case is interesting as Switzerland has gone particularly far in activity testing by adopting new rules linking benefit eligibility closely to participation in ALMP measures. Swiss ALMP measures entail both training courses and

employment programs.

Using nonexperimental data from administrative records Lalive et al. (2000) employ a "timing-of-events" duration method to study the impact of ALMPs on unemployment duration. In addition to the conventional procedure of modeling the mechanism determining selection into treatment together with the process of exit from unemployment, their approach makes explicit use of the information contained in the timing of the treatment. Treatment can be started at various points in time during an unemployment spell, and variation in that timing can be exploited to identify the treatment effect. Within their duration analysis framework Lalive et al. (2000) allow for unobserved heterogeneity. This relaxation of the conditional independence assumption, however, needs a much more restrictive specification of both the selection process into the programs as well as of the dependence of labor market outcomes on individual characteristics and time.

Lalive et al. (2000) find the following results: (i) after participation in ALMP the transition rate to jobs increases for Swiss women, but not for Swiss men. The job hazard rate, though, is strongly reduced during participation. Taken together, the authors conclude that programs prolong unemployment duration for men, but tend to shorten durations for women. (ii) Once the unemployment spell comes close to the running out of unconditional benefit entitlement the job hazard rate increases strongly, both for women and for men. (iii) The authors do not find any important selectivity effects regardless of gender.

Another econometric evaluation of Swiss ALMP is a study by Gerfin and Lechner (2000) that is part of the same research program initiated by the Swiss government. Thus, their data come from the same source as those used in Lalive et al. (2000). However, while the latter restrict their sample to inflows into unemployment between December 1997 and March 1998 – due to their econometric analysis being based on duration models –, Gerfin and Lechner (2000) ground their evaluation approach on the stock of those having been unemployed for less than a year in December 1997.

The data originate from administrative unemployment and social security records, and are claimed to be "unusually informative". This claim seems to be justified, as for a merged random subsample of about 25,000 observations the data contain information on individual labor market histories and earnings for at most 10 years prior to the current unemployment spell. Moreover, the data include a variety of sociodemographic characteristics, regional information, subjective valuations of placement officers, sanctions imposed by the placement office, and information on previous and desired jobs. The authors are thus confident that after

controlling for this wealth of information there will be little unobserved heterogeneity left that is systematically correlated with program participation and labor market outcomes.

The active labor market programs can be grouped into three broad categories, (a) training courses, (b) employment programs, and (c) temporary employment with wage subsidy, which the authors abbreviate to "temporary wage subsidy". A basic difference between (b) on the one hand and (c) on the other is that the former takes place outside the "regular" labor market, while the latter must be a regular job.

Training courses consist of a large variety of 16 types of courses ranging from basic courses to specific work-related training. In the analysis they are aggregated to five rather homogeneous groups. *Employment programs* can take place in both private and public institutions. A main feature is that they should be as similar as possible to regular employment, but still extraordinary, i.e. not in competition with other firms. Moreover, during an employment program the unemployed has to continue her job search and must accept any suitable job offer. While participation in both training and employment programs does not extend the benefit entitlement period, a *temporary wage subsidy* might do so if its cumulated duration exceeds 12 months. The aim of a temporary wage subsidy is to encourage job seekers to accept job offers that pay less than their unemployment benefit by compensating the difference with additional payments. As the income generated by the scheme is higher then the unemployment benefit for remaining unemployed, it is financially attractive for both the unemployed and the placement office.

The evaluation strategy of the paper follows a propensity score matching setup considering multivalued treatment. For an individual it is thus possible to have potentially been in one of nine mutually exclusive "treatments": one of five distinct types of training courses (basic, language, computer, further vocational, and other), one of two types of employment programs (private or public sector), temporary wage subsidy, or not participating in any program. The analysis considers employment as the outcome variable. The results show that for the respective participants in the programs temporary wage subsidy is superior to almost all other programs, with a mean gain between approximately 6% to 22% points. In fact, temporary wage subsidy is the only program that dominates non-participation. Summarizing all pair-wise and composite effects Gerfin and Lechner (2000) thus find that temporary wage subsidy is the most effective program, while employment programs as well as basic and language courses display negative treatment effects. For the other training courses the authors find mixed results.

Two aspects are particularly noteworthy about this study. First, it is based on an unusually informative data set, showing that good data can go a long way in helping to set up a solid evaluation study. Second, it finds the rather interesting results that a traditional employment program – probably due to taking place in a sheltered labor market – exhibits negative effects, while a rather unique program of temporary wage subsidy seems to be a powerful instrument. The only concern regards potential negative incentive effects of temporary wage subsidy in terms of underbidding of the wages set in collective bargaining, and avoidance of dismissal protection.

3.5.4 Transition Countries

From its very beginning the transition process of the formerly communist countries of Central and Eastern Europe has been accompanied by vivid economic research trying to grasp a deeper understanding of the developments in these countries. In this respect, also the implementation and evaluation of ALMP measures has attracted considerable interest. We focus on recent experiences from two countries, Poland and Slovakia. For a survey of earlier results and their implicit lessons for OECD countries see Boeri (1997).

Poland. In their microeconometric evaluation of the effectiveness of ALMP measures in Poland, Kluve, Lehmann and Schmidt (1999)[34] apply the method of matching as a nonexperimental substitute for randomization in labor market programs. Using retrospective data from the 18th wave of the Polish Labour Force Survey (PLFS) as of August 1996, the authors focus on a supplementary questionnaire containing information on individual labor market histories in monthly representation, and implement a conditional difference-in-differences estimator of treatment effects.

Along the lines of other transition countries during the 1990s Poland – facing high and persistent unemployment rates – also introduced a variety of ALMP programs. These include *Training, Intervention Works* (wage subsidy to private employers), and *Public Works* (a public sector employment program). The PLFS data follow individuals for a period of 56 months (January 1992 to August 1996) entailing information on their respective labor market state for every single month. Kluve et al. (1999) condense this information to a multinomial variable of labor market outcome (employed, unemployed, out-of-the-labor-force).

Furthermore, they construct treatment groups for each policy measure, respectively,

and individually and dynamically (exact point in time of program entry) match controls to program participants. Doing this the dynamic 'moving window' feature accounts for changing macroeconomic environment. The matching criteria applied in this procedure require the control to (i) display identical characteristics with respect to gender, education, region, and marital status, to (ii) hold an identical 12-month employment history preceding treatment, and to (iii) display minimum age deviation. The conditioning on pre-treatment labor market histories is a feature particularly noteworthy, as the employment record preceding entry into treatment has been found to be an important determinant of program participation (Heckman and Smith (1999); cf. also Card and Sullivan (1988) for an early application).

Having constructed treatment and matched comparison group following this procedure, Kluve et al. (1999) use the trinomial labor market outcome variable described above to analyze the effect of ALMP measures on employment and unemployment rates. Exploiting the history structure they take into account short-term (9 post-treatment months) and medium-term (18 post-treatment months) effects. The authors' findings suggest that training has a positive effect on the employment probability for both men and women. This effect is slightly more pronounced for women. Therefore, this ALMP measure clearly seems to improve the efficiency of the Polish labor market. Regarding intervention works there is no overall treatment effect for participating women, while the authors report strong negative treatment effects on the employment rates of men who took part in either intervention works or public works (Public works for women are not being analyzed due to small sample size). As participation in any of these two ALMP measures entitles the participant to a new round of unemployment benefits, intervention works and public works seem to be a common intermediate stage between two spells of unemployment benefit receipt, where the individual entered the program after having exhausted his benefit eligibility. Hence, while stigmatization might have some role to play, Kluve et al. (1999) attribute most of the negative overall treatment effects of these programs to 'benefit churning'.

The Slovak Republic. Van Ours (2000) studies treatment effects of training programs and subsidized jobs that were part of the Slovak program of Active Labor Market Policy. He also focuses on individual treatment effects. Of the ALMP measures introduced (and occasionally re-organized) in the 1990s in Slovakia, the author analyses three main program types: (i) *Training*, (ii) *Socially Purposeful Jobs* (SPJ), (iii) *Publicly Useful Jobs* (PUJ). The latter two

[34] This study is included as chapter 4 in this thesis.

are both programs of temporary subsidized employment, where SPJ were mainly created in the private sector and concerned higher qualified functions, while PUJ were low ranking jobs in the public sector best described as "community works". The SPJ program has been the most extensive Slovak ALMP program both in terms of expenditures and participants.

Using nonexperimental data from the unemployment register, van Ours (2000) focuses on a sample of three Slovak districts with detailed labor market information on male workers that started their unemployment spell in 1993. He investigates whether the exit rate to regular jobs increases if an unemployed person enters a PUJ, a SPJ or a training program. Furthermore he investigates a possible relation of the separation rate from a new job to whether or not the worker previously participated in an ALMP. In an event history model of labor force dynamics the author exploits information with respect to the duration of unemployment, the duration of the stay in an ALMP, the destinations after that, and the duration of subsequent employment spells.

In multivariate duration models, the variation in the durations at which treatment is administered to individuals along with data on the corresponding pre- and post-treatment durations can be exploited to identify the treatment effect. In order to account for possible selectivity in the inflow into ALMP, van Ours (2000) establishes a model considering the effect of ALMP on the transition rate from unemployment to a job and also the effect of ALMP on the separation rate once workers have found a job.

He finds that in Slovakia short-term subsidized jobs seem to be the most efficient active labor market policy. Workers that are or have been on a short-term subsidized job have a higher job finding rate than other unemployed workers, and once they find a job their job separation rate is lower than that of workers not having been on a short-term subsidized job. On the other hand, van Ours (2000) reports long-term subsidized jobs to have a negative effect on the job finding rate, and no effect on the job separation rate. His additional finding of a positive effect of training on the job finding rate is attributed to the possibility of reversed causality: some workers enter a training program only after they are promised a job. Training does not seem to affect the job separation rate.

3.5.5 Expert opinions

The complexity of the unemployment problem makes it highly unlikely that labor market experts will unanimously agree on the particular set of interventions that should be implemented to alleviate the situation. Each expert would presumably favor his or her own

mix of policy measures for tackling the unemployment problem. However, despite this lack of unanimity, among the most promising avenues for identifying appropriate policy measures is the attempt to extract expert consensus or agreement from the heterogeneous individual recommendations. This is the innovative avenue taken by Profit and Tschernig (1998). At a conference drawing together many labor economists familiar with the German labor market, these authors collected and analyzed a survey of expert opinions on the perceived desirability of various types of labor policy measures[35]. While the self-selected nature of the set of respondents is obvious, it is nevertheless instructive to review this analysis.

The responses to the questionnaire suggest that economic experts do not expect a single set of measures to work best in alleviating the unemployment problem. Instead, a wide variety of measures is implicated at being promising, among them increased investment into education and training in general, and increasing training and qualification programs in particular. Among the most favored measures were also interventions altering the incentives of labor market participants, albeit only represented by measures implying a more restrictive environment, namely a stricter administration of unemployment benefits, enhanced monitoring of the unemployed, and a reduction of unemployment benefit levels, not by increased subsidies for low wage earners or subsidies for promoting (self-)employment. The direct creation of public sector jobs was not among these most favored policy proposals.

In general, stronger distortion of individual decisions, for instance by increased centralization of wage bargaining, by stricter regulation of standard weekly hours or by the discouragement of overtime, did not appeal to the respondents. They were similarly disfavorable to a general expansion of public activity, both by expansionary monetary or fiscal policy. Instead, economists seem to favor the de-regulation of various aspects of labor and goods markets, for instance the de-regulation of small businesses and the de-regulation of part-time work. Certainly for the German labor market the actual experience with any of these concrete measures is limited. More importantly even, the empirical knowledge is extremely scarce, since historically there has not been any systematic attempt at a scientific evaluation of policy measures, let alone conceptually convincing experimental evidence. By and large, though, the international experience with labor market policies is relatively consistent with the responses analyzed in Profit and Tschernig (1998), although this result should certainly not be exaggerated.

[35] The authors repeated this exercise at a recent meeting of the European Economic Association, but due to technical problems the response rate was prohibitively low.

3.6 Collecting the Evidence

This section extracts the principal findings from the detailed review in the previous section. The main features of the specific country studies are summarized in Table 3.1.

Looking at the types of programs, we find a substantial variety across countries. Unsurprisingly, however, even though regulations differ greatly in their detail, a broad categorization into the two types of "training-based" programs and "subsidy-based" programs seems apt. The target group usually consists of unemployed individuals that either receive unemployment benefits (in most cases) or are at least eligible for the receipt. Many programs, in particular in Northern Europe, focus on youth unemployment. Few of the programs make finer distinctions in targeting, like e.g. the study by van den Berg and van der Klaauw (2000) that analyses Counseling & Monitoring for unemployed workers with "relatively good labor market prospects".

This study is also the only study in our review originating from a social experiment. All other papers are set in a nonexperimental context. Most of the studies analyze treatment effects on either employment (unemployment) rates or employment (unemployment) durations or hazards, respectively. The minority of studies (Bell et al. 1999, Larsson 2000, Lechner 2000, Raaum and Torp 2001, Regnér 2001) also considers wages as outcome variables of interest.

With respect to estimation methods, we find two methods of predominant use in Europe: duration models and matching methods. This also implies two fundamentally different approaches to underlying assumptions. Either one explicitly models the selection process and incorporates unobserved heterogeneity relying on some functional form assumptions (as in most duration analyses). Or one matches individuals under the conditional independence assumption claiming that one considers all relevant variables, and that selection is on observables (as in the analyses focusing on post-treatment employment rates). Both approaches have their advantages and disadvantages. However, there may be data situations where one method seems to be more appropriate than the other (e.g. matching in the case of Gerfin and Lechner 2000). Classic selection models – which played a major role in early evaluation research (cf., e.g., Björklund and Moffitt 1987) – nowadays seem to be of subordinate importance, and are rarely applied. This may be due to their exclusive focus on earnings as outcome of interest.

Table 3.1 Overview of recent European evaluation studies

Study	Country	Measure	Target Group	Type	Obser-vation period	Outcome of interest	Estimation method	Results	Notes
Larsson (2000)	Sweden	2 programs: Youth Practice and Labour Market Training	Young unemployed	Non-experimental	1991-1997	Annual earnings, re-employment probability, probability of regular education	Propensity score matching (multivalued treatment), OLS, probit	Both programs: short-run 0 to –, long-run 0 to slight +; youth practice better than labour training	Heterogeneity problems
Sianesi (2001)	Sweden	Various ALMP measures condensed into one "treatment"	Unemployed	Non-experimental	1994-1999	Various measures of labour market status, in particular employment probability	Propensity score matching	At best 0 effects (if cycling excluded), otherwise –	- Assumption of one "treatment" ? - Negative effects due to cycling behaviour
Regnér (2001)	Sweden	Training	Unemployed	Non-experimental	1987-1992	Earnings	Selection models: linear control, fixed effect, random growth	0 or – effects	Cycling behaviour
Raaum and Torp (2001)	Norway	Labour market training	Unemployed	Non-experimental	1989-1994	Earnings	Selection models: linear control, fixed effect, random growth (2)	+ for the linear control model, other models rejected	Internal comparison group

Table 3.1 [ctd]

Study	Country	Measure	Target Group	Type	Obser-vation period	Outcome of interest	Estimation method	Results	Notes
Jensen (1999) Jensen, Svarer Nielsen, Rosholm (2000)	Denmark	Youth Employment Program	Unemployed low-educated youth	Non-experimental	1996	Unemployment duration	Competing risks duration model	Sign. increase in transition rate U→S, weaker: U→E	U→S relevant question?
Rosholm (1999)	Denmark	Employment subsidy (public and private)	Unemployed (UI benefit eligible)	Non-experimental	1983-1990	Unemployment hazard, Employment hazard	Duration (MPH)	Private sector: U→E generally +, E→U strongly -, public sector: U→E mostly -, E→U strongly -	-UI benefits restored -Selection issue? -50% remain in subsidised firm -Stigmatisation (public sectors)
Bell, Blundell, van Reenen (1999)	UK	Temporary wage subsidy, training ("New Deal")	Young unemployed	Non-experimental	1997-1998	Productivity =wages	Trend-adj. difference-in-differences	Productivity effects relatively modest (comp. to size of subsidy)	Complementary method: ex ante general eq. model of labour market
van den Berg, van der Klaauw (2000)	Netherlands	Counseling & Monitoring	UI recipients (w/ relatively good labour market prospects)	Social experiment	1998-1999	Unemployment hazard	Duration models, limited dep. variable models	No sign. effect on individual transition rate U→E (still: program cost effective)	Choice of target group?

Table 3.1 [ctd]

Study	Country	Measure	Target Group	Type	Obser-vation period	Outcome of interest	Estimation method	Results	Notes
Brodaty , Crépon, Fougère (2001)	France	Youth employment programs: "workplace" training programs (private s.), "workfare" programs (public s.)	"The most disadvantaged and unskilled young workers"	Non-experimental	1986-1988	Employment status	Propensity score matching (multivalued treatment)	On-the-job training in private sector + (=higher amount of vocational & specific training)	
Lechner (2000)	Germany	Training and Retraining	Workers in East Germany	Non-experimental	1990-1994	Employment probability, earnings, career prospects	Partial propensity score matching	Short-term -, long-term 0, "waste of ressources"	- Built from scratch - Reduced official unemployment rate (main goal)
Lalive, Zwei-müller, van Ours (2000)	Switzerland	Benefit receipt cond. on ALMP participation	Unemployed UI recipients	Non-experimental	1997-1999	Unemployment duration	Duration model	Unemployment duration: men ↑, women ↓	
Gerfin, Lechner (2000)	Switzerland	Training (5 types), employment programs (private + public), temporary wage subsidy	Unemployed UI recipients	Non-experimental	1997-1998	Employment	Propensity score matching (multivalued treatment)	Temporary wage subsidy ++, employment programs –, training mixed	Excellent data base

Table 3.1 [ctd]

Study	Country	Measure	Target Group	Type	Obser-vation period	Outcome of interest	Estimation method	Results	Notes
Kluve, Lehmann, Schmidt (1999)	Poland	Training, IW (wage subsidies private sector), PW (public sector employment program)	Unemployed	Non-experimental	1992-1996	Employment rates, unemployment rates	Exact covariate matching	Training: men & women +, IW: women 0, men -, PW: men -	Benefit churning
van Ours (2000)	Slovak Rep.	Training, SPJ (Socially purposeful jobs), PUJ (publicly useful jobs)	Unemployed	Non-experimental	1993-1998	Job finding rate, job separation rate	Duration model	Short-term subs. jobs +, long-term subs. jobs -, training +	Training: reversed causality
Profit, Tschernig (1998)	Germany		Questionnaire to labour economists on what measure to apply best					No single measure, but various: +: Investm. in education & training in general, training & qual. programs in particular, incentives (stricter UI regulations) -: public sector job creation, subsidies for low wage earners	

Columns 1-9 are self-explaining. "Notes" refers e.g. to special features of the study, or to potential problems for interpreting results. All of these points are explained in detail in the corresponding text. The study by Bonnal et al. (1997) has been omitted from the table as the results are in line with Brodaty et al. (2001).

Regarding overall results, we find a surprising coincidence with the answers given by economists in the questionnaire of Profit and Tschernig (1998). Above all programs with a large training content seem to be the measures that are most likely to improve employment probability. Of course, this is not true for all types of courses (Gerfin and Lechner 2000), but given that both direct job creation and employment subsidies in the public sector almost always seem to fail (Rosholm 1999, Brodaty et al. 2001, Kluve et al. 1999) – in particular if they are only meant to keep unemployed off the register (Lechner 2000) – this is a rather robust result. In general, private sector programs seem to be far better than public sector programs. However, overall treatment effects appear to be rather modest, so that one should not expect too much from European ALMP.

In detail, we find three other results noteworthy: First, the highly positive effect of the temporary wage subsidy in Switzerland, a program that encourages job seekers to accept job offers that pay less than their unemployment benefit by compensating the difference with additional payments. As the income generated by the scheme is higher than the unemployment benefit for remaining unemployed, it is financially attractive for both the unemployed and the placement office. This seems to be a promising alternative measure of active labor market policy, and it would be interesting to see whether other countries would make similarly positive experiences with it.

Second, an approach of "Counseling and Monitoring" as analyzed by van den Berg and van der Klaauw (2000) seems promising. The fact that they do not find a significantly positive effect seems to be only due to an inappropriate choice of target group, as those unemployed "with relatively good job prospects" are most likely not the ones that would benefit most from C&M in their job search.

Third, we note that in quite many cases unemployment benefit regulations seem to be closely connected with program effects. There are two points to this issue. First, with respect to a tightening of rules it remains unclear, whether positive treatment effects can at all be induced by a measure that forces individuals into participation (cf. Lalive et al. 2000). Second, it seems to be a major distorting factor for treatment effectiveness if program participation restores benefit receipt eligibility. Given the substantial number of studies providing evidence for this hypothesis – Rosholm (1999), Kluve et al. (1999), Sianesi (2001), Regnér (2001) – this appears to be one of the most robust results of current evaluation research. In fact it is surprising that such regulations are still common practice in many European countries, as too generous unemployment benefit systems have frequently been

identified as one labor market feature in Europe associated with high unemployment (cf. for instance Nickell 1997).

3.7 Lessons for Economic Policy

Even before the emergence of this latest generation of European evaluation studies any broad overview of the European evidence would have suggested that one cannot conclude on any particular *Active Labor Market Policy* to yield consistently greater employment impacts than another. Throughout Europe there have been examples of studies supporting significantly positive effects on employment rates, but also some rather disappointing results. Frequently training has been estimated to have a positive impact on employment rates, but this has rarely been the case for its effect on wages.

Many of the more recent studies present a leap forward in terms of data quality and methodological rigor. Eventually, at least a few experimental studies have been performed or are about to get started. Moreover, nonexperimental analysis has been developed further, approaching – as far as the nonexperimental data allow – the credibility of experimental results. Unfortunately, these advances imply that many of the more optimistic results in the earlier European evaluation literature might be overstated.

The fact that the return to program participation varies tremendously across target groups, time, and place has been one of the most important insights from the many studies on the impact of labor market programs in the US. The recent generation of European evaluation studies confirms this conclusion. Not only are the results heterogeneous across economies, there is also convincing evidence that even within a country the precise formulation of target group matters dramatically for the results to be expected (cf. Bell et al. 1999, van den Berg and van der Klaauw 2000). Moreover, the economic environment and most importantly the regulations governing the social insurance system seem to be important elements in the performance of labor market programs. For instance, it may be very difficult to generate a positive effect on the future economic performance of program participants, if participation in the labor market program restores eligibility of unemployment benefits – then the major purpose for entering treatment may be *benefit churning*, not the genuine desire to improve one's labor market prospects.

This appears to be one of the findings that hold across economies. Besides,

unsurprisingly, it may well be that the idiosyncratic nature of labor market problems needs idiosyncratic approaches to their solution. Each European country certainly does have some distinct labor market difficulties that a common monocausal approach will fall short of addressing. There is no universal panacea in the set of ALMP that can cure unemployment for each and every country. And simple co-ordination of National Action Plans may not be sufficient. But nonetheless – despite heterogeneity of countries, of programs, of effects – current evaluation research can help avoid an ALMP practice of "shots in the dark", and does provide guidance with respect to sensible program implementation: At least we know some strategies that do not seem to work at all (tying together benefit receipt eligibility and program participation), and we know some schemes that might be worth trying for every European economy (Temporary wage subsidy, New Deal, C&M).

The attempt of the annual EU Joint Employment Reports at identifying examples of "good practice" takes a similar approach. What it absolutely lacks, however, is the deliberate evaluation effort. We have shown that both sides – policy and science – move in the right direction, but we are surprised – and discontented – to find that they do so separately. From the very first step, the planning and implementation of ALMP measures should go hand in hand with their evaluation. Only a thorough scientific evaluation accompanying the intervention can make an innovative program worthwhile. The Joint Employment Reports prove that European policy makers have the explicit desire to know about the effectiveness of labor market programs. And science has the means to provide just that knowledge. Finding for every single European economy one or more programs that work, rather than following the apparent practice of conducting untested large-scale interventions, might be an important avenue to a Europe less vulnerable to high and persistent unemployment.

At this juncture, one can only hope that European policy makers will increase their desire to include independent researchers in the effort of evaluating labor market policy interventions. Undoubtedly, this would revitalize and decisively support the feedback structure of the Joint Employment Reports, and thus help keep the Luxembourg Process on the right path.

Chapter 4

Active Labor Market Policies in Poland: Human Capital Enhancement, Stigmatization or Benefit Churning?

Together with Hartmut Lehmann and Christoph M. Schmidt

Abstract. This chapter provides microeconometric evidence on the effectiveness of Active Labor Market Policies in Poland. We sketch the theoretical framework of matching estimators as a substitute for randomization in labor market programs. Using retrospective data from the 18th wave of the Polish Labor Force Survey we implement a conditional difference-in-differences matching estimator of treatment effects. Treatment and control groups are matched over individual observable characteristics and pre-treatment labor market histories to minimize bias from unobserved heterogeneity. We also require that observations on controls are from the same regional labor market and from an identical phase of the transition cycle. Considering as the outcome a multinomial variable of labor market status, our first important finding suggests that training of men and women has a positive effect on the employment probability. For men public works and intervention works have negative treatment effects, while participation in intervention works does not affect women's employment probabilities. We attribute the negative treatment effects for men to benefit churning rather than to stigmatization of intervention and public works participants.

4.1 Background

The transition to a market economy had serious repercussions in the Polish labor market. Most significantly, open unemployment rose from virtually zero to a peak of around 16% at the end of 1993, declining slightly and hovering since then around 13%. Like in most transition countries unemployment in Poland can be characterized as a "stagnant pool" (Boeri 1994). This stagnancy came about because of very low outflow rates from unemployment and led to unemployment persistence and a large share of long-term unemployment. Individuals with unfavorable demographic and skill characteristics, i.e. workers who are old and have obsolete skills, whilst being an important element of the total stock of unemployment dominate long-term unemployment (Góra and Schmidt 1998 and Lehmann 1998).

Given this context, Active Labor Market Policies (ALMP) might in principle have an important role to play in combating unemployment in general and long-term unemployment in particular. Further training and re-training measures might help to solve skill mismatch, while subsidized employment in private and public firms and direct public job creation could be useful instruments in the re-building of human capital of some of the long-term unemployed. Measures of this kind are then meant to boost outflow rates from unemployment, in particular from long-term unemployment, thus raising labor turnover and improving the performance of the labor market. This rationale for labor market intervention by the government was developed in mature market economies. Whether it can be easily carried over into the context of a labor market in transition is an important and contentious issue that we do not further pursue here. This chapter has a more modest aim and focuses on the evaluation of the three most important Polish ALMP, i.e. publicly financed further training and re-training, intervention works[1] and public works.

Evaluation of ALMP in a labor market in transition has its own difficulties. In most transition countries rules for the assignment of ALMP measures and for the monitoring of the unemployed are either not well developed or not strongly enforced, leading often to large unforeseen distortions. Another difficulty is the absence of a stationary environment in which the evaluation takes place. In addition, the quality of the data used for the evaluation is often quite poor. These difficulties seem to be mirrored in the literature that exists on the evaluation of Polish ALMP. Góra et al. (1996) and Góra and Schmidt (1998) look at the rather loose

[1] "Intervention works" basically entails wage subsidies to boost employment of the unemployed in private or public firms.

application of assignment and monitoring rules in Poland and show some of the distortions arising from this. Most of the studies that have tried to econometrically evaluate Polish ALMP have certainly been plagued by data problems. These problems are mainly responsible for the not entirely convincing model specifications that underlie the impact analysis of Polish ALMP that have been undertaken in the past (cf. e.g. Puhani and Steiner 1997 and O'Leary 1998).

In our analysis we use data from the supplement to the August 1996 wave of the Polish Labor Force Survey (PLFS). This supplemental data set has a detailed retrospective part on monthly labor market histories over the period from January 1992 to August 1996 which can be linked to the August 1996 PLFS quarterly wave. This linked data set allows separate assessment of the three ALMP measures and lends itself to an evaluation procedure that matches controls to the treated individuals, both conditional on pre-treatment histories as well as on observable characteristics. Recent developments in the econometric evaluation literature suggest that this kind of matching seems to perform well in controlling for unobserved heterogeneity and self-selection (Heckman et al. 1997, 1998). Parallel to our study Puhani (1998) employs a similar but distinct variant of this matching approach.

The next section presents the theory underlying our matching approach. Section 4.3 gives a brief account of the three ALMP measures evaluated. Section 4.4 discusses the labor market histories in the August 1996 supplement and the matching algorithm, while section 4.5 looks at the empirical results. Matching estimates of treatment effects on the employment and unemployment rates as well as on employment retention and job accession rates are presented for the three ALMP measures, taking into account two regional taxonomies and two variants of matching analyzing short- and medium-term treatment effects, respectively. Section 4.6 concludes.

4.2 Application of Matching Methods

Two problems affect the evaluation of measures of Active Labor Market Policy in transition economies. The first is the usual evaluation problem that characterizes all nonexperimental analyses of interventions. Since counterfactual outcomes under no intervention cannot be observed for individuals receiving the intervention, one has to find an appropriate group of controls that, together with some identifying assumptions, facilitates the construction of the

desired counterfactual. Matching estimators have recently received a lot of attention in the econometric literature as one serious alternative nonexperimental evaluation approach (cf. Heckman et al. 1997, 1998, Angrist 1998). The second problem stems from transition itself. Interventions administered at different points of the transition cycle may have very distinct effects and nonexperimental controls observed at a different time period may not be appropriate. Since the interventions in our data are widely dispersed over the observation period, a matching approach that very stringently enforces the same temporal structure across intervention group and control group is a particularly promising evaluation strategy. This is the approach chosen in our application.

Our empirical work is in the spirit of Card and Sullivan (1988) who analyze the effect of the CETA training program on employment status by using conditional difference-in-differences matching estimators that match over labor market histories. We extend their analysis by considering a richer variable of labor force status (employment, unemployment and out-of-the-labor-force), by matching over observable individual characteristics and by adapting the analysis to the temporal structure of our data. All of their trainees received the treatment within a single year; however, they did not consider the timing and the duration of the treatment any further. In contrast, we establish the exact beginning and the duration of an intervention, and match controls accordingly.

The formal development of matching techniques, in particular the role of exclusion restrictions and of time-persistent individual heterogeneity has been discussed recently by Heckman et al. (1997 and 1998) who explicitly derive a non-parametric conditional difference-in-differences estimator of treatment effects. Another recent application of matching methods can be found in Lechner (1997) who analyzes the effects of training in the East German labor market.

The PLFS data provide information on labor force status at the individual level, together with information on individual and household characteristics. In the empirical analysis, the three ALMP interventions under scrutiny, *training*, *intervention works*, and *public works*, are considered separately. Thus, for purposes of the formal exposition, we only need to consider a single intervention. Furthermore, in our matching approach we will explicitly require that individuals who receive treatment are matched with individuals from the identical set of observed pre-treatment and post-treatment months. Any reference to the time period is therefore omitted from the exposition as well.

Denote the state associated with receiving the intervention with "*1*", and the state

associated with not receiving the intervention with "*0*". Assume that there are N_1 individuals in the intervention sample, with indices $i \in I_1$, and N_0 individuals in the sample of potential controls, with indices $i \in I_0$. Receiving the intervention is indicated by the individual indicator variable D_i (*1* = yes, *0* = no). Denote the potential labor market outcomes in post-treatment quarter q $(q = 1, 2, ..., Q)$ by Y_{qi}^1 if individual i received treatment, and Y_{qi}^0 if individual i did not receive treatment. These outcomes are defined as multinomials with three possible realizations ("*0*" = out-of-the-labor-force, "*1*" = employed, "*2*" = unemployed). While only one of these two outcome variables can actually be observed for each individual i, this being denoted Y_{qi}, the definition of potential outcomes allows for the formal construction of the unobservable counterfactual outcome $Y_{qi}^0 \mid D_i = 1$.

For purposes of evaluating the impact of the intervention, the post-intervention labor market success of each individual i will be summarized by the individual's average employment and unemployment rates, taken over the Q quarters following the intervention. Using indicator function $\mathbf{1}(.)$, these outcomes are $\frac{1}{Q}\sum_q \mathbf{1}(Y_{qi} = 1)$ for employment rates and $\frac{1}{Q}\sum_q \mathbf{1}(Y_{qi} = 2)$ for unemployment rates, respectively. These formulations extend those of Card and Sullivan (1988) from a binomial to a multinomial setting.

Using the indicator of intervention status D_i and the index $k \in \{1,2\}$ observed outcomes for individual i could be written as

$$(4.1) \quad \tfrac{1}{Q}\sum_q \mathbf{1}(Y_{qi} = k) \;=\; \tfrac{1}{Q}\left(D_i \sum_q \mathbf{1}(Y_{qi}^1 = k) + (1 - D_i)\sum_q \mathbf{1}(Y_{qi}^0 = k)\right) \quad ,$$

and the impact of the intervention on the average labor market status of individual i could be expressed as

$$(4.2) \quad \Delta_{ki} = \tfrac{1}{Q}\left(\sum_q \mathbf{1}(Y_{qi}^1 = k) - \sum_q \mathbf{1}(Y_{qi}^0 = k)\right)$$

for average employment rates $(k = 1)$ and for average unemployment rates $(k = 2)$.

Unfortunately, we can never observe Y_{qi}^1 and Y_{qi}^0 simultaneously for a given individual, and neither the joint distribution of the two outcomes across the intervention

sample. Instead, we have to focus on evaluation parameters for which we can construct counterfactuals by invoking appropriate identification assumptions. Our interest here is on the mean effects of treatment on the treated,

$$(4.3) \quad E(\Delta_{ki} \mid X_i, h_i, D_i = 1) \quad = \quad E\left(\tfrac{1}{Q}\left(\sum_q \mathbf{1}(Y_{qi}^1 = k) - \sum_q \mathbf{1}(Y_{qi}^0 = k)\right) \mid X_i, h_i, D_i = 1\right) \quad ,$$

that is the mean of the average employment and unemployment rates, respectively, over the population of the treated, conditional on observable individual characteristics X_i and previous labor market history h_i which is captured by a sequence of labor market states in the four quarters preceding the intervention. This slightly extends the discussion of Heckman et al. 1997 to a pair of evaluation parameters. Conditioning on previous labor market history was advocated by Card and Sullivan (1988) and by Heckman et al. (1997, 1998), accounting for the panel nature of their data.

More specifically, we will concentrate on average treatment effects over the joint support *S* of *X* and *h* given *D=1*,

$$(4.4) \quad M_k = \frac{\int_S E(\Delta_k \mid X, h, D = 1)\, dF(X, h \mid D = 1)}{\int_S dF(X, h \mid D = 1)} \quad .$$

In the absence of observations on the labor market status Y_{qi}^0 that recipients of the intervention would have realized had they not received the intervention, one needs to invoke appropriate identification assumptions in the construction of estimates for M_1 and M_2. The objective is to replace those expected values whose sample counterparts are unobservable by expected values whose counterparts can be constructed from sample data. In randomized experiments, if several conditions regarding the timing of the randomization, the process of sample attrition and the impact of randomization itself on individual behavior are met (cf. Heckman 1996a, Heckman et al. 1997), the counterfactual expected values under no intervention can be estimated for intervention recipients by the mean values of the outcome for randomized-out would-be recipients. Nonexperimental methods instead use data on non-recipient control groups to estimate the required counterfactuals.

The principal idea of matching is to assign to (preferably) all individuals *i* in the intervention sample as matching partners one or more individuals from the nonexperimental

control sample who are similar in terms of their observed individual characteristics (cf. Heckman et al. 1997). Within each matched set of individuals, one can then estimate the impact of the intervention on individual *i* by the difference over sample means, and one can construct an estimate of the overall impact by an average over these individual estimates. Matching estimators thereby approximate the virtues of randomization mainly by balancing the distribution of observed attributes across treatment and control groups, both by ensuring a common region of support for individuals in the intervention sample and their matched controls and by re-weighting the distribution over the common region of support.

As Heckman et al. (1997) point out, the strong assumptions traditionally invoked in the matching literature, conditional independence of the labor market status Y_{qi}^0 and of the treatment indicator D_i, given individual observable characteristics X_i, are not necessary to ensure identification of the mean effects of treatment on the treated. Instead, weaker mean independence assumptions are all that is needed for our matching estimators to identify the desired evaluation parameters,

$$(4.5) \quad E(\mathbf{1}(Y_{qi}^0=k)\,|\,X_i,h_i,D_i=1) \;=\; E(\mathbf{1}(Y_{qi}^0=k)\,|\,X_i,h_i,D_i=0) \quad .$$

That is, given the observable individual characteristics X_i and previous labor market history h_i that together form the basis for the individual matches, the fact that an individual received the intervention is assumed not to carry further information on the distribution of his or her no-intervention outcome.

In our empirical analysis, we use a variant of a nearest-neighbor matching estimator that implicitly performs a conditional difference-in-differences comparison between individuals in the intervention sample and their matched controls. For any treatment history h for which at least one match could be found, we estimate the impact of the intervention by

$$(4.6) \quad \hat{M}_{kh} \;=\; \frac{1}{N_{1h}} \sum_{i\in I_{1h}} \left[\frac{1}{Q}\sum\nolimits_q \mathbf{1}(Y_{qi}^1=k) \;-\; \sum_{j\in I_{0h}\,|\,X_j\in C(X_i)} \frac{1}{n_{i0}} \left(\frac{1}{Q}\sum\nolimits_q \mathbf{1}(Y_{qj}^0=k)\right)\right] \quad ,$$

where N_{1h} is the number of individuals with history h who receive the intervention ($N_1=\sum_h N_{1h}$), I_{1h} is the set of indices for these individuals, $C(X_i)$ is the appropriate

neighborhood of individual *i*'s characteristics X_i, and n_{i0} is the number of controls with history *h* who are falling within this neighborhood ($N_0 = \sum_i n_{i0}$), with the set of indices for control-individuals with history *h* being I_{0h}. The variance of this expression is then estimated as a function of the estimated probabilities from the underlying multinomial models[2].

The overall effect of the intervention is estimated in a last step by calculating a weighted average over the history-specific intervention effects,

$$(4.7) \qquad \hat{M}_k = \sum_h \left[\frac{N_{1h}}{\sum_h N_{1h}} \hat{M}_{kh} \right] ,$$

using the treatment group sample fractions as weights. The variance is derived as the corresponding weighted average of the history-specific variances.

The main impact of the intervention on labor market outcomes might arise from a positive one-shot effect at the end of the treatment period. One would like to know, however, whether workers who received the intervention were more successful in holding on to employment after the first post-intervention quarter than workers in the control group (cf. Card and Sullivan 1988). Hence, define the *job retention rate* as the probability of holding on to a job until post-intervention quarter *Q* conditional on being employed in the first quarter after treatment. The intervention effect is then

$$(4.8) \qquad \begin{aligned} E(r_i \,|\, X_i, h_i, D_i = 1) = E\,(& \mathbf{1}(Y_{2i}^1 = 1 \wedge \ldots \wedge Y_{Qi}^1 = 1) \\ & - \mathbf{1}(Y_{2i}^0 = 1 \wedge \ldots \wedge Y_{Qi}^0 = 1) \,|\, Y_{1i} = 1, X_i, h_i, D_i = 1), \end{aligned}$$

conditional on observable individual characteristics X_i and previous labor market history h_i. The average impact *r* over the joint support *S* of *X* and *h* given *D=1* is then defined similarly to the intervention effect in equation (4.4).

For pre-intervention history *h* we estimate the impact of the intervention on retention rates as

[2] Whenever feasible we based the estimation on unrestricted multinomial models.

$$(4.9)\qquad \hat{r}_h = \frac{\sum_{i\in I_{1h}} \mathbf{1}(Y_{1i}^1 = 1 \wedge Y_{2i}^1 = 1 \wedge \ldots \wedge Y_{Qi}^1 = 1)}{\sum_{i\in I_{1h}} \mathbf{1}(Y_{1i}^1 = 1)} - \frac{\sum_{j\in I_{0h}} \mathbf{1}(Y_{1j}^0 = 1 \wedge Y_{2j}^0 = 1 \wedge \ldots \wedge Y_{Qj}^0 = 1)}{\sum_{j\in I_{oh}} \mathbf{1}(Y_{1j}^0 = 1)}.$$

The variance of this expression is calculated using the delta method. The overall effect of the intervention, $\hat{r}$, is estimated as in equation (4.7) by calculating a weighted average of the history-specific effects, deriving the variance of this weighted average accordingly.

By raising their rates of access to a new job, the intervention might also exert a positive influence on workers who were unemployed at the end of the treatment period. Hence we define the *job accession rate* as the probability of starting a new job and holding on to it until post-intervention quarter Q conditional on being unemployed in the first quarter after treatment. The intervention effect is then

$$(4.10)\qquad \begin{aligned} &E(a_i \,|\, X_i, h_i, D_i = 1) = \\ &\quad E\Big(\textstyle\sum_r \mathbf{1}(Y_{1i}^1 = 2 \wedge \ldots \wedge Y_{r-1,i}^1 = 2 \wedge Y_{ri}^1 = 1 \wedge \ldots \wedge Y_{Qi}^1 = 1) \\ &\quad - \textstyle\sum_r \mathbf{1}(Y_{1i}^0 = 2 \wedge \ldots \wedge Y_{r-1,i}^0 = 2 \wedge Y_{ri}^0 = 1 \wedge \ldots \wedge Y_{Qi}^0 = 1) \,\Big|\, Y_{1i} = 2, X_i, h_i, D_i = 1\Big), \end{aligned}$$

conditional on observable individual characteristics X_i and previous labor market history h_i, where r is the first quarter of being employed, $1 < r \le Q$. The average impact over the joint support S of X and h given $D_i=1$ is then defined accordingly (see equation (4.4)).

For pre-intervention history h we estimate the impact of the intervention on accession rates as

$$(4.11)\qquad \begin{aligned} \hat{a}_h = &\frac{\sum_{i\in I_{1h}} \sum_r \mathbf{1}(Y_{1i}^1 = 2 \wedge \ldots \wedge Y_{r-1,i}^1 = 2 \wedge Y_{ri}^1 = 1 \wedge \ldots \wedge Y_{Qi}^1 = 1)}{\sum_{i\in I_{1h}} \mathbf{1}(Y_{1i}^1 = 2)} - \\ &\frac{\sum_{j\in I_{0h}} \sum_r \mathbf{1}(Y_{1j}^0 = 2 \wedge \ldots \wedge Y_{r-1,j}^0 = 2 \wedge Y_{rj}^0 = 1 \wedge \ldots \wedge Y_{Qj}^0 = 1)}{\sum_{j\in I_{oh}} \mathbf{1}(Y_{1j}^0 = 2)}. \end{aligned}$$

The variance of this expression is again calculated using the delta method. As for the other intervention effects, the overall effect of the intervention, $\hat{a}$, and its variance are estimated by

calculating the appropriate weighted averages of the history-specific effects and their variances.

4.3 ALMP Measures in Poland

The ALMP measures that we analyze – training, intervention works and public works – have been described at length in Lehmann (1998), Puhani and Steiner (1997) and Góra et al. (1996) for example. We therefore only briefly discuss these measures here. Besides presenting the evolution of expenditures on these measures during the period of interest (1992-1996) we concentrate on those institutional aspects of the design and implementation of the programs that are central in the context of this chapter.

Table 4.1 Distribution of resources between passive and active labor market programs [a]

	1992	1993	1994	1995	1996
Total expenditures[b]	2 282.75	2 370.23	2 527.87	2 796.59	2 825.39
PLMP	1 969.74	1 988.95	2 117.21	2 404.27	2407.97
ALMP	107.48	263.40	323.43	338.41	302.65
of which:	%	%	%	%	%
Intervention Works	43.74	38.54	43.02	41.25	34.39
Public Works	16.20	33.76	36.85	34.36	29.56
Training/Retraining	17.96	12.68	10.46	8.47	6.41
Loans (Self-employment)	14.91	9.03	6.11	5.91	6.84
Other	7.19	5.99	3.56	10.00	22.80
Participant inflows of major ALMP programs in % of labor force					
Intervention Works	0.8	1.2	1.8	2.0	1.5
Public Works	0.2	0.4	0.6	0.7	0.6
Training/Retraining	0.4	0.4	0.5	0.5	0.5

[a] in thousand Polish zlotys.
[b] in 1992 constant prices.
Source: Polish Ministry of Labor and Social Affairs.

Expenditures on labor market policies have only slightly risen over the period 1992-96 as can be seen in Table 4.1. Apart from 1992 when expenditures on ALMP amounted to only 5% of PLMP, the ratio of expenditures on the types of programs has been about 1:8 throughout the period. In an international comparison of Visegrad countries, Poland is roughly in line with Hungary that also spends predominantly on PLMP, but spends relatively less than the Czech Republic or the Slovak Republic. In relation to western OECD economies with similar unemployment levels, Poland spends little on ALMP and has low inflow rates into ALMP schemes. Of the three programs analyzed intervention works and public works have received the bulk of funds, while we see a monotonic decline of the relative fraction of expenditures going to training.

The main objective of *training* and *re-training* courses is to solve skill mismatch. By increasing the human capital of the unemployed in skills that employers in the expanding sectors want, the chances of the unemployed to enter a regular job are meant to increase and bottlenecks in the supply of certain skilled workers are meant to be eliminated. Popular courses are in the fields of data processing, accounting and secretarial work, as well as in tailoring and welding. The length of the courses is relatively short, in our sample the mean length being 2.6 months in the case of male and 2.5 months in the case of female trainees. The courses are organized by the local labor offices (LLOs) or by private agencies, which are then paid by the LLOs, or take place directly in firms. Trainees receive 115 percent of the amount of unemployment benefit, part of which has to be repaid if they do not complete the course.

Intervention works (wage subsidies) have two major goals. First, by hiring an unemployed person on a subsidized job he or she can enhance or regain human capital that might enable him or her to subsequently enter a regular job. Secondly, entrepreneurs can learn about the productivity of a worker without paying him or her a full wage. Incentives to the firm are structured in such a way that ensures the longest possible employment relationship. The longer a previously unemployed worker is kept in an intervention works slot the higher the cumulative subsidy going to the firm will be. Workers have an incentive to hold on to such a subsidized job for at least 6 months as, in the period under study, an employment relationship of this length entitled workers to another round of 12 months benefit receipt. The modal length of intervention works jobs in our sample is 6 months for men (63%) and women (48%), with very few jobs below this duration. So, most participants in intervention works qualify in principle for another round of benefit payment.

Public works are directly created public jobs that are mainly but not exclusively

targeted at the long-term unemployed. Like intervention works they are meant to enhance the human capital of participants, many of the jobs offered are, however, of a very low skill nature. The focus of these public works is the amelioration of the environment and the improvement of local infrastructure. Public works are organized by the LLOs in cooperation with municipal authorities. The incentive structures facing employers and workers are similar to the ones in connection with intervention works. The cumulative subsidy is larger the longer a previous unemployed is kept in a directly created public job, while such a worker qualifies for another round of benefit receipt if he or she remains in the job for at least 6 months. Not surprisingly the modal length of public works jobs is 6 months (53%).

Since the end of 1991 unemployment benefits have been limited to 12 months and been paid as a flat rate amounting to slightly below the minimum wage. Unemployed persons who exhaust their benefits have to rely on social assistance which is often, however, either only sporadically paid or paid out in the form of material help (Góra and Schmidt 1998). So, in many cases the only route to prolonged income support at a decent level is involvement in an ALMP measure, which entitles the unemployed to a further 12 months of benefit payment.

4.4 The Data: Labor Market Histories and the Matching Algorithm

The Polish Labor Force Survey (PLFS) is a quarterly survey, which was started in May 1992 and which has been structured as a rotating panel since its fifth wave (May 1993). Supplements on labor market policies were introduced in August 1994 and in August 1996. These supplements make it possible to generate a database for the evaluation of labor market policies. This chapter focuses on the 18th wave of the PLFS, taken in August 1996, in connection with its complementing supplement. Since we are interested in the effectiveness of ALMP measures offered by Local Labor Offices (LLOs) to the unemployed or to the previously unemployed, we use a sub-set of the full PLFS data set generated from this wave. For the construction of treatment and control groups we select those respondents who were registered at least once as unemployed between January 1992 and August 1996.

The supplement to the 18th wave includes individual labor market histories containing information on an individual's labor market state in every single month from January 1992 to

August 1996. Table 4.2 shows the various labor market states in these histories. Not all these labor market states are mutually exclusive, as e.g. training and registered unemployment, intervention works and employment are logical double entries. We re-coded such double entries very carefully to ensure consistency of the data.[3]

We created sub-samples of ALMP participants, choosing those who had been offered participation in training, intervention works or public works by their LLO and who had accepted the offer. The sample sizes for these sub-samples are 241, 532 and 93 respectively. We then generated a corresponding sub-sample of potential controls by selecting all those who at least once had been registered as unemployed since January 1992 and excluding all individuals in the ALMP sub-samples. This control sample has a size of 7784 records.

Table 4.2 Individual labor market history outcomes in August 1996 PLFS supplement

Code	Outcome
1	Employment
2	Temporarily not in work (for "objective" reasons)
3	Participation in training course
4	Social assistance recipient
5	Taking care of small child
6	Registered unemployed
7	Unemployment benefit recipient
8	Intervention works (wage subsidies)
9	Public works

Source: PLFS supplement August 1996

As was discussed formally in some detail in section 4.2, we match participants and controls not only across certain observable characteristics, but also across their pre-treatment history. Since we want to use as many treatment and control cases as possible we apply a "moving window" to the data as shown in Figure 4.1.[4] Given that an individual participated in a labor market program at a particular point in time for a particular number of months, we require a

[3] A detailed account of the transformation and re-coding of the data can be found in Kluve (1998). Exploiting the panel nature of the data, the author also shows in this study that recall error is a minor problem.

[4] Figure 4.1 and what follows focus on training. The same matching procedures apply for intervention works and public works.

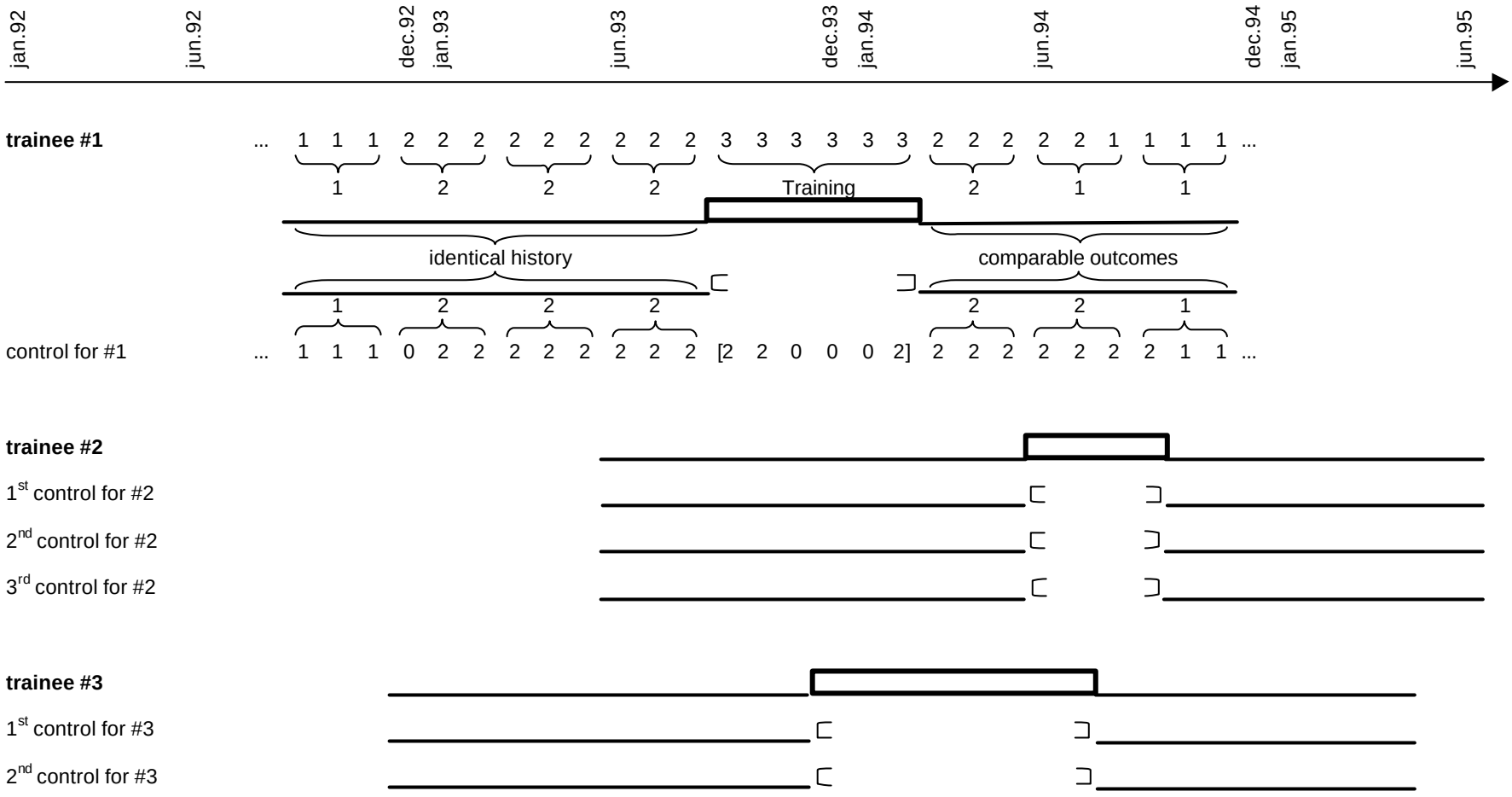
control to have an identical pre-treatment labor market history at the same point in time. Also, we compare employment outcomes for exactly the sequence of months that started when the participant's program spell ended. Since non-participants might have an advantage insofar as their job search activities are not restricted during the participants' program spells, evaluation of an ALMP measure at the individual level should take this into account. For example, the impact of a training measure for an unemployed will consist of two countervailing effects with respect to employment

On the one hand, during the training spell the unemployed individual will not be able to engage in job search as vigorously as his or her not-participating colleague, i.e. *ceteris paribus* participation lowers the probability of finding employment after the end of the program. On the other hand, training is meant to increase the participant's human capital and should therefore *ceteris paribus* increase the probability of finding employment after completing the program. In this study we are interested in the *net* overall impact that arises out of these two countervailing effects.

In matching across individual pre-program histories as well as in analyzing the post-treatment outcomes we are mainly interested in the labor market outcomes "employed" (="1"), "unemployed" (="2"), and "out-of-the-labor-force" (="0"). These realizations, "0", "1" and "2", are recorded on a quarterly basis, where quarters are those three-month-intervals either ending in the month immediately preceding training or beginning in the month immediately succeeding training. If those three months that form a quarter do not contain identical entries, the interval is assigned the value of the event appearing twice, e.g. "212" constitutes a "2". If the interval reads "021", it constitutes a "1", since being in a "successful" labor market state, i.e. being employed, during one month, constitutes a corresponding "successful" quarter. Extending conventional matching to a dynamic setting implies that an individual trainee's record has to meet the following requirements:

i. An entry "training" must exist for at least 1 month in the 56 months observed.
ii. The record must have a complete 12-month pre-training history.
iii. The record must have a complete 9-month post-training history (i.e. $Q=3$).
iv. The beginning and the end of the training spell must be defined.
v. Demographic information that we use in the matching algorithm must be complete.

Figure 4.1 Matching over identical individual labor market histories applying a "moving window"



As far as condition (i) is concerned we retain those observations with exactly one training spell and we select that training spell of those few individuals with multiple spells that is the longest and that has a complete pre- and post-training history. Requirements (ii) and (iii) imply that very early (1992) or very recent (1996) spells of training cannot be part of our analysis. The information stored in (iv) is crucial for our dynamic matching algorithm and essential for the control of macroeconomic effects, while the demographic information (v) includes the following categorical variables: gender (male/female), marital status (married/not married), education (high = university; low = primary school or below; medium = all other), and age.

Furthermore, in accordance with Heckman et al. (1997), who emphasize the need to control for local labor market conditions, we perform the matching algorithm for two different regional taxonomies. Taxonomy 1 considers a regional dummy distinguishing Warsaw/not Warsaw. This distinction takes account of a more dynamic labor market in the capital, while taxonomy 2 accounts for an exact regional match across all 49 voivodships. Theory predicts that taxonomy 2 will yield the least biased estimators. We will see, though, that matching conditional on exact voivodship matches reduces the number of participants who find matching partners in the control sample.

All those observations meeting the above-described requirements with respect to characteristics and, after the appropriate transformation from months into quarters, with respect to 4 pre-training and 3 post-training quarterly labor market outcomes constitute the data set of trainees for our analysis.

Matching proceeds then as follows. We match participants with all those controls who satisfy our requirements in terms of observable attributes and identical pre-treatment history.[5] In this, the 4-quarter pre-treatment history has to be identical for program participant and corresponding control group member, while the control needs to have a complete 9-month-history starting with the first month of the trainee's 9-month post-training history. Also, the control has to be identical in the following categorical variables: gender, marital status, education, and region.

If these requirements are met, we choose controls with the minimum distance in years of age. Most, i.e. 97.6 per cent of the matched controls, do not deviate more than 5 years from their treatment group matching partners, but we do allow for a maximum distance of 20 years.

This procedure, which has the most stringent matching requirements compared with other possible matching algorithms we employed, yields sufficiently large treatment and matched control samples. Due to the stringent matching requirements this procedure also generated the most robust results.

Using *Q=3* post-treatment quarters focuses on the short-term treatment effects. To estimate medium-term treatment effects we apply a second matching procedure ("second match") with the same stringent requirements, but where we compare labor market outcomes for *Q=6* post-treatment quarters.

4.5 Results

Basic demographic characteristics and labor market outcomes are described for the full sample and the various sub-groups that are used in the analysis in Tables 4.3 and 4.4. *Potential controls* are all those who between 1992 and 1996 were unemployed at least once but did not participate during this period in any of the three ALMP. Looking at men (Table 4.3), the demographic characteristics indicate that relative to the full subsets of trainees, intervention and public works participants (columns 3 – 5 in the upper panel) the group of potential controls (column 2 in the upper panel) is slightly younger than trainees and intervention works participants, but on average even more than four years younger than public works participants. Marital status is quite similar across the three participant groups, but for trainees the marriage rate is 7 percentage points higher than for controls.

The difference in educational attainment across treatment groups is striking: The fraction of trainees with non-compulsory education[6] is slightly higher than that among potential controls, but substantially lower among intervention works and public works participants. So, on this measure, unemployed individuals are targeted for training who have slightly more human capital than the average unemployed while for intervention works and public works we see individuals targeted with significantly less human capital than the average unemployed. The differences in observable characteristics are on the whole maintained as we move from the full subset of treatment groups to the smaller subsets of

[5] We apply a procedure of sampling with replacement, since we allow for an observation in the comparison group sample to control for more than one trainee, if he or she meets the necessary requirements. Such a constellation, however, rarely occurs.

treatment groups that are used for the two types of matches (columns 1 – 6 of lower panel).

One particularly interesting labor market outcome is the employment rate, which is shown here for August 1992 and August 1996. Recall that our matching algorithm requires the pre-treatment history to be 4 quarters long and the post-treatment history to have a length of at least three quarters. Hence, the two employment rate estimates shown are free from any bias arising from the participation in an ALMP scheme.

Inspection of these employment rates shows large variations across the various treatment groups and over time. Trainees have much higher employment rates than individuals who participate in other schemes. It is also noteworthy, that whilst the employment rate of the full sub-sample of trainees has risen by 3 percentage points from August 1992 to August 1996, during the same period it has fallen by 13 and 22 percentage points for the full sub-samples of intervention works and public works participants, respectively. A naive way to evaluate the three Polish ALMP measures would consist in constructing a difference-in-differences estimator based on these differences and the difference between the August 1996 and August 1992 employment rates of the set of potential controls. Such a rough difference-in-differences estimator would indicate that training raises the probability of employment for men by 4 percentage points, while intervention works and public works lower this probability by 12 and 21 percentage points respectively. Such a crude approach would, however, tell us little about the true impact of these programs at the individual level.

Women hardly participate in the public works program; Table 4.4 and all subsequent tables therefore only report the involvement of women in training and in intervention works. Regarding employment outcomes the same pattern of discrepancies in observable characteristics arises across the various sub-samples for women as for men. Inspection of Tables 4.3 and 4.4 generates additional interesting information regarding differences between men and women. In Poland women have a higher incidence of unemployment than men which is reflected in the larger female pool of those in this 18th wave of the PLFS who have been unemployed at least once between 1992 and 1996. Women tend to be represented more substantially in training schemes, whereas men dominate intervention works.

[6] Compulsory education is defined here as primary school attainment or less.

Table 4.3 Demographic characteristics and employment rates for full sample, potential controls and treatment group sub-samples – MEN

	Full Sample[a]	Potential Controls[b]	Treatment Groups (Full Sub-sample)[c]		
			Training	IW	PW
Sample Size	3829	3369	94	307	75
Demographic Characteristics:[d]					
1. Average Age	33.4	33.1	35.2	35.5	37.7
2. Fraction Married	0.586	0.581	0.649	0.625	0.613
3. Fraction of Non-Compulsory Education[e]	0.784	0.801	0.830	0.616	0.613
Employment Rates:[f]					
1992	0.686	0.697	0.671	0.504	0.647
1996	0.679	0.684	0.700	0.372	0.423

	Treatment Groups (Suitable for 1st Match)[g]			Treatment Groups (Suitable for 2nd Match)[h]		
	Training	IW	PW	Training	IW	PW
Sample Size	53	164	45	31	102	20
Demographic Characteristics:						
1. Average Age	36.6	36.5	37.2	38.3	36.4	36.7
2. Fraction Married	0.660	0.665	0.556	0.742	0.676	0.650
3. Fraction of Non-Compulsory Education	0.868	0.616	0.689	0.839	0.627	0.550
Employment Rates:						
1992	0.726	0.436	0.651	0.700	0.439	0.650
1996	0.755	0.410	0.556	0.742	0.436	0.650

[a] Individuals at least once registered as unemployed since January 1992 (Observations containing histories with less than 15 entries are omitted).
[b] Individuals at least once registered as unemployed since January 1992 who did not participate in an ALMP program.
[c] Individuals who participated in the corresponding ALMP program.
[d] At time of survey, i.e. August 1996.
[e] Excludes all individuals with primary school attainment or less.
[f] Employment rates are calculated for August 1992 and August 1996.
[g] Observations that were used for the short-term matching analysis ($Q=3$).
[h] Observations that were used for the medium-term matching analysis ($Q=6$).

Table 4.4 Demographic characteristics and employment rates for full sample, potential controls and treatment group sub-samples – WOMEN[a]

	Full Sample	Potential Controls	Treatment Groups (Full Sub-sample)	
			Training	IW
Sample Size	4230	3808	147	225
Demographic Characteristics:				
1. Average Age	32.7	32.6	32.6	33.6
2. Fraction Married	0.697	0.698	0.673	0.644
3. Fraction of Non-Compulsory Education	0.812	0.811	0.966	0.747
Employment Rates:				
1992	0.489	0.500	0.381	0.360
1996	0.477	0.472	0.529	0.396

	Treatment Groups (Suitable for 1st Match)		Treatment Groups (Suitable for 2nd Match)	
	Training	IW	Training	IW
Sample Size	68	111	42	71
Demographic Characteristics:				
1. Average Age	32.8	36.0	34.2	36.9
2. Fraction Married	0.676	0.694	0.714	0.732
3. Fraction of Non-Compulsory Education	0.956	0.676	0.929	0.690
Employment Rates:				
1992	0.476	0.387	0.550	0.380
1996	0.591	0.417	0.650	0.471

[a] See footnotes of Table 4.3. Public Works not reported due to small sample size.

The higher educational quality of female participants in both training courses and intervention works is striking. For example, taking the full sub-sample of treatment groups, in training and intervention works women have fractions of program participants with noncompulsory education that are 13 percentage points higher than the corresponding fractions of male participants.

Employment rates by contrast are substantially lower for women than for men. This is due to higher female unemployment rates as well as lower female participation rates during

the period under study. Finally, our naive difference-in-differences estimator would establish that for women training and intervention works raise the probability of employment by about 17 percentage points and 6 percentage points respectively.

Treatment effects are estimated as the average difference within matches of employment and unemployment rates during either three or six consecutive quarters after the intervention. The matching estimators are conditioned on observable individual characteristics, identical pre-treatment histories and the local labor market. The first estimator considers short-run effects, while the second estimator attempts to capture more persistent, medium-term effects of treatment.

Tables 4.5 and 4.6 illustrate for men and women, respectively, how the weighted total treatment effect of an ALMP measure on employment and unemployment rates is constructed from the detailed results for specific labor market histories. First, for each pre-treatment history a history-specific treatment effect is calculated as a simple average of the differences in the post-treatment rates of individual treatments and their matched controls. In this calculation, these individual differences are constructed as the differences between the outcome for the treated individual and the average outcome over all his or her matching partners.

Second, the history-specific estimates are condensed into an estimate of the overall effect. This overall effect is calculated as the weighted average of the history-specific estimates, where the weights are given by the fractions of the participants' sample. As regional taxonomy 2 is more disaggregated, the number of histories of individuals in the treatment group and of matching partners declines. However, the difference is not severe for the individuals in the treatment group, so the precision of the estimate of the total effect suffers only slightly as we reduce bias when moving from the aggregate to the more sophisticated regional taxonomy.

As can be seen in Tables 4.5 and 4.6, the vast majority of participants in intervention works was unemployed at least for one quarter during the year preceding the beginning of the treatment spell, and roughly 75 percent of participants were unemployed throughout the year preceding the beginning of their spell on intervention works. However, a small number of participants who were offered a slot on the scheme by their LLO comes from inactivity or from employment. As we define any measure offered by the LLOs as an ALMP measure we include these histories in our calculations. The dominance of the unemployment state in the pre-treatment histories also characterizes the other ALMP schemes.

Table 4.5 Average post-treatment employment rates and treatment effect by pre-treatment labor market history: short-term effects – Intervention Works – MEN

Regional Taxonomy 1

	treatment group			matched controls				
history	**N**	**rate**[a]	**std.err.**	**N**	**rate**	**std.err.**	**effect**[b]	**std.err.**
0000	2	0.000	0.000	9	0.208	0.287	-0.208	0.287
0022	1	0.000	0.000	1	0.000	0.000	0.000	0.000
1110	1	0.333	0.000	1	0.000	0.000	0.333	0.000
1111	10	0.900	0.095	184	0.727	0.141	0.173	0.170
1112	4	0.333	0.204	6	0.417	0.247	-0.084	0.320
1122	10	0.167	0.085	21	0.475	0.158	-0.308	0.179
1222	5	0.400	0.219	17	0.678	0.209	-0.278	0.303
2211	2	0.333	0.236	4	0.944	0.162	-0.611	0.286
2212	1	0.000	0.000	1	0.000	0.000	0.000	0.000
2221	1	0.000	0.000	1	1.000	0.000	-1.000	0.000
2222	125	0.107	0.025	496	0.385	0.044	-0.278	0.051
total[c]	162			741			-0.248	0.044

Regional Taxonomy 2

	treatment group			matched controls				
history	**N**	**rate**	**std.err.**	**N**	**rate**	**std.err.**	**effect**	**std.err.**
0000	2	0.000	0.000	2	0.500	0.354	-0.500	0.354
1111	10	0.900	0.095	12	0.783	0.130	0.117	0.161
1112	3	0.444	0.240	3	0.222	0.240	0.222	0.339
1122	4	0.083	0.072	4	0.417	0.247	-0.334	0.257
1222	2	0.000	0.000	2	0.833	0.264	-0.833	0.264
2211	2	0.333	0.236	2	0.833	0.264	-0.500	0.354
2222	100	0.127	0.030	108	0.385	0.049	-0.258	0.057
total	123			133			-0.236	0.051

[a] Average employment rate in the 3 post-treatment quarters.
[b] Difference between rates of treatment group and matched control group.
[c] Total effect = weighted average of effects for individual histories using participants' sample fractions as weights.

Even with a casual glance at Tables 4.5 and 4.6 one notes the widely differing treatment effects of the individual matches conditioned on pre-treatment histories. Not only do they have very different magnitudes, but also the estimated treatment effect often changes its sign as we go from one history to another. We also see that most individuals in the treatment group are concentrated in a few histories. In previous work (Kluve et al. 1998) we checked the sensitivity of the estimates of the overall treatment effects to the presence of these important histories. To this purpose we removed one of these important histories at a time and re-

estimated the overall treatment effects. The results of this procedure indicate that the total treatment effect estimates are quite robust: in the case of significant estimates removing an important history typically affects the magnitude of the overall estimates, while leaving the sign unchanged.

Table 4.6 Average post-treatment employment rates and treatment effect by pre-treatment labor market history: short-term effects – Intervention Works – WOMEN[a]

Regional Taxonomy 1

	treatment group			matched controls				
history	**N**	**rate**	**std.err.**	**N**	**rate**	**std.err.**	**effect**	**std.err.**
0000	4	0.417	0.217	46	0.137	0.172	0.280	0.277
0002	2	0.167	0.118	4	0.083	0.195	0.084	0.228
0022	1	0.667	0.000	1	0.000	0.000	0.667	0.000
1111	8	0.583	0.164	52	0.812	0.138	-0.229	0.214
1112	3	0.667	0.272	9	0.176	0.220	0.491	0.350
1122	7	0.381	0.171	8	0.476	0.189	-0.095	0.255
1222	2	1.000	0.000	4	0.333	0.333	0.667	0.333
2000	1	1.000	0.000	1	0.000	0.000	1.000	0.000
2111	1	1.000	0.000	1	1.000	0.000	0.000	0.000
2211	5	0.400	0.219	9	0.333	0.211	0.067	0.304
2221	2	0.000	0.000	2	0.667	0.333	-0.667	0.333
2222	73	0.247	0.046	643	0.263	0.052	-0.016	0.069
total	109			780			0.010	0.056

Regional Taxonomy 2

	treatment group			matched controls				
history	**N**	**rate**	**std.err.**	**N**	**rate**	**std.err.**	**effect**	**std.err.**
0000	3	0.556	0.240	4	0.333	0.272	0.223	0.363
0002	1	0.000	0.000	1	0.667	0.000	-0.667	0.000
1111	6	0.667	0.192	7	0.639	0.196	0.028	0.274
1112	2	0.500	0.354	3	0.083	0.195	0.417	0.404
1122	2	0.500	0.354	2	0.167	0.264	0.333	0.442
1222	2	1.000	0.000	2	0.833	0.264	0.167	0.264
2000	1	1.000	0.000	1	0.000	0.000	1.000	0.000
2111	1	1.000	0.000	1	1.000	0.000	0.000	0.000
2211	2	0.000	0.000	2	0.500	0.354	-0.500	0.354
2221	1	0.000	0.000	1	0.333	0.000	-0.333	0.000
2222	68	0.265	0.049	83	0.255	0.053	0.010	0.072
total	89			107			0.026	0.062

[a] See footnotes of Table 4.5.

The overall treatment effects of the three ALMP programs on employment and unemployment rates are displayed in Tables 4.7 and 4.8, where each total treatment effect estimate is constructed as shown in Tables 4.5 and 4.6. Let us first consider Table 4.7 that presents short-term effects.

Table 4.7 Overall treatment effects on employment and unemployment rates according to treatment, gender and regional taxonomy: SHORT-TERM effects

Training

		N	N	employment rate		unemployment rate	
		treatment	controls	effect	std.err.	effect	std.err.
Regional							
Taxonomy 1	all	118	956	0.005	0.051	0.015	0.052
	men	52	394	-0.041	0.079	0.024	0.080
	women	66	562	0.042	0.065	0.008	0.066
Regional							
Taxonomy 2	all	87	111	0.138	0.059	-0.092	0.059
	men	36	39	0.148	0.092	-0.139	0.091
	women	51	72	0.130	0.070	-0.058	0.073

Intervention Works

		N	N	employment rate		unemployment rate	
		treatment	controls	effect	std.err.	effect	std.err.
Regional							
Taxonomy 1	all	271	1521	-0.144	0.035	0.160	0.036
	men	162	741	-0.248	0.044	0.252	0.045
	women	109	780	0.010	0.056	0.025	0.057
Regional							
Taxonomy 2	all	212	240	-0.126	0.040	0.161	0.041
	men	123	133	-0.236	0.051	0.244	0.052
	women	89	107	0.026	0.062	0.045	0.063

Public Works

		N	N	employment rate		unemployment rate	
		treatment	controls	effect	std.err.	effect	std.err.
Regional							
Taxonomy 1	men	45	223	-0.156	0.078	0.159	0.078
Regional							
Taxonomy 2	men	33	35	-0.131	0.087	0.152	0.088

If matching is conditioned on the first regional taxonomy, no discernible effect of training on employment and unemployment rates can be observed. When local labor market conditions are seemingly better controlled for, we find a statistically significant positive overall treatment effect of training on the employment rate. This positive effect can still be established when the estimation is done separately for men and women, although the separate treatment effects for each gender are less well defined than the overall effect. Also, while training raises the average employment rate by some 15 percentage points for male trainees and by 13 percentage points for female participants, training measures have a negative, but statistically not significant impact on the unemployment rates of participants. So, the higher employment rates could result from training measures preventing workers from flowing out of the labor force rather than from lowering unemployment rates among active workers.

Irrespective of the regional taxonomy, according to our estimates intervention works have a large negative and statistically significant effect on the employment rate of men. This negative overall treatment effect of approximately minus 24 percentage points has its counterpart in a positive and significant overall treatment effect on the unemployment rate, which is, in absolute value, of the same magnitude. It is noteworthy, though, that women's employment and unemployment rates do not seem to be affected by the participation in this program. For men, the overall treatment effects of public works display a similar pattern as the effects of intervention work. Public works seem to depress the employment rate of participants and raise their unemployment rate, even if the magnitude of these effects is somewhat smaller.

Turning to Table 4.8, we find that in the medium term the employment rate of women is raised by 17 percentage points through participation in a training measure, while for men the overall treatment effect amounts to only 10 percentage points and is also not well defined. As in the short term the unemployment rates of both men and women are not affected by training in the medium term. On the other hand, males who participate in intervention works apparently have a more negative labor market experience even in the medium term. The overall treatment effect estimates are, however, in absolute value, some 7 percentage points lower than the effects in the short-term. A final interesting result from Table 4.8 is the large positive and statistically significant overall treatment effect of public works on the unemployment rate if we use the second regional taxonomy. The number of cases in the treatment and control groups is very low, though, and this result needs to be interpreted with some caution.

Table 4.8 Overall treatment effects on employment and unemployment rates according to treatment, gender and regional taxonomy: MEDIUM-TERM effects

Training

		N	N	employment rate		unemployment rate	
		treatment	controls	effect	std.err.	effect	std.err.
Regional							
Taxonomy 1	all	71	481	0.046	0.059	-0.024	0.059
	men	31	241	-0.010	0.089	-0.013	0.089
	women	40	240	0.090	0.075	-0.032	0.077
Regional							
Taxonomy 2	all	50	64	0.141	0.070	-0.123	0.071
	men	21	23	0.103	0.110	-0.119	0.108
	women	29	41	0.168	0.084	-0.125	0.089

Intervention Works

		N	N	employment rate		unemployment rate	
		treatment	controls	effect	std.err.	effect	std.err.
Regional							
Taxonomy 1	all	170	871	-0.075	0.039	0.112	0.039
	men	100	423	-0.182	0.050	0.202	0.050
	women	70	448	0.077	0.058	-0.016	0.060
Regional							
Taxonomy 2	all	128	146	-0.060	0.044	0.122	0.045
	men	73	80	-0.170	0.059	0.176	0.060
	women	55	66	0.086	0.062	0.051	0.067

Public Works

		N	N	employment rate		unemployment rate	
		treatment	controls	effect	std.err.	effect	std.err.
Regional							
Taxonomy 1	men	20	102	-0.142	0.094	0.146	0.094
Regional							
Taxonomy 2	men	13	14	-0.154	0.109	0.218	0.108

In summary, from an efficiency point of view training appears to be an ALMP program that performs well in Poland. Both men and women raise their chances of being employed in the short-term if they participate in this program, while women in particular benefit also in the medium term. Previous microeconometric work did not find such a beneficial treatment effect (cf. Puhani and Steiner 1997). In contrast, both subsidized employment (intervention works)

and direct public employment (public works) are highly inefficient when targeted at men. Whereas these measures are meant to raise the human capital of participants and thus *ceteris paribus* raise their employment rate, our estimates imply that they do exactly the opposite in the Polish case. It is certainly ironic, that the very ALMP program that seems to improve the performance of unemployed individuals in the Polish labor market, training, has experienced sharp expenditure cuts in recent years, while this has not been the case for the apparently ineffective intervention works and public works.

Additional information about the effects of active labor market programs is provided by employment retention and job accession rates. Given that an individual is employed in the first quarter following treatment, we ask whether the ALMP program has raised the probability that he or she will hold on to the job in the short-term or in the medium term (*job retention*). Given that the first quarter after treatment was spent in unemployment, we analyze the conditional probability of starting a new job and holding on to it (*job accession*). Since the employment stock at a particular point in time is crucially affected by job retention and job accession, looking at the overall treatment effects on these two rates might also help us better understand what lies behind the overall treatment effects on the employment rate. Given our definitions of retention and accession rates, the number of individuals that we match might be too small for serious statistical analysis. In the case of public works this is precisely what happens and we do not analyze this ALMP measure when estimating the treatment effects on retention and accession rates.

The short-term positive effect on the employment rate of both men and women that we found in the case of training and the short-term negative effect on the employment rate of men in the case of intervention works are reflected in the results reported in Table 4.9. The increased employment rate of male trainees appears to be supported by an improved retention rate, while the employment rate of female training participants apparently benefits from better access to jobs. Intervention works, on the other hand, significantly depress the job accession rate of men. If male participants find themselves unemployed in the first quarter after the end of the program, they have a much lower probability of flowing to a regular job than if they had not participated.

The medium-term estimates of the overall effects on retention and accession rates shown in Table 4.10 are in general less well defined. There are some interesting results, nevertheless.

Table 4.9 Estimated retention and accession rate treatment effects according to treatment, gender and regional taxonomy: SHORT-TERM effects

Training

		N	N	retention rate		N	N	accession rate	
		treatment	controls	effect	std.err.[a]	treatment	controls	effect	std.err.
Regional									
Taxonomy 1	all	107	932	0.022	0.066	109	946	0.125	0.065
	men	47	384	0.082	0.069	45	356	-0.004	0.084
	women	55	541	-0.018	0.095	60	555	0.207	0.068
Regional									
Taxonomy 2	all	76	99	0.136	0.104	85	109	0.231	0.090
	men	32	35	0.326	0.127	32	35	0.162	0.097
	women	40	59	-0.078	0.171	48	69	0.252	0.069

Intervention Works

		N	N	retention rate		N	N	accession rate	
		treatment	controls	effect	std.err.	treatment	controls	effect	std.err.
Regional									
Taxonomy 1	all	263	1513	0.076	0.056	265	1515	-0.100	0.025
	men	146	707	0.008	0.102	158	735	-0.121	0.032
	women	105	776	0.114	0.068	98	725	-0.067	0.045
Regional									
Taxonomy 2	all	204	231	0.109	0.074	209	237	-0.092	0.039
	men	112	122	0.130	0.124	121	131	-0.179	0.059
	women	80	97	0.011	0.098	81	98	0.023	0.051

[a] Standard errors are obtained by the delta method.

The treatment effect of training on the accession rate of women is still positive and large, though not significant at conventional levels. Most noteworthy is the large negative effect of training on the retention rate of women when the second regional taxonomy is used. This effect might be worrying as it seems to imply that the type of training courses women are offered makes it difficult for them to hold on to a job for more than a year. In contrast, there is a positive treatment effect of intervention works on the retention rate of women in the medium term. So, if women are retained after the treatment spell ended, which is one of the effects intended by the program, they have a greater probability of holding on to a job for at least 18 months than if they had not taken part in such a program.

Polish training measures seem to enhance the human capital of unemployed workers of either gender and thus improve their chances to find employment in regular jobs. In

contrast, for men, participation in either intervention or public works appears to lower the likelihood of finding regular work. One reason often given in the literature is that participation in such employment programs carries a stigma. Because of asymmetric information employers do not know the productivity of new workers, some of whom they might hire from the pool of the unemployed. Prospective employers might then perceive participants in such employment programs as low productivity workers or workers with tenuous labor market attachment.

Table 4.10 Estimated retention and accession rate treatment effects according to treatment, gender and regional taxonomy: MEDIUM-TERM effects

Training

		N	N	retention rate		N	N	accession rate	
		treatment	controls	effect	std.err.[a]	treatment	controls	effect	std.err.
Regional									
Taxonomy 1	all	60	455	-0.027	0.103	63	463	0.091	0.091
	men	27	235	0.021	0.153	29	239	-0.052	0.134
	women	30	216	-0.109	0.130	33	222	0.231	0.098
Regional									
Taxonomy 2	all	42	56	-0.017	0.163	45	59	0.163	0.143
	men	17	19	0.240	0.251	21	23	0.008	0.211
	women	23	35	-0.345	0.155	23	35	0.259	0.162

Intervention Works

		N	N	retention rate		N	N	accession rate	
		treatment	controls	effect	std.err.	treatment	controls	effect	std.err.
Regional									
Taxonomy 1	all	162	862	0.187	0.075	161	827	0.007	0.044
	men	91	408	0.105	0.133	96	417	-0.032	0.057
	women	65	442	0.237	0.091	55	371	0.031	0.081
Regional									
Taxonomy 2	all	116	133	0.177	0.104	123	140	0.044	0.055
	men	67	74	0.154	0.160	70	77	0.099	0.072
	women	46	56	0.194	0.155	45	55	0.046	0.096

[a] Standard errors are obtained by the delta method.

While such stigmatization might affect some workers in Poland, it cannot fully explain our results. If stigmatization were the full story, why do women placed on intervention works

apparently escape this stigmatization? A competing explanation of the negative treatment effects of intervention and public works for men could be benefit churning. There is widespread anecdotal evidence that officials in Polish LLOs place some of the unemployed into these schemes so that they re-qualify for benefit payment. In Table 4.11 we try to provide some evidence for this type of interaction of ALMP programs and unemployment compensation.

Table 4.11 Benefit churning - unemployment benefit situation of program participants

	training	intervention works			public works
	all	all	men	women	men
N total	121	275	164	111	45
N conditional[a]	23	130	89	41	19
No. in benefits - 1st month[b]	16	86	66	20	11
No. in benefits - 2nd month	12	93	70	23	12
No. in benefits - 3rd month	11	94	70	24	12
No. in benefits - 4th month	5	92	70	22	11
No. in benefits - 5th month	7	91	68	23	11
No. in benefits - 6th month	4	85	62	23	10
No. in benefits - 7th month	5	83	60	23	11
No. in benefits - 8th month	4	85	61	24	11
No. in benefits - 9th month	2	85	62	23	11
No. in benefits – all 9 months[c]	0	68	53	15	8
No. in benefits – at least 1 month[d]	16	102	76	26	13

[a] No. of program participants in each ALMP sub-sample conditional on a 6-month pre-treatment history with all 6 months unemployed and at least one of these months receiving benefits.
[b] No. of program participants conditional on ([a]) who received benefits in the first month after treatment.
[c] No. of program participants conditional on ([a]) who received benefits for all 9 months succeeding treatment.
[d] No. of program participants conditional on ([a]) who received benefits for at least one of these 9 post-treatment months.

The second row of Table 4.11 shows those participants in the various programs who, prior to treatment, were unemployed continuously for at least six months and who also received benefits for at least one month. The second to last row of the table gives the number of

participants who right after treatment were receiving benefits continuously for nine months. An individual who appears in both of these rows might be thought of as someone engaging in benefit churning. There is a circular flow that takes an individual from a long unemployment spell with some benefit payment through an ALMP program and then immediately after its termination back to another unemployment spell with continuous benefit payment for at least 9 months. Churning rates thus understood turn out to be 0 percent for trainees[7], 60 percent and 42 percent for male participants of intervention works and public works respectively, and 37 percent for female participants of intervention works.

While these back-of-the-envelope calculations are based on small numbers, the large fractions of male "benefit churners" still make the point convincing that benefit churning apparently contributes to a large extent to the poor performance of both the intervention works and public work programs. As income support for those on long unemployment spells is rather poorly developed in Poland (Góra and Schmidt 1998), officials of LLOs seem to consider males, often heads of households, particularly worthy to receive prolonged income support.

4.6 Conclusion

In this study we implement a conditional difference-in-differences matching estimator to evaluate, at the micro level, the effectiveness of three measures of Active Labor Market Policy in Poland: training and re-training, subsidized employment ("intervention works") and direct public employment ("public works"). Our approach is insofar innovative as we apply a "moving window" technique to the data to account for a changing macroeconomic environment. Most importantly, we match simultaneously on observable characteristics and pre-treatment labor market histories and thus ensure that selection bias and bias due to unobserved heterogeneity are minimized.

Individual treatment effects are estimated by estimating the difference in employment and unemployment rates of those subsets of treatment and control groups that have an identical pre-treatment history. Employment and unemployment rates are averaged over 3 post-treatment quarters to characterize short-term effects, while averaging over 6 post-

[7] Participation in a training scheme normally does not imply an employment relationship, i.e. does not carry with it a new entitlement of benefit receipt.

treatment quarters is the basis for analyzing the impact of these policies in the medium term. The overall treatment effect of an ALMP measure is then calculated from the weighted sum of the individual effects, where the weights are the fractions of the treatment group belonging to each pre-treatment history.

How effective are Polish ALMP programs? Training and re-training is the ALMP measure that performs well from an efficiency point of view. Our estimates suggest that the short-term post-treatment employment rates of both female and male participants are higher than they would have been had these individuals not participated in the program. Key ingredients of these results are higher employment retention rates in the case of men and higher job accession rates for female trainees. In the medium-term we see a statistically significant positive treatment effect only on the female employment rate; this rate, averaged over 6 post-treatment quarters, is raised through training participation by an estimated 17 percentage points. These beneficial effects of the Polish training and re-training program which could not be found in previous econometric work are in line with Puhani's (1998) findings who uses the same data set. So, this ALMP measure clearly seems to improve the efficiency of the Polish labor market and more resources should be dedicated to this program in future.

In contrast, the Polish employment programs seem to be burdened by major distortions. Despite their intention to enhance or rebuild the human capital of unemployed individuals, we find neither positive nor negative overall treatment effects for women who participate in intervention works, but find strong negative overall treatment effects on the employment rate of men who take part in intervention and public works. These negative effects are somewhat reduced from minus 24 percentage points to minus 17 percentage points as we move from the short-term to the medium-term perspective. Our estimates also show that we obtain corresponding positive overall treatment effects on the male unemployment rate that are, in absolute value, of the same magnitude as the negative treatment effects on the employment rate.

Combining this information with the evidence of a sharply depressed job accession rate for male participants of intervention works leads us to believe that Polish employment programs are often the intermediate stage between two spells of unemployment benefit receipt. We cite some numbers on this "recycling" of unemployment compensation recipients which takes place above all via intervention works. These numbers strengthen our conviction that while stigmatization might have some role to play, benefit churning explains most of the

negative overall treatment effects of these programs. Out of "social considerations" officials in LLOs deem males as heads of households particularly worthy of prolonged income support from the state. On our evidence, a reform of the Polish employment programs seems to be needed that eliminates the distortions arising from interactions between the unemployment compensation system and these programs.

Chapter 5

Disentangling Treatment Effects of Polish Active Labor Market Policies: Evidence from Matched Samples

Together with Hartmut Lehmann and Christoph M. Schmidt

Abstract. This chapter estimates causal effects of two Polish active labor market policies – Training and Intervention Works – on employment probabilities. Using data from the 18th wave of the Polish Labor Force Survey we discuss three stages of an appropriately designed matching procedure and demonstrate how the method succeeds in balancing relevant covariates. The validity of this approach is illustrated using the estimated propensity score as a summary measure of balance. We implement a conditional difference-in-differences estimator of treatment effects based on individual trinomial sequences of pre-treatment labor market status. Our findings suggest that Training raises employment probability, while Intervention Works seems to lead to a negative treatment effect for men. Furthermore, we find that appropriate subdivision of the matched sample for conditional treatment effect estimation can add considerable insight to the interpretation of results.

5.1 Introduction

The evaluation of active labor market policy (ALMP) in the transition countries of Central and Eastern Europe faces serious methodical obstacles. Most importantly, studies typically have to rely on nonexperimental data, a feature they share with most evaluation studies on measures of active labor market policy in OECD countries. In fact, nonexperimental settings are still predominant in any European country study, as large-scale – or any – experimental studies similar to those conducted in the US have remained highly uncommon.

Apart from this more general drawback early evaluation studies on transition countries frequently had to be based on yet inadequate data: certainly, first of all, local national statistics offices had to gather experiences in generating data sets. Moreover, as the urge to evaluate programs already emerged almost simultaneously with the introduction of the data sets and the introduction of the policy measures themselves, early studies could not exhaust any long-term data. And yet another distinct feature of policy evaluation in a transition country is the need to control for the – in early years after transition – quickly changing macro environment, in particular if one aims at estimation of individual treatment effects.

The transition countries of Central Europe display a U-shaped pattern of output over the first years of transition, showing an initial contraction in economic activity after the onset of reform followed by, in the Polish case, robust expansion (cf. Blanchard 1997). The effectiveness of ALMP measures depends – *ceteris paribus* – on the tightness of the labor market and, therefore, on the point on the U-curve where the economy is located. Evaluating the effects of ALMP measures administered over several years without controlling for the large moves along the U-curve observed in Central European transition countries would severely bias the results.

This study focuses on the evaluation of active labor market policy in Poland, with an emphasis on two major points. First, with regard to the implicit missing data problem in any nonexperimental evaluation study, we explore the potential of different matching procedures to achieve covariate balance, and we demonstrate how in our case exact matching methods may in an intuitively appealing way resolve the dilemma of constructing an adequate counterfactual. To this end we discuss three stages of a matching procedure that is meticulously adapted to the specific nature of the data. Our arguments are illustrated by comparing covariate balance and balance in estimated propensity scores – a summary measure of balance – across post-match samples.

Second, we discuss our evaluation results in detail, confirming earlier results on Polish ALMP (cf. Kluve, Lehmann and Schmidt 1999, Puhani 1998). We place particular emphasis on the necessity of considering subsets of the population of treatment units in the interpretation of results. We argue emphatically that a careful interpretation of results is as important as the devotion of effort to constructing an adequate comparison group, an idea that frequently seems to be overlooked in applied work. Specifically, we demonstrate that – even though an appropriate matching method does control for the relevant variables – once the comparison group is found, the analysis is not complete. Instead, pursuing the estimation of conditional treatment effects for appropriately defined subsamples may be useful to avoid otherwise misleading results.

The chapter is organized as follows: Section 5.2 presents a brief description of the data and gives a short exposition of the evaluation problem, showing how matching on covariates and/or the propensity score can identify the treatment effect. Section 5.3 explains how our matched samples were constructed and to what extent the matching methods applied succeed in balancing observable covariates. Section 5.4 focuses on developing our matching estimator of treatment effects, on interpreting treatment effect estimates, and on the importance of conditioning treatment effect estimates on covariates for interpretation purposes. Section 5.5 concludes.

5.2 Data and Methods

5.2.1 The Data

We employ data from the 18th wave of the Polish Labour Force Survey (PLFS) as of August 1996. The PLFS is a quarterly rotating panel introduced in May 1992. The distinct feature of the August 1996 wave is a supplementary questionnaire containing retrospective questions on individual labor market behavior. From these questions, individual labor market histories in quarterly structure have been constructed. The individual histories cover the 56-month-period from January 1992 to August 1996. Yet, the retrospective data required considerable preparatory work.

First, out of an initial number of 48,385 observations 11,102 individual labor market histories lacked any entry, and were omitted from the analysis. The vast majority of these are

individuals who were inactive in August of 1996. From the remaining data we had to exclude both treatment participants with too early (before January 1993) or too late (after November 1995) treatment spells since in our econometric approach we condition on pre-treatment histories spanning one year and look at post-treatment labor market outcomes averaged over three quarters. Incomplete spells containing too little information were also excluded from the analysis.

Our analysis focuses on individuals who experienced at least one spell of unemployment during the observation period. For both treated units and potential comparison units this ensures consideration of individuals potentially eligible for participation in ALMP measures offered by the employment offices. Since we focus on two distinct ALMP programs, Training and Intervention Works, the resulting samples of treatment participants for both measures and their potential comparisons are substantially smaller than the initial data set. We discuss sample composition in more detail in section 5.3.1.

Secondly, in order to be able to handle such rich data, we had to condense the information contained in individual labor market histories. Monthly entries entail, e.g., "employed", "unemployed", "receiving unemployment benefits", "maternal leave", etc. Furthermore, individual histories indicate whether and when an individual took part in an ALMP course. We compress the 30 possible monthly states occurring in the data into the three labor market states "employed" (henceforth denoted "1"), "unemployed" (denoted "2"), and "out-of-the-labor-force" (denoted "0"). Information on treatment participation is stored separately. Kluve et al. (1999) give a more detailed account of data transformation and adaptation. The resulting structure of individual spells for treatment and potential comparisons will be illustrated further in section 5.3.2.

In our estimation of individual treatment effects we consider two distinct measures of Polish ALMP: Training and Intervention Works[43]. For more information on institutional details, on ALMP regulations and descriptions of courses we refer to earlier papers on the topic (Kluve et al. 1999, Puhani 1998, Góra and Schmidt 1998). For our purposes in this study it is mainly important to note the distinct nature of the two programs. Training is meant to enhance, or at least sustain, individual human capital during a period of unemployment. The Polish Training measure for the unemployed is training off-the-job whose final aim is raising the unemployed person's probability of re-employment in a regular job.

[43] A third measure of Polish ALMP, Public Works (=direct job creation in the public sector), has been left out in this study for the sake of brevity, and due to small sample sizes. Cf. also Kluve et al. (1999), Puhani (1998).

Wage subsidy schemes like the Polish Intervention Works also have a human capital enhancing or -preserving aspect. However, the enhancement or preservation of a person's human capital takes place on-the-job. This human capital component of the program is thought to increase the chances of a participant to find regular, non-subsidized employment at the same firm or elsewhere after the end of the program. In addition, if there is asymmetric information about the productivity of potential employees, wage subsidy schemes are designed to facilitate temporary job matches that might translate into regular and lasting matches at the same firm once the subsidy ends. A crucial feature of ALMP regulation in the reported period, however, was that participation in Intervention Works was considered by the law like any other employment spell entitling individuals to a new round of benefit receipt, given the subsidized job lasted at least six months. Taking part in a Polish training measure for the unemployed did, on the other hand, not entitle a person to renewed benefit payments since this training was done off-the-job.

5.2.2 Matching as a Substitute for Randomization

Program evaluation aims at estimating causal effects, i.e. changes in the variable of interest that are due to treatment participation. The formal setting is cast into the statistical "potential outcome framework" for causal inference based on Neyman (1923 [1990], 1935), Fisher (1935) and Rubin (1974, 1977). Let us consider a population indexed by i, and let Y_{i1} denote the variable of interest given individual i participated in a program, indicated by $D_i=1$. Likewise, let Y_{i0} denote the outcome if $D_i=0$, i.e. if individual i was not a participant, and define the single unit treatment effect as $D_i=Y_{i1}-Y_{i0}$. However, outcomes Y_{i1} and Y_{i0} are "potential" in that we can never observe both of them simultaneously for one individual. The parameter of interest in nonexperimental studies is the mean effect of treatment on the treated population:

$$(5.1) \quad \Delta|_{D=1} = E(\Delta_i \mid D_i = 1) = E(Y_{i1} \mid D_i = 1) - E(Y_{i0} \mid D_i = 1)$$

The equation shows the inherent missing data problem, as we cannot observe the non-treatment outcome Y_{i0} for treatment participants $D_i=1$. We thus have to rely on establishing a convincing substitute for $E(Y_{i0}|D_i=1)$ in equation (5.1) in order to identify the desired parameter.

In an experimental study randomization ensures that potential outcomes Y_{i1} and Y_{i0} are

independent of treatment assignment D_i, i.e. $Y_{i1}, Y_{i0} \perp D_i$. Hence, program participants and comparison group do not systematically differ from each other, yielding the expectation of Y_{i0} for the comparison group as a substitute for the expectation of Y_{i0} of the treated group. Thus,

(5.2a) $$E(Y_{i0} \mid D_i = 1) = E(Y_{i0} \mid D_i = 0) = E(Y_i \mid D_i = 0),$$

where Y_i is the actually observed value of the outcome variable, i.e. $Y_i = D_i Y_{i1} + (1-D_i) Y_{i0}$. Thus, randomization ensures identification of the desired parameter $D|_{D=1}$ from equation (5.1). Randomization also implies an assumption referred to as stable-unit-treatment-value assumption (SUTVA, see e.g. Rubin 1980): Potential outcomes for each individual are not related to the treatment status of other individuals, i.e. $Y_{i0}, Y_{i1} \perp D_j \quad \forall\, i \neq j$.

Given a nonexperimental setting it appears appropriate to substitute for missing randomized-out controls by constructing a set of potential comparison units for whom we observe the same set of pre-treatment covariates X_i as for the treated units. The following proposition given in Rubin (1977) extends the above framework to nonexperimental studies:

> *If for each unit we observe a vector of covariates X_i, and $Y_{i0}, Y_{i1} \perp D_i | X_i$ holds $\forall\, i$, then the population treatment effect for the treated $D|_{D=1}$ is identified: it is equal to the treatment effect conditional on covariates and assignment to treatment $D|_{D=1,X}$ averaged over the distribution $X|D_i=1$.*

Such a construction of counterfactual outcomes can only be sensible if conditioning is on variables that themselves are not the outcome of treatment participation. Post-treatment employment success is a case in point: by matching individuals who are or are not successful the effect of treatment will necessarily be derived to be zero. Similar conceptual reservations would hold for characteristics of post-treatment jobs such as industry or working hours.

Consequently, conditional on observable covariates assignment to treatment can be considered as having been random, and unobservable characteristics possibly influencing treatment participation are ruled out. In fact, by this proposition comparing a program participant with a comparison individual displaying the same observable characteristics is like comparing the two in a randomized experiment. We thus merely need to estimate $E(Y_{i0} \mid X_i, D_i=0)$, so that

(5.2b) $E(Y_{i0} \mid D_i = 1) = E_X\left(E(Y_{i0} \mid X_i, D_i = 0) \mid D_i = 1\right),$

identifying the mean effect of treatment on the treated of equation (5.1) for a nonexperimental setting: constructing the appropriate weighted average over conditional (on X) no-treatment outcomes mimics randomization by balancing all relevant covariates.

Ideally, in order to implement a procedure for estimating the conditional treatment effect $D|_{D=1,X}$, we could simply match treated and comparison units on their covariate vector X_i. While exact matching on X_i achieves an exact balancing of attributes, it suffers from the fact that X_i might be of high dimension or contain continuously-distributed variables, so that some treated units might not find comparisons. To avoid the problem of matching on a high-dimensional X_i, the method of propensity score matching has been proposed by Rosenbaum and Rubin (1983). Define the propensity score as $p(X_i)=Pr(D_i=1|X_i)=E(D_i|X_i)$, i.e. the conditional probability of receiving treatment given a set of covariates. Then the conditional independence result from above extends to the propensity score: $Y_{i0},Y_{i1} \perp D_i|X_i \Rightarrow Y_{i0},Y_{i1} \perp D_i|p(X_i)$.

The reduced dimension comes at a cost, however. The propensity score is not known and has to be estimated. Also, in samples of limited size, for some i and j it may occur that $p(X_i)=p(X_j)$ even if $X_i \neq X_j$, resulting in imperfect balancing of the distributions of covariates. Thus, the small sample performance of propensity-score matching might be quite dismal. In fact, the literature indicates that the trade-off between exact matching and propensity score matching[44] is one of truly empirical nature: The decision for one approach or the other should depend heavily on the data, e.g. the number of observations, the dimension of X, time structure of variables, etc., and certainly it should depend on what the researcher believes (and justifies) to be the adequate *modus operandi* in each specific case.

Angrist and Hahn (1999) make this point forcefully by stating that existing theory provides little in the way of specific guidelines as how to choose between the two. On the one hand Hahn (1998) proves that exact matching is asymptotically efficient while propensity score matching is not, and concludes that asymptotic arguments would appear to offer no justification for anything other than full control for covariates. On the other hand Angrist and Hahn (1999) show that in some plausible scenarios estimators controlling for the propensity

[44] In practice, matching algorithms are manifold, including e.g. exact matching, matching within calipers (fixed or flexible), minimum-distance matching, or optimal full matching minimizing total distance. For further reference see Gu and Rosenbaum (1993), Rosenbaum (1995a), and Augurzky (2000).

score can be more efficient than exact-matching estimators. The latter seems to be valid in particular when cell-sizes are small, the explanatory value of the covariates is low conditional on the propensity score, and/or the probability of treatment is far from ½.

Still, what counts in practice is how well balance is achieved, so that the researcher can indeed "compare the comparable" (Heckman et al. 1997). Any matching procedure allowing for any distance in either X or $p(X)$ must be aware of that. And including a weak predictor of $p(X)$ into the estimation might be more harmful than covariate matching on a reduced set of comparisons. Thus, both Dehejia and Wahba (1998) – based on an empirical study – and Augurzky and Schmidt (2000) – based on a simulation study – argue that it is more important to achieve balance of relevant covariates rather than painstakingly modeling the selection process.

5.3 Analyzing Matched Samples

5.3.1 Composition of Matched Samples

For each of the two measures under scrutiny – Training and Intervention Works – we start the construction of matched samples from an initial sample consisting of treated individuals and untreated potential comparison individuals, where every observation is required to have at least one spell of unemployment. From this starting point we subsequently impose stronger restrictions on X (i.e. enlarge the dimension of the matching criteria) step-by-step, in order to obtain three samples of matched treatment-comparison units for each of the two measures:

> *Sample A: A comparison unit is matched to a treated unit if his or her labor market history is observed without substantial gaps from a year before up to the beginning of treatment and from the end of treatment until 9 months later. None of the observed individual characteristics is used as a matching criterion.*
>
> *Sample B: A comparison unit is matched to a treated unit if requirement (A) is met, and if he or she is identical in observable characteristics age, gender, education, marital status, and region.*
>
> *Sample C: A comparison unit is matched to a treated unit if requirements (A) and (B)*

are met, and if he or she displays an identical 4-quarter (12-month) pre-treatment labor market history at the exact same point in time as the treated unit. [45]

Samples (A) through (C) are constructed applying an exact-matching-within-calipers algorithm. For all three samples, if a treated individual finds any matching partner among the potential comparisons, this observation is retained. All algorithms allow for an oversampling procedure, i.e. a treated unit may be assigned more than one comparison unit. While we could have sharpened the matching criteria in a different order, this sequence reflects our conviction that *timing* is the pivotal aspect of comparison group construction in a transition economy.

The firmness in requirements (A) to (C) increases substantially. While under the weak precondition of Sample (A) no treated unit is lost in the matching process, and almost all potential comparisons are used, under requirement (C) some treated units do not find matching partners, and the number of matched comparison units is far smaller. Thus, algorithm (C) proceeds with replacement: some comparison units are matched to more than one treated individual. Samples (A) and (B) are constructed from potential comparison units with replacement, too, but here we use only the join of sets over matched comparison units.

Table 5.1 presents resulting sample sizes, as well as means of relevant variables. We observe that there is a reduction in the number of treated units who find matching partners from (A) to (C) of almost one third for Training, and almost one quarter for Intervention Works. Due to matching-with-replacement, samples (C) contain comparison units matched to more than one treated unit. With less than one percent, the number is very low for Training, and with approximately one tenth it is also fairly low for Intervention Works. Table 5.1 also shows that Training participants on average are better educated, somewhat younger and more likely to be female than Intervention Works participants.

Throughout, we focus our attention on exact matching procedures. In sample (B), the number of matching variables is limited, and they are all categorical variables. Moreover, exact matching performs quite well: despite the substantial number of cells, approximately 9 out of 10 of treated units find a comparison unit. With regard to sample (C), our exact matching approach is a very practical device to account for the pre-treatment employment sequence. Further illustration is provided in the next sections.

[45] We consider 6 age categories, 3 education categories, gender, marital status, and 49 regions, resulting in 3528 different cells for sample (B). Including a 4-quarter sequence of a trinomial labor market outcome variable (cf. section 5.3.2) increases the number of cells to $3528*3^4=285{,}768$ cells for sample (C).

Table 5.1 Composition of matched samples

		Training		Intervention Works	
		treated	**untreated**	**treated**	**untreated**
Initial Sample	Observations	121	7177	275	7177
Sample A	Observations	121	6751	275	6757
	age	34.5	33.1	36.3	33.1
	%education[a]	91.7	80.7	64.0	80.7
	%female	56.2	53.0	40.4	53.0
	%married	66.9	65.8	67.6	65.6
Sample B	Observations	114	983	244	1354
	age	34.0	33.0	36.0	34.7
	%education	93.9	98.9	69.3	87.4
	%female	56.1	62.1	40.6	51.9
	%married	65.8	23.2	70.5	77.8
Sample C	Observations	87	111	212	240
	[Individuals][b]		[110]		[211]
	age	33.4	33.8	36.0	35.2
	%education	96.6	97.3	71.2	74.2
	%female	58.6	64.8	42.0	44.6
	%married	67.8	70.3	70.3	70.4

[a] Excluding individuals with only primary school attainment or less.
[b] Number of observations that the algorithm matched exactly once.

5.3.2 Timing of interventions

In our preferred sample (C) we require treated and matched comparison units to display an identical pre-treatment history. To achieve comparability across samples (A) to (C), we impose the requirement on samples (A) and (B) that we observe any history at all in the year preceding treatment, although the precise information *what* history was experienced exactly is not used in matching. Moreover, to allow an assessment of post-treatment labor market performance, we require for treated units and all comparison samples that we observe a post-treatment sequence of labor force status variables in the nine months after treatment. In accordance with our preparatory data work, we condense the monthly information for treatment units to a sequence of three quarters of a multinomial outcome variable (0,1,2) denoting labor force status (out-of-the-labor force, employed, unemployed).

Correspondingly, for those comparison units eventually matched to the treated units, a comparable three-quarter post-treatment multinomial sequence of labor force status is computed as well, again starting at the exact point in time when the treatment spell of the corresponding treated unit ended. Our analysis thus incorporates individual treatment duration by conditioning on a complete (i.e. without major gaps) pre-treatment labor market history being observed before month "start" and comparing labor force status outcomes after month "stop". Thus, treated units and matched comparison units are always being compared during the same period. Figure 5.1a illustrates this procedure for samples A and B, in which the timing structure is considered, but the contents of individually matched labor force status histories does not matter. Figure 5.1b proceeds to depict the case for inclusion of exact pre-treatment histories in matching for sample (C)

We thus take advantage of the specific nature of the data with monthly information on employment status for a 56-month period, considering the exact timing of "start" and "stop" of treatment – a feature that is neither common nor possible in many studies, even those focussing on duration data. Moreover, given the rapid upward moves of the Polish economy along the positive section of its U-shaped curve of output between 1992 and 1996, we can assume that labor market tightness has increased in Poland in the reported period. Hence, the fact that we are able to compare treated and comparison units individually at the same point of time seems particularly valuable.

There might be other ways to solve the crucial problem of finding the "starting point of treatment" for comparison units. In principle, one could first match on characteristics X or the propensity score conditioned on characteristics, $p(X)$, and then directly impose requirements on comparable timing. A procedure following such a "partial balancing score" is for instance used by Lechner (2000). It seems more natural to us, however, to incorporate timing as a principal component of matching.

5.3.3 Covariate Balance

In section 5.2.2 we have emphasized that balance in all relevant factors – observed as well as unobserved – is the principal objective in experiments, and in its observational counterpart, the matching approach. In this section we examine whether the particular matching procedures we applied here indeed succeed in balancing the distributions of pre-treatment covariates between treatment units and their comparisons. Figures 5.2 and 5.3 show the distributions of the two principal covariates age and region for treated and comparison units

Figure 5.1a Matching applying a "moving window" in samples (A) and (B)

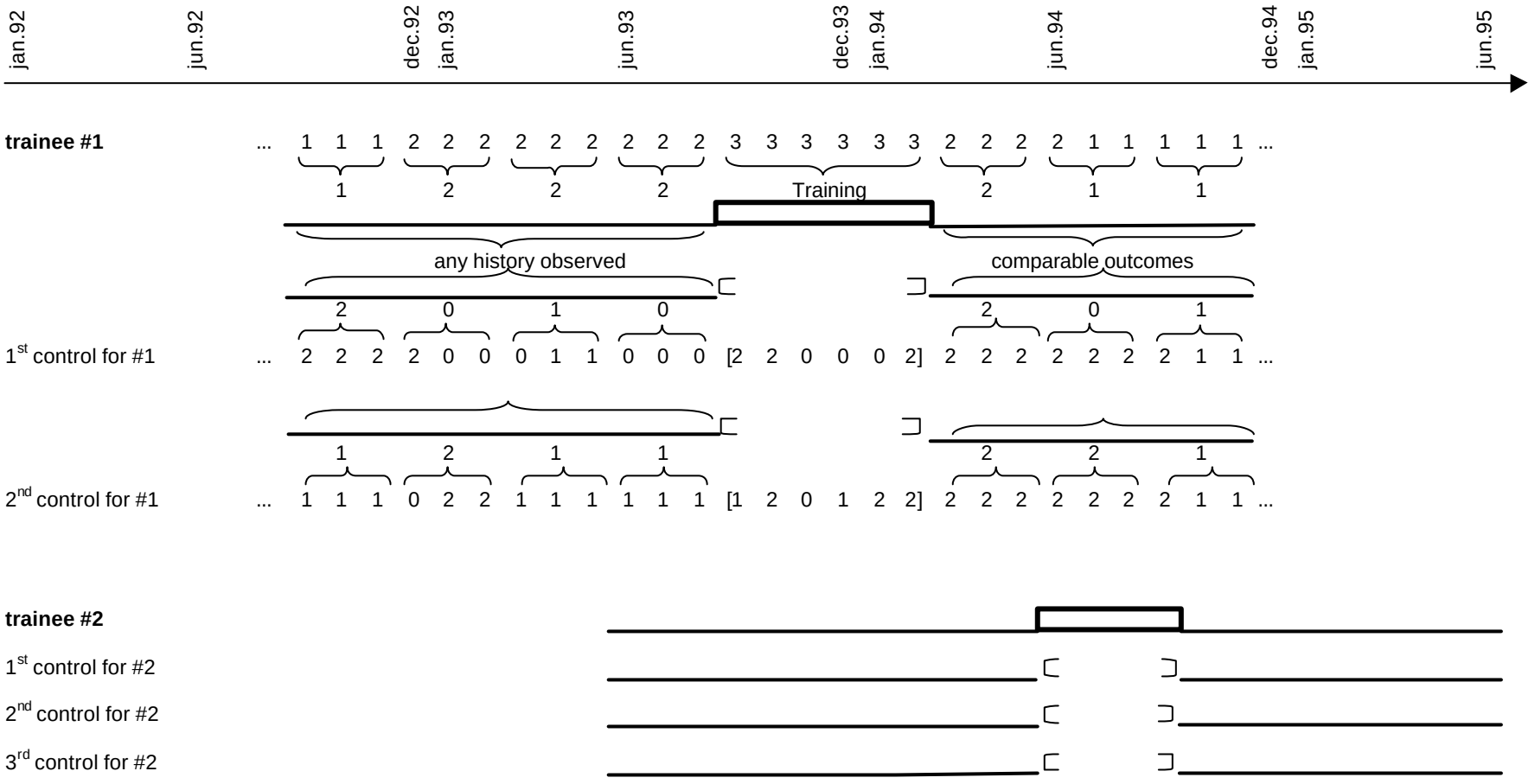
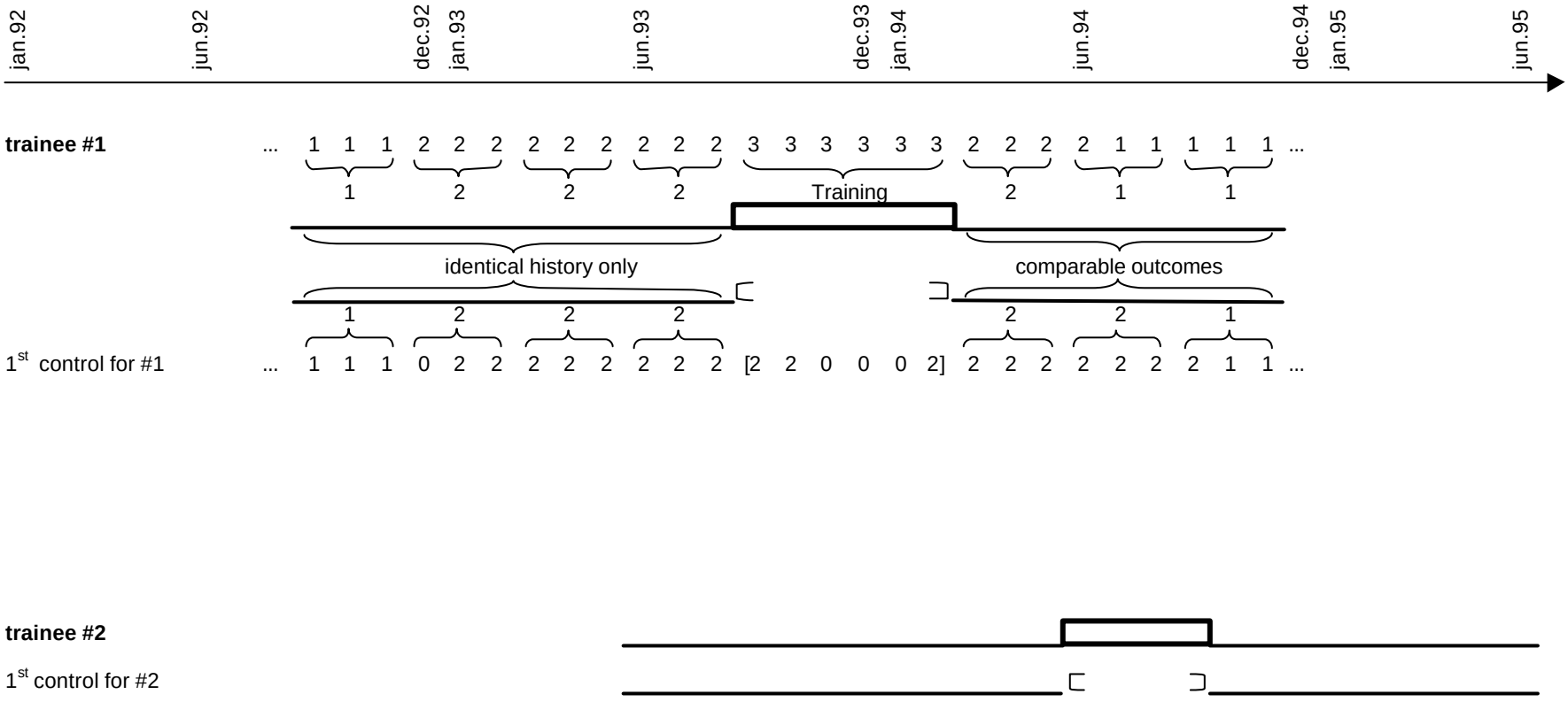


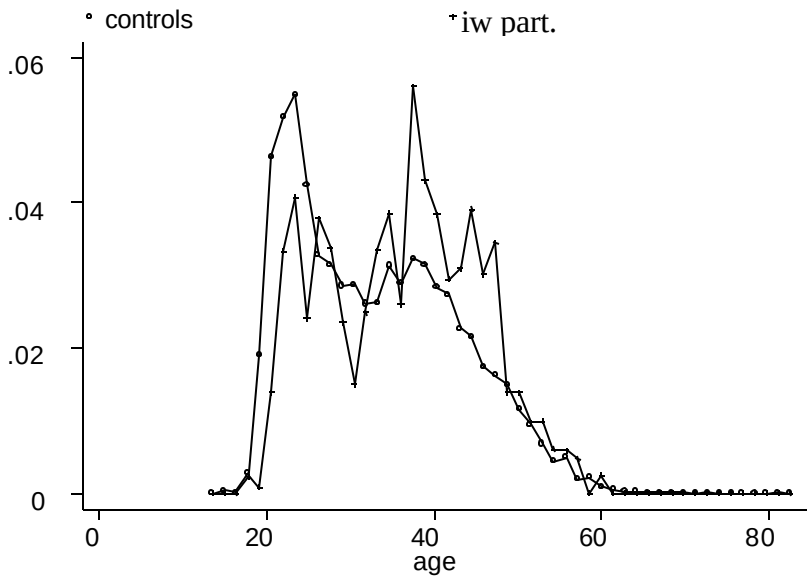
Figure 5.1b Matching over identical individual labor market histories applying a "moving window" in sample (C)



when matching is according to requirements (A) and the analyzed treatment is Intervention Works. By contrast to sample (A), samples (B) and (C) match on these individual characteristics. The figures illustrate by how much matching on the correct timing alone would miss out on balancing individual characteristics.

Figure 5.2 Distribution of age – Intervention Works

Sample A



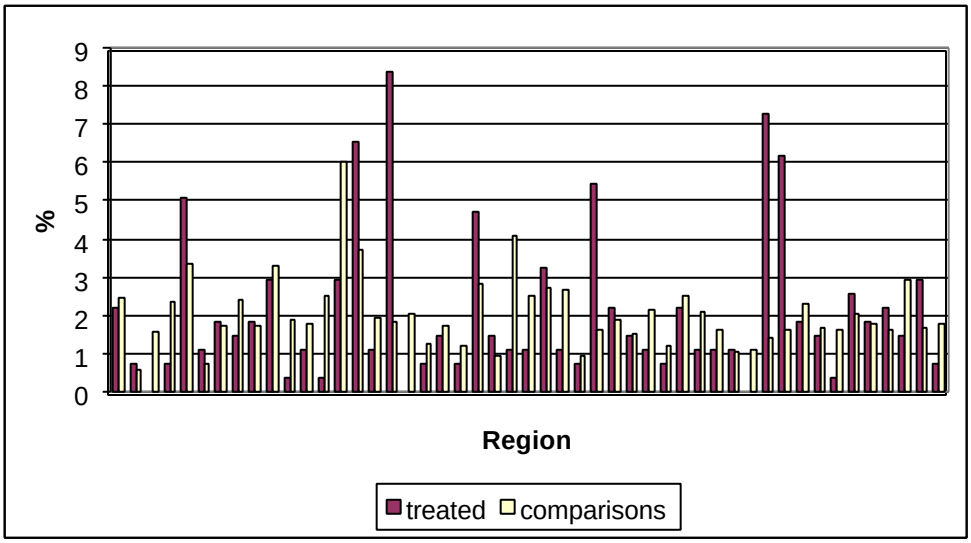
Kernel density estimates of the relevant variable for treated and comparison units by STATA using an Epanechnikov kernel and total bandwidth of (.5). Density estimates are not bound, their purpose is for illustration only.

Figure 5.2 shows that if not accounting for age, the young would be over-represented among the comparisons, and the mature (35-50, say) workers would be over-represented among the treated units.

Figure 5.3 plots the frequency distribution for the 49 Polish voivodships. Including regional indicators among the matching covariates is firmly advocated by Heckman et al. (1997) in order to control for the local labor market. This is the more imperative in the Polish case, since local labor market conditions are quite heterogeneous in any typical transition country. The matching criteria for samples (B) and (C) achieve complete balance – besides oversampling of comparison units – in the distribution of voivodships for treated and comparison units, while sample (A) displays considerable imbalance. Thus, if regional information were left out of the matching algorithm, regional balance would not be assured.

Figure 5.3 Distribution of region – Intervention Works

Sample A



Region = 49 voivodships.

With respect to further socio-demographic characteristics, 59.6% of Intervention Works participants are male, while there are only 47% men in comparison sample (A). Regarding the three education categories, the middle category comprises 63.6% of Intervention Works participants, and there is only one single individual out of the 275 treated (=0.36%) in the top category. Among comparison units in (A), 2.4% and 78.4% are in the top and middle categories, respectively. Table 5.1 shows that sample (B) and in particular sample (C) achieve balance in terms of sex and education.

5.3.4 Pre-Treatment Histories

The literature on program participation has always been concerned with the focal problem of controlling for observable characteristics, unobserved heterogeneity, and selection bias. Mainly affecting a difference-in-differences estimation approach, Ashenfelter (1978) pointed to a potentially serious limitation of this procedure when he observed a relative decline in pretreatment earnings for participants in subsidized training programs. This empirical regularity has been called "Ashenfelter's dip" and has been confirmed by subsequent analyses of many other training and adult education programs (cf. Bassi 1983, Ashenfelter and Card 1985, LaLonde 1986, Heckman, LaLonde, and Smith 1999). For instance, Ashenfelter and Card (1985) apply a model that focuses on earnings changes as the determinants of

participation. This line of thought was a natural consequence of Ashenfelter's discovery and resulted in analyses using earnings histories to eliminate differences between participants and nonparticipants[46]. Clearly, the fact whether the pre-program earnings dip is transitory or permanent determines what would have happened to participants had they not participated, and the validity of any estimation approach depends on the relationship between earnings in the post-program period and the determinants of program participation (Heckman and Smith 1999).

This rather established observation that it is earnings dynamics that drive program participation has lately been put into serious question by Heckman and Smith (1999), who argue that it is rather labor force dynamics that determine participation in an ALMP program. This point had implicitly been made before by Card and Sullivan (1988), who analyze training effects conditional on pre-program employment histories. Furthermore, Heckman and Smith (1999) argue for a distinction between employment dynamics – indicating whether an individual is employed or not – and labor force dynamics, incorporating also whether a nonemployed person is either unemployed or out-of-the-labor-force. Their conclusion is "that labor force dynamics, rather than earnings or employment dynamics, drive the participation process" (Heckman and Smith 1999). Therefore, we extend the "employment history setting" considered in Card and Sullivan (1988) to a "labor force status history setting" reflecting also movements in and out of inactivity.

We consider the 12-month labor market history of every single treated unit directly preceding the exact point in time – i.e. month – that the individual entered the program. As for the post-treatment outcomes, we condense the monthly information to a sequence of four quarters of a multinomial outcome variable (0,1,2) denoting labor force status (out-of-the-labor-force, employed, unemployed). For each treated unit in succession, the matching algorithm for sample (C) computes labor market histories for all potential comparison units at this point in time and matches those units who – in addition to the correspondence in the other covariates – display identical "pre-treatment" histories. For illustration see Figure 5.1b.

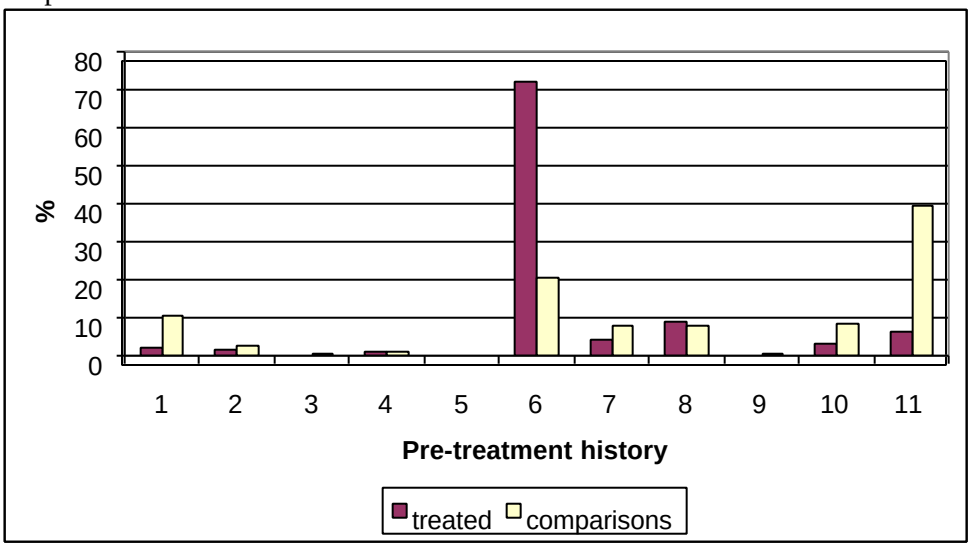
Figures 5.4 and 5.5 draw the distributions of pre-treatment labor market histories for samples (A) and (B) for both Intervention Works (Figure 5.4) and Training (Figure 5.5). Representing a 12-month labor force status sequence with 4 quarterly realizations of a trinomial variable (0,1,2) yields 81 possible sequences ("0000" to "2222"). For the purpose of illustrating the balanced distributions – and only for that purpose – we classify these 81

[46] Heckman and Smith (1999) attribute this emphasis also to the limited data available to "early analysts".

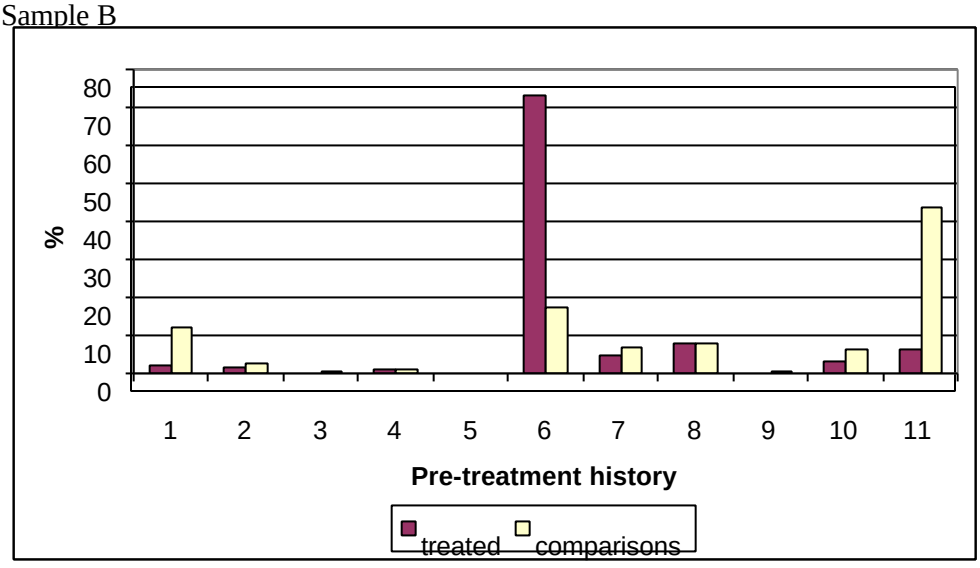
sequences into 11 categories (see Appendix A), so that on the abscissa the low categories contain "inactive" sequences (mostly '0's), the middle categories comprise "unemployed" sequences ('2's), and the high categories represent "employed" sequences ('1's). Categories 1, 6, and 11 exclusively embody the straight sequences (i.e. "0000", "2222", and "1111", respectively).

Figure 5.4 Distribution of pre-treatment labor market history by sample – Intervention Works

Sample A



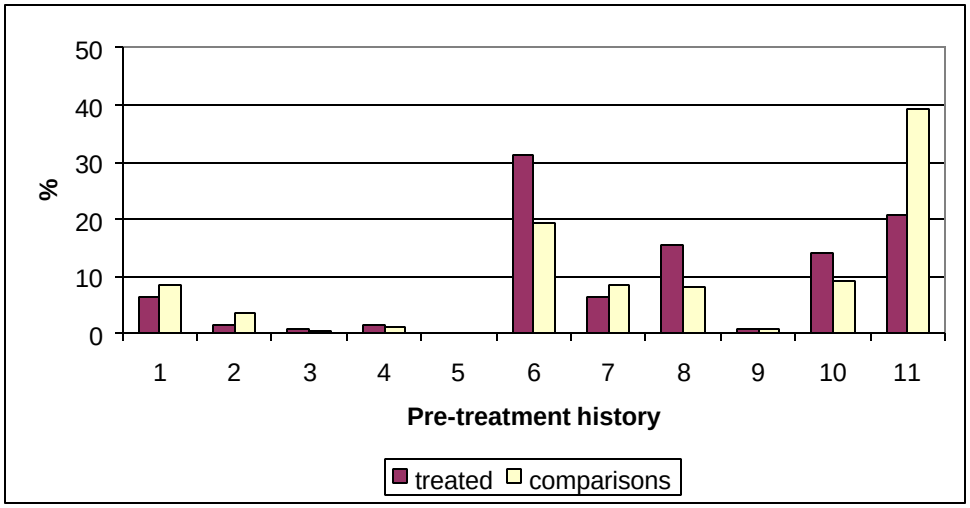
Sample B



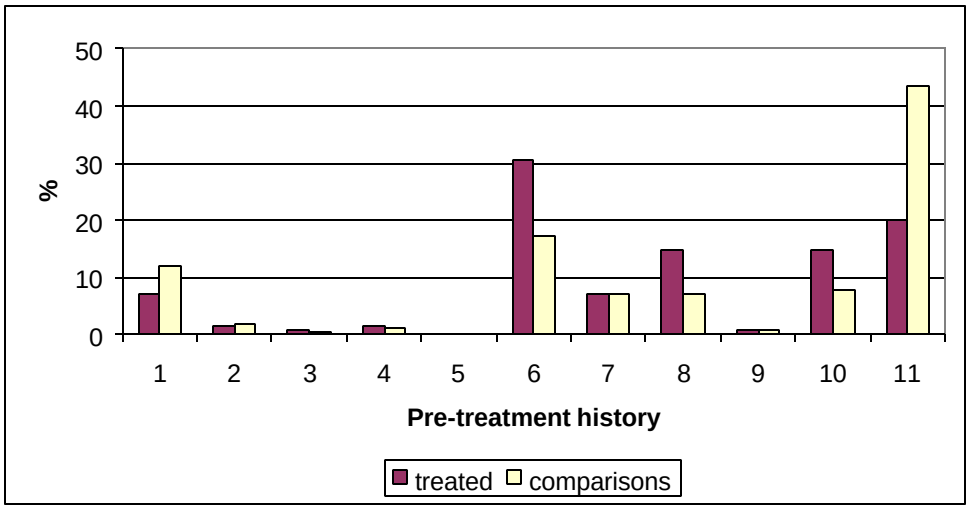
The 3^4 possible labor force status sequences are classified into 11 categories (see text and Appendix A).

Figure 5.5 Distribution of pre-treatment labor market history by sample – Training

Sample A



Sample B



The 3^4 possible labor force status sequences are classified into 11 categories (see text and Appendix A).

Thus, of the three peaks we observe in most of the graphs in figures 5.4 and 5.5, the left peak represents the area of "inactive" histories, because histories with a low order number contain many '0's. Accordingly, the peak in the middle expresses "unemployed" histories, and the peak to the right depicts "employed" histories. In terms of balancing of distributions, the picture is almost the same for figures 5.4 and 5.5. Both samples (A) and (B) display only limited accordance in pre-treatment histories for treated and comparison units. The figures also show that treatment individuals in Training are quite different from those in Intervention

Works. For the Training participants, the fractions of "employed" and "unemployed" histories are quite close to each other, while in the Intervention Works sample we observe a far larger fraction of "unemployed" histories among the treated. Moreover, for both Training and Intervention Works the comparison samples (A) and (B) are too "successful" in that they contain too many "employed" sequences relative to "unemployed" sequences in order to be comparable to the treated units, where "unemployed" sequences dominate.

5.3.5 Propensity Score Balance

The preceding sections were concerned with balance in selected individual characteristics. It is instructive to also provide a *summary measure of balance*, the propensity score. While the estimation of propensity scores is usually a principal step in the construction of matched samples – with the hope that the resulting matched sample displays a balance in all relevant characteristics but no possibility to test this presumption – we can use our samples to directly analyze balance in the propensity score. Correspondingly, we predict post-match propensity scores for samples (A) and (B), based on estimates derived from sample (A). We follow a probit specification with interaction terms between some of the covariates,

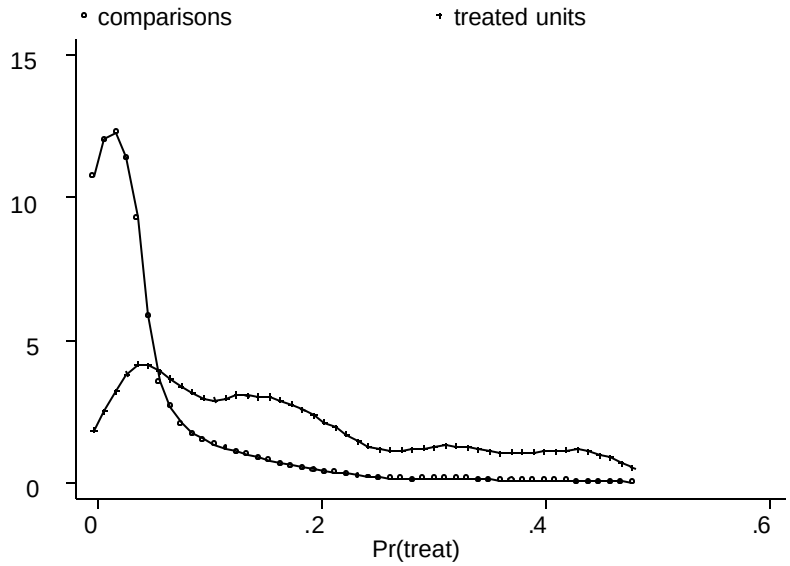
$$(5.3) \quad \mathrm{P}(D=1|X)=\Phi(a_0+a_1X+a_2X\otimes X)$$

where F denotes the cumulative normal density function, X is the vector of covariates, and $X \otimes X$ indicates all relevant interactions across covariates. Regressors comprise indicator variables capturing age, education, gender, and region. Moreover, corresponding to the condensation of pre-treatment labor market histories into 11 distinct "types" in section 5.3.4, there are 10 indicators of pre-treatment history among the regressors. Finally we interacted age, gender and education in a saturated fashion.

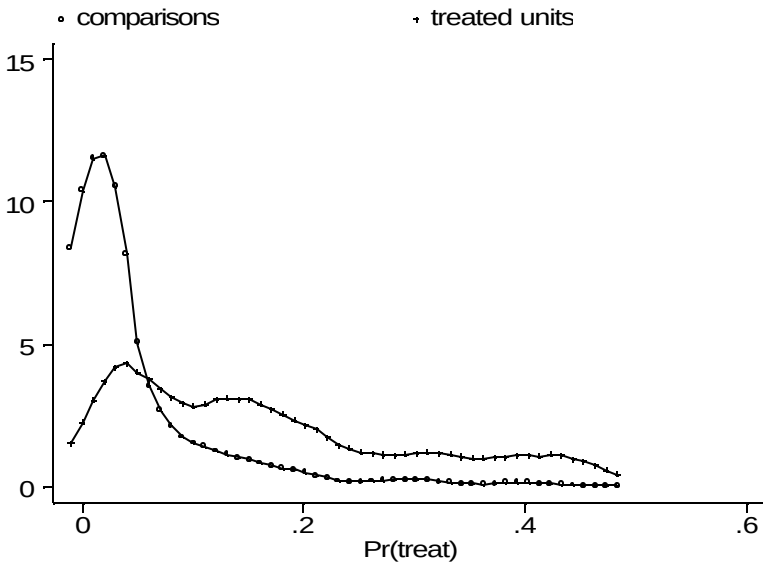
This model is estimated using the treatment units (yielding the value "1") and comparison sample (A) providing the "0" observations. Note that we observe both the individual characteristics and the pre-treatment histories also with comparison sample (A), although this information is utilized only in the construction of comparison samples (B) and (C), respectively. The resulting coefficients are employed to predict propensity scores in samples (A) and (B). Figures 5.6 and 5.7 document the distribution of propensity scores in these comparison samples – relative to the corresponding distribution among treatment units – for the two measures under study.

Figure 5.6 Distribution of estimated propensity score by sample – Intervention Works

Sample A



Sample B



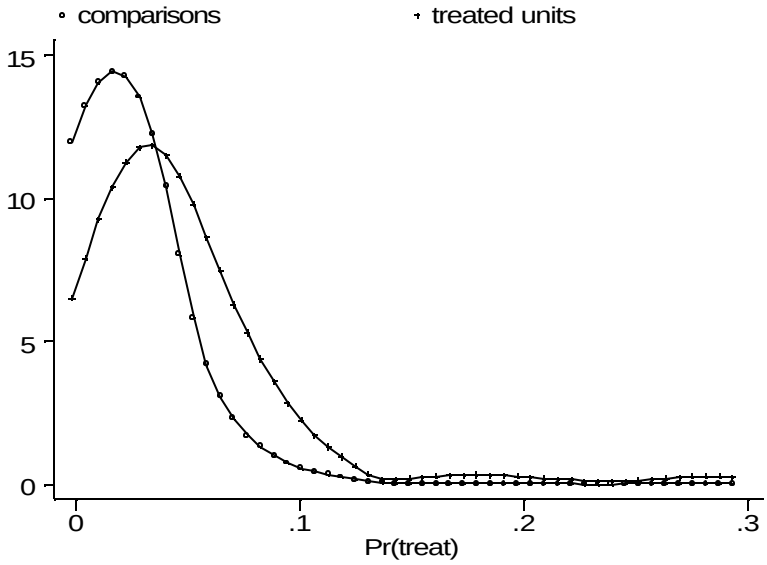
Kernel density estimates of the propensity score for treated and comparison units by STATA using an Epanechnikov kernel and total bandwidth of (.02). Density estimates are not bound, their purpose is for illustration only. Y-axis denotes percentages.

Note that the density for treated units is not scaled relative to the number of observations in the comparison pool, so that the figure depicts the distribution of scores rather than the proportion of treated units to comparison units. In both figures 5.6 and 5.7 the comparison units gather at the low end of the estimated score. Whereas for Intervention Works treated units are distributed rather evenly, with the peak to the low end and then slightly declining

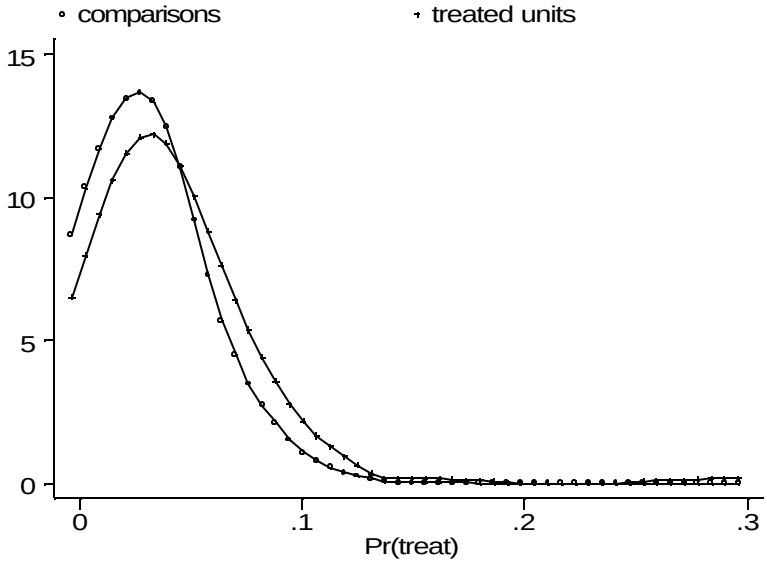
towards the upper tail, the majority of treated units for Training also displays relatively low scores, with an overall distribution quite close to that of comparison units. We find relatively little change in balance from (A) to (B) for both Training and Intervention Works. For Training the distributions are rather balanced – for Intervention Works, however, the substantial imbalance in pre-treatment histories clearly finds expression in the score distributions for (A) and (B) that do not yet control for this imbalance.

Figure 5.7 Distribution of estimated propensity score by sample – Training

Sample A



Sample B



Kernel density estimates of the propensity score for treated and comparison units by STATA using an Epanechnikov kernel and total bandwidth of (.02). Density estimates are not bound, their purpose is for illustration only. Y-axis denotes percentages.

5.4 Empirical Results

5.4.1 Distributions of Outcomes

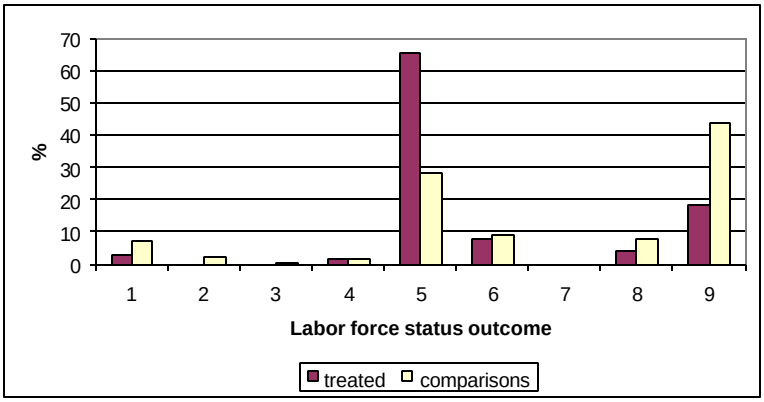
To illustrate the substantial heterogeneity of labor market outcomes following Intervention Works and Training, Figures 5.8 and 5.9 plot distributions for the post-treatment employment success of treated units and comparisons in samples (A) to (C). There are 27 possible labor market status sequences capturing employment performance in the three quarters succeeding treatment (cf. also Figures 5.1a,b). Similar to the presentation of pre-treatment labor market histories we classify these 27 possible sequences of 3 quarterly realizations of a trinomial variable into 9 categories for illustration purposes. This categorization is outlined in Appendix A. Once more, low categories contain "inactive" sequences (category 1="000"), middle categories include "unemployed" sequences (category 5="222"), and high categories comprise "employed" histories (category 9="111"). Accordingly, in the graphs the left peak depicts "inactive" sequences, the middle peak "unemployed" sequences, and the peak to the right "employed" histories.

Looking at the Intervention Works samples in Figure 5.8, we find that in all samples the "unemployed" sequences are clearly predominant for the treated units. At the same time, comparison units display rather successful labor market histories in samples (A) and (B). For our preferred comparison sample (C) this picture changes considerably, and a larger fraction of comparison units also displays "unemployed" histories. However, the comparison group still fares visibly better than the program participants. Attributing the most reliable results to sample (C), we would conclude that during the 9 months directly succeeding participation in Intervention Works the treated units on average were marginally – possibly insignificantly – less successful in finding employment than the comparison units.

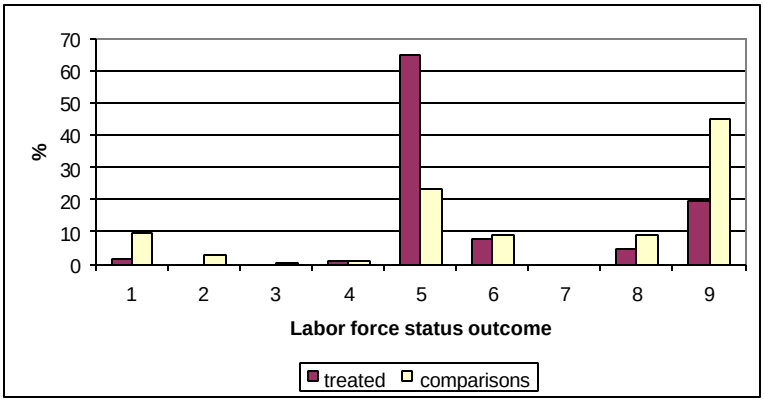
For the Training samples shown in Figure 5.9 we find slightly different results. Similar to pre-treatment sequences of these samples (Figure 5.5), the "employed" and "unemployed" peaks have approximately the same height also for the post-treatment sequence. But while for samples (A) and (B) the "employed" peak is higher for comparison units than for treated units, and the "unemployed" peak is higher for treated units than for comparisons, this relation switches for sample (C). In (C) treated units display on average a slightly more successful post-treatment labor market sequence than corresponding comparisons. We would thus attribute a slightly – possibly insignificant – positive treatment effect to Training.

Figure 5.8 Distribution of post-treatment labor market sequence by sample – Intervention Works

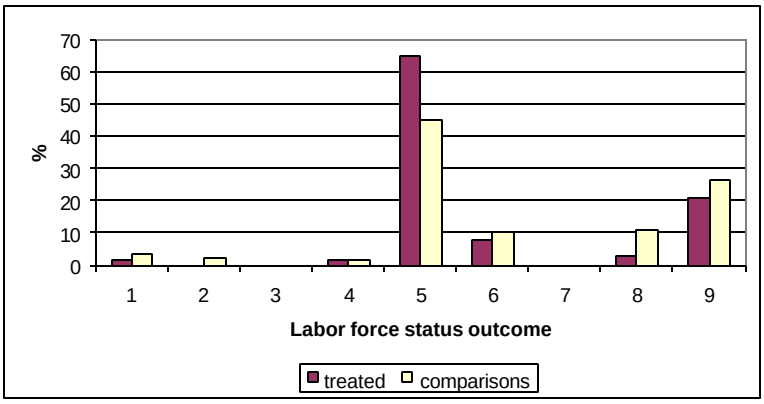
Sample A



Sample B



Sample C

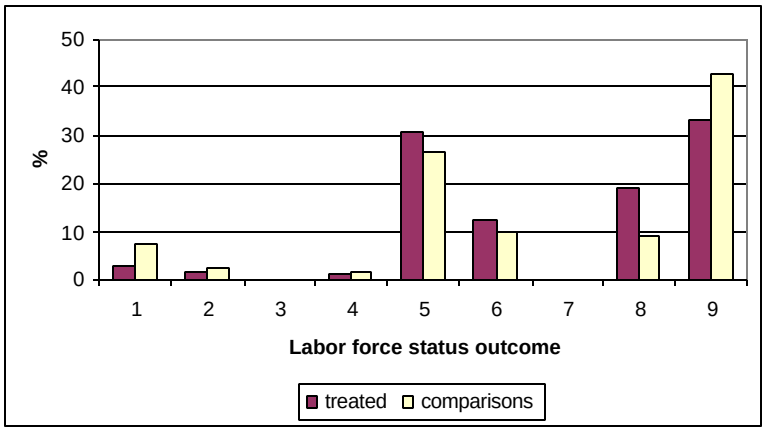


The 3^3 possible labor force status sequences are classified into 9 categories (see text and Appendix A).

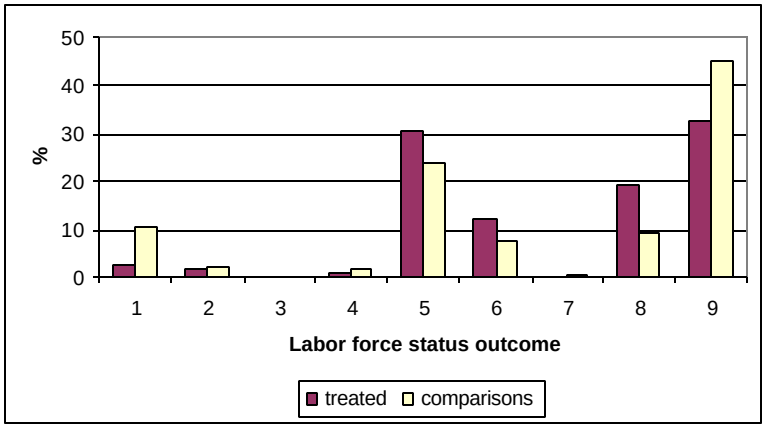
Taken together, Figures 5.8 and 5.9 display three important patterns. First, moving from (A) to (C) we do not observe much variation in the distributions for treated units. Thus, the fact

Figure 5.9 Distribution of post-treatment labor market sequence by sample – Training

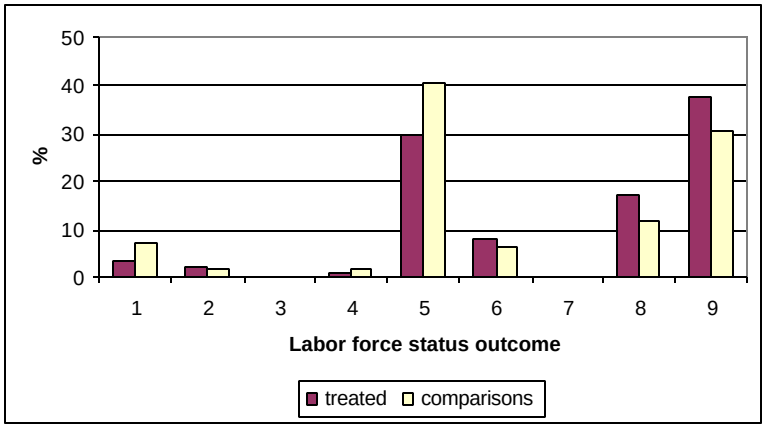
Sample A



Sample B



Sample C



The 3^3 possible labor force status sequences are classified into 9 categories (see text and Appendix A).

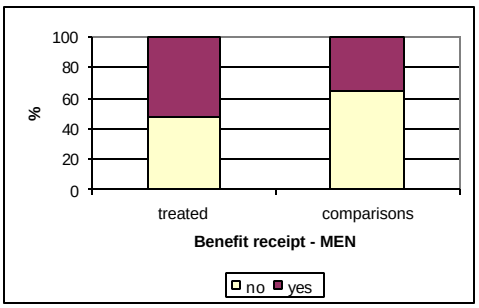
that we lose some treated units while increasing matching requirements does not seem to play an important role. Second, without conditioning on pre-treatment labor market histories the

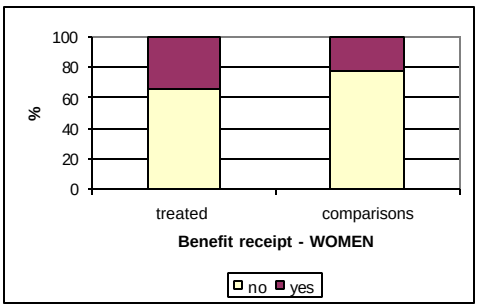
comparison samples apparently contain too many "successful" individuals – a pattern which we already observed for pre-treatment labor force status sequences in Figures 5.4 and 5.5. For samples (A) and (B) this would result in a far too negative estimate of treatment effects. Third, across comparison units and treated units we observe clearly more "successful" outcomes for Training than for Intervention Works. This is not surprising, as we noticed a similar relation for pre-treatment labor market history distributions (Figures 5.4 and 5.5).

In Figures 5.10 and 5.11 we address the idea that participation in Intervention Works might primarily be a vehicle to renew eligibility for unemployment benefits. Recall that according to Polish ALMP regulations Intervention Works renews benefit receipt eligibility, whereas Training does not. Figures 5.10 and 5.11 perform a simple before-after comparison of the variable "unemployment benefit receipt" for both ALMP measures, and for men and women separately. The top panel of each figure indicates benefit receipt in at least two of the three months directly *preceding* treatment. The middle panel shows benefit receipt in at least two of the three months directly *succeeding* treatment. The bottom panel plots benefit receipt in at least two months of each of the three quarters succeeding treatment, i.e. at least 6 out of 9 months. We focus on sample (C) for both measures.

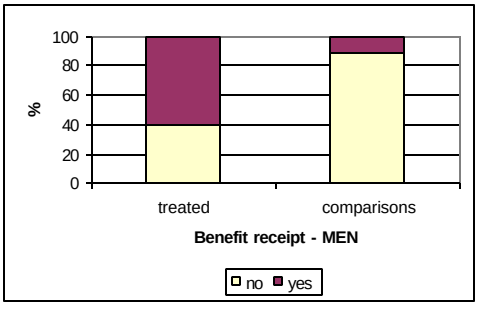
Figure 5.10 Distribution of benefit receipt by sex for sample C – Intervention Works

During 3 months BEFORE treatment:





During 3 months AFTER treatment:



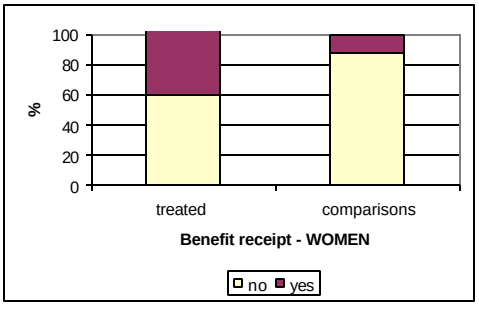


Figure 5.10 [ctd]

During 9 months AFTER treatment:

The upper panel indicates benefit receipt (="yes") during at least two of the last three months preceding treatment. The middle panel indicates benefit receipt during at least two of the first three months succeeding treatment. The bottom panel indicates benefit receipt during at least two of the three months in each of the three quarters succeeding treatment.

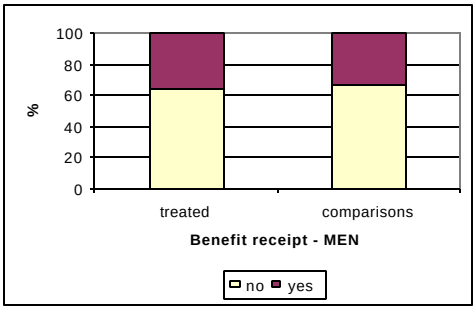
Figure 5.10 shows for Intervention Works that a substantial fraction of both treated and comparison units received pre-treatment benefits, although benefits do seem to play a more important role for treated units. This pattern is more pronounced for men. In the middle and bottom panel this situation aggravates substantially. While both short-term and medium-term benefit receipt played a minor role for comparison units, we observe that approximately 60% of the treated males received unemployment benefits in the quarter directly following treatment, and that more than half of the treated males received benefits during the whole 9-month post-treatment period. For females, this pattern is not quite as severe, but still post-treatment benefit receipt plays a major role for Intervention Works participants.

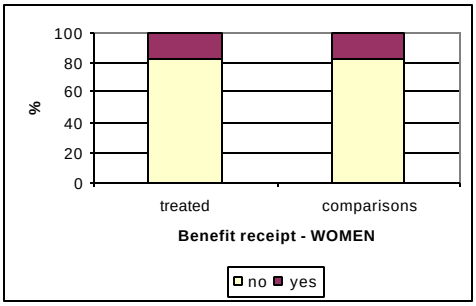
The situation for the Training sample is quite different. As Figure 5.11 shows, unemployment benefits do play some role for both treated and comparison units during the one quarter directly before and after participation, at least for the males. However, in the medium run this effect diminishes, and only very few observations in the treatment and comparison group display benefit receipt for the whole 9-month period following treatment. This pattern is even less pronounced for women than for men.

As a result, figures 5.8 through 5.11 indicate that individuals involved in Training measures seem to be generally more successful before and after the treatment than those participating in Intervention Works. However, these patterns are difficult to reconcile on the basis of a more favorable impact of Training. Rather, this simple evidence suggests that substantial benefit churning seems to take place in the case of Intervention Works, but not in the case of Training.

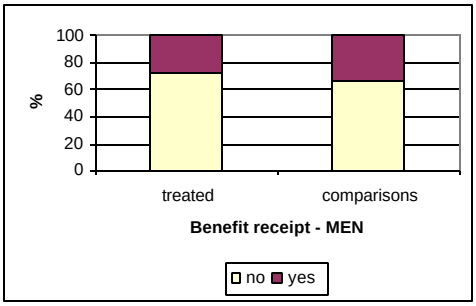
Figure 5.11 Distribution of benefit receipt by sex for sample C – Training

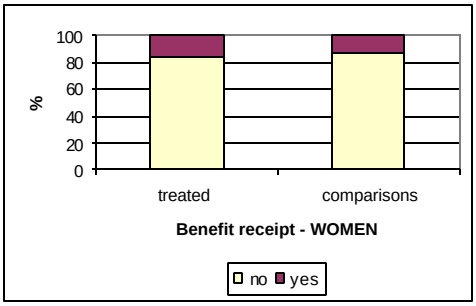
During 3 months BEFORE treatment:



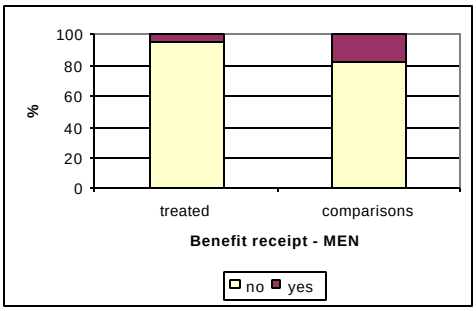


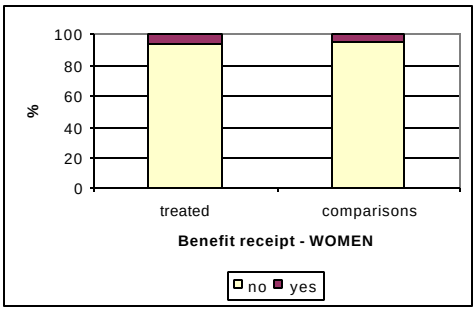
During 3 months AFTER treatment:





During 9 months AFTER treatment:





The upper panel indicates benefit receipt (="yes") during at least two of the last three months preceding treatment. The middle panel indicates benefit receipt during at least two of the first three months succeeding treatment. The bottom panel indicates benefit receipt during at least two of the three months in each of the three quarters succeeding treatment.

5.4.2 Treatment Effect Estimation

Our aim is to identify treatment effects of two different measures of Polish active labor market policy, Intervention Works and Training, which we consider separately in the empirical analysis. For purposes of the formal exposition of our estimation approach we consider a single generic intervention. Furthermore, we explicitly require that treated units be matched with comparison units from the identical set of observed pre-treatment and post-

treatment months. Any reference to the time period is therefore omitted from the formal exposition as well.

In addition to the terminology introduced in section 5.2, let N_1 denote the number of treated units, with indices $i \in I_1$, and N_0 the number of potential comparison units, with indices $i \in I_0$. Potential labor market outcomes in post-treatment quarter q $(q = 1, 2, 3)$ are denoted by Y_{qi}^1, if individual i received treatment, and by Y_{qi}^0, if individual i did not receive treatment. Outcomes are defined as multinomials with three possible realizations ('0'=out-of-the-labor-force, '1'=employed, '2'=unemployed), extending the formulations of Card and Sullivan (1988) from a binomial to a trinomial setting.

We can only observe one of the two potential outcomes Y_{qi}^1 and Y_{qi}^0 for a given individual. This actual outcome is denoted by Y_{qi}. The objective is then to formally construct an estimator of the mean of the unobservable counterfactual outcome $E(Y_{qi}^0 | D_i=1)$. Following the quarterly sequence of labor market outcomes might be too detailed, though, for a direct economic interpretation of results. Thus, to condense the available information further, the post-intervention labor market success of each individual i is summarized by the individual's average employment rate over the three quarters following the intervention. Using indicator function $\mathbf{1}(.)$, these employment rate outcomes are $\frac{1}{3}\sum_q \mathbf{1}(Y_{qi}=1)$.[47] *Observed* outcomes for individual i can then be written as

$$(5.4) \quad \frac{1}{3}\sum_q \mathbf{1}(Y_{qi}=1) \;=\; \frac{1}{3}\left(D_i \sum_q \mathbf{1}(Y_{qi}^1=1) + (1-D_i)\sum_q \mathbf{1}(Y_{qi}^0=1)\right) \quad ,$$

and the impact of the intervention on the average labor market status of individual i can be expressed as

$$(5.5) \quad \Delta_i = \frac{1}{3}\left(\sum_q \mathbf{1}(Y_{qi}^1=1) - \sum_q \mathbf{1}(Y_{qi}^0=1)\right)$$

[47] Kluve et al. (1999) extend this setting to considering both employment and unemployment rates, so that corresponding outcomes would be $\frac{1}{3}\sum_q \mathbf{1}(Y_{qi}=w)$, where $w \in \{1,2\}$. Comparing employment and unemployment rate treatment effects shows for instance that exits to inactivity play a much larger role for women than for men. Moreover, Kluve et al. (1999) also consider the medium run, i.e. 6 post-treatment quarters, while we focus on the short-term case here. The extension to any number of post-treatment periods is straightforward.

for average employment rates. The parameters of interest in our evaluation analysis are weighted population averages over these individual treatment effects, the mean effect of treatment on the treated for types of individuals characterized simultaneously by specific sets of characteristics *X*; and labor market histories before treatment h_i,

$$(5.6) \quad E(\Delta_i \mid X_i, h_i, D_i = 1) \quad = \quad E\left(\frac{1}{3}\left(\sum_q \mathbf{1}(Y_{qi}^1 = 1) - \sum_q \mathbf{1}(Y_{qi}^0 = 1)\right) \mid X_i, h_i, D_i = 1\right) \quad .$$

The less inclusive the chosen set of characteristics conditioned upon – i.e. the more specific characteristics are included in X – the larger is the population of treated individuals over which the conditional mean is taken. As laid out above, previous labor market histories h_i are captured by the sequence of labor market states in the four quarters preceding the intervention.

Our approach to combine the population averages of the treatment effects for individuals in a given history-specific "cell" – characterized by demographic and other characteristics, in particular labor market history – gives us considerable flexibility in addressing the economic interpretation of results. The standard approach to evaluation would be to consider the distinction of type-history cells primarily as a device to achieve comparability of treatment and comparison units (see below). The ultimate interest there typically lies in the average treatment effects over the joint support of *X* and *h* given *D=1*,

$$(5.7) \quad M = \sum_s w_s E(\Delta \mid s, D = 1),$$

with *s* indicating any possible combination of *X* and *h*, and w_s representing the corresponding relative frequency in the treatment sample. By contrast to this standard approach, in what follows we will consider appropriate subsets of this joint support.

How does our particular observational approach – matching – facilitate the estimation of these parameters of interest? In randomized experiments the counterfactual expected values under no intervention can simply be estimated for intervention recipients by the mean values of the outcome for randomized-out would-be recipients. As we have shown in section 5.2, matching methods can recover the desired counterfactual for a nonexperimental comparison group: Within each matched set of individuals, one can estimate the treatment impact on individual *i* by the difference over sample means, and one can construct an estimate of the

overall impact by forming a weighted average over these individual estimates.

Matching estimators thereby approximate the virtues of randomization mainly by balancing the distribution of observed attributes across treatment and comparison groups, both by ensuring a common region of support for individuals in the intervention sample and their matched comparisons and by re-weighting the distribution over the common region of support. The central identification assumption is that of mean independence of the labor market status Y_{qi}^0 and of the treatment indicator D_i, given individual observable characteristics. In our specific application these conditioning characteristics are the demographic and regional variables X_i and the pre-treatment history h_i, i.e. from equation (5.2) in our case,

$$(5.8) \quad E(\mathbf{1}(Y_{qi}^0=1) \mid X_i, h_i, D_i=1) \quad = \quad E(\mathbf{1}(Y_{qi}^0=1) \mid X_i, h_i, D_i=0) \quad .$$

Thus, by conditioning on previous labor market history we exploit the longitudinal nature of our data.

In a standard difference-in-differences approach pre-treatment and post-treatment outcomes are typically treated symmetrically; the identifying assumption is that the change in outcomes that treated individuals would have experienced had they not received treatment, would have been the same change – on average – that untreated individuals experience during the same period. This assumption accounts for the phenomenon that treatment units typically experience lower pre-treatment outcomes, even though they might be otherwise identical to comparison units. It does not lend itself naturally to the analysis of categorical outcome variables, though. In this context, a natural generalization of the difference-in-differences idea is to condition on the specific realization of the outcome variable in the pre-treatment period, as we do here. This is possible, since due to the categorical nature of the outcome the conditioning remains tractable. Card and Sullivan (1988) and Heckman et al. (1997) advocate such difference-in-differences approaches (cf. also Schmidt 1999).

Our matching estimator is one of oversampling exact covariate matching within calipers, allowing for matching-with-replacement. Our particular attention to pre-treatment labor market histories implements this idea of a generalized difference-in-differences juxtaposition between treated units and comparison units. Due to the relevance of the previous history for subsequent labor market success – state dependence is one of the issues most

discussed in the labor literature – we also emphasize this variable in the construction of the estimates. Specifically, for any treatment history h for which at least one match could be found, we estimate the impact of the intervention by

$$(5.9) \quad \hat{M}_h = \frac{1}{N_{1h}} \sum_{i \in I_{1h}} \left[\frac{1}{3} \sum_q \mathbf{1}(Y_{qi}^1 = 1) - \sum_{j \in I_{0h} \mid X_j \in C(X_i)} \frac{1}{n_{i0}} \left(\frac{1}{3} \sum_q \mathbf{1}(Y_{qj}^0 = 1) \right) \right] ,$$

where N_{1h} is the number of individuals with history h who receive the intervention ($N_1 = \sum_h N_{1h}$), I_{1h} is the set of indices for these individuals, $C(X_i)$ defines the caliper for individual i's characteristics X_i, and n_{i0} is the number of comparisons with history h who are falling within this caliper, with the set of indices for comparison-individuals with history h being I_{0h}. The standard error of the estimated treatment effect is then constructed as a function of the underlying multinomial probabilities. This procedure is outlined in Appendix B.

The overall effect of the intervention is estimated in a last step by calculating a weighted average over the history-specific intervention effects,

$$(5.10) \quad \hat{M} = \sum_h \left[\frac{N_{1h}}{\sum_h N_{1h}} \hat{M}_h \right] ,$$

using the treated units' sample fractions as weights. The variance is derived as the corresponding weighted average of the history-specific variances.

5.4.3 Treatment Effect Results

In this section we analyze the treatment effect estimates which we obtain by applying the estimator developed in the previous section. Table 5.2 presents average treatment effects on the post-intervention employment rate for Intervention Works sample (C). The structure of the table shows how the total treatment effect (-.126) is being calculated by computing history-specific effects first. As explained above, for each treated unit, if he or she has more than one matched comparison unit, the comparison units' employment rates are averaged and handled as if they were the employment rate of only a single unit. The total effect is the weighted average of the history-specific effects using the treated units' sample fractions as weights.

Table 5.2 Average post-treatment employment rate treatment effect by pre-treatment labor market history for comparison sample C – Intervention Works

	treated units			comparison units				
job history	**N**	**rate[a]**	**std.err.**	**N**	**rate**	**std.err.**	**effect[b]**	**std.err.**
0000	5	0.333	0.189	6	0.400	0.219	-0.067	0.289
0002	1	0.000	0.000	1	0.667	0.471	-0.667	0.471
1111	16	0.813	0.098	19	0.729	0.111	0.084	0.148
1112	5	0.467	0.202	6	0.167	0.167	0.300	0.262
1122	6	0.222	0.150	6	0.333	0.192	-0.111	0.244
1222	4	0.500	0.250	4	0.833	0.186	-0.333	0.312
2000	1	1.000	0.000	1	0.000	0.000	1.000	0.000
2111	1	1.000	0.000	1	1.000	0.000	0.000	0.000
2211	4	0.167	0.144	4	0.667	0.236	-0.500	0.276
2221	1	0.000	0.000	1	0.333	0.471	-0.333	0.471
2222	168	0.183	0.027	191	0.333	0.036	-0.150	0.045
total[c]	212			240			-0.126	0.040

[a] Average employment rate in the three post-treatment quarters.
[b] Difference between rates of treated units and matched comparison units.
[c] Total effect is the weighted average of the effects for the individual histories using the treated units' sample fractions as weights.

Besides treatment effect calculation Table 5.2 shows which labor market state sequences occurred in the data, thus picking up the theme of figure 5.4. We observe the same predominance of "unemployed" histories which we already noticed in the figure. The total treatment effect casts a rather negative picture on the Intervention Works program, suggesting that participation tends to lower post-treatment employment prospects. In principle, this finding would conclude our analysis: we have described the nonexperimental context of the study, we have shown by what means we overcome the problem of constructing the desired counterfactual, and we have applied the appropriate estimation methods in order to obtain credible treatment effect estimates. As far as the data permit, the causal effect of Intervention Works participation is identified. Or is it?

In fact, looking at Table 5.3 we find that there may be more to it. First, we report treatment effect estimates for comparison samples (A) and (B) obtained by taking sample averages over the average employment rate in the three post-treatment quarters. The estimates are far more negative than the one obtained using sample (C), clearly reflecting the over-representation of "successful" labor force status sequences in the respective comparison samples (cf. Figures 5.4 and 5.8). Furthermore, in accordance with our discussion of expression (5.7), in Table 5.3 we subdivide the matched Intervention Works comparison

sample (C) with respect to various covariates, and we compare the conditional treatment effect for the subsample to the full sample estimate. Even a simple subdivision by gender reveals an interesting finding: The significantly negative full sample effect consists of a – more or less – zero treatment effect for women and a considerably larger negative effect for men. On the other hand, a subdivision by date of program entry that parts the observation period into two halves does not reveal any apparent influence of changes in the macroeconomic environment.

Table 5.3 Average post-treatment employment rate treatment effect for subsamples – Intervention Works

Subdivision by	Categories	treated units	matched comparison units	effect[a]	std.err.
Sample A	-	275	6757	-.285	.026
Sample B	-	244	1354	-.291	.031
Sample C:	-	212	240	-.126	.040
Gender	Men	123	133	-.236	.051
	Women	89	107	.026	.062
Date of Program Entry	≤ June 1994	116	137	-.135	.052
	≥ July 1994	96	103	-.115	.056
Program Entry & Gender	≤ June 1994 Men	66	73	-.295	.069
	≤ June 1994 Women	50	64	.076	.079
	≥ July 1994 Men	57	60	-.167	.073
	≥ July 1994 Women	39	43	-.038	.089
Labor market history	1111	16	19	.084	.148
	2222	168	191	-.150	.045
Labor market history & Gender	1111 Men	10	12	.117	.161
	1111 Women	6	7	.028	.274
	2222 Men	100	108	-.258	.057
	2222 Women	68	83	.010	.072

[a] Average employment rate in the three post-treatment quarters.

The next step is to further refine cells and classify the sample by both gender and date of program entry. These subsamples indicate that post-treatment employment prospects for male Intervention Works participants were quite unfavorable in the second period after July 1994, but particularly severe during the first period until June 1994. For women the time period distinction leads to the opposite result, but both the positive effect of the first half and the negative effect of the second half are small and insignificant. This also points to the fact that, as we increase the number of subdivisions, subsample sizes decrease and standard errors increase.

Classification by labor market history allows us to look at the two major labor force status sequences that drive the peaks from Figures 5.4 and 5.5. For "employed" (1111) histories subsample sizes are rather small and the effects not well defined. For the subsample of "unemployed" (2222) histories, which entails almost 80% of total treated and comparison units, we find a significantly negative treatment effect close to the full sample effect. This is certainly no surprise, as the estimate of the full sample effect is dominated by the "2222" subsample effect. If we further classify by labor market history and gender, treatment effects for the "1111" subsample remain insignificant for both men and women, while the "2222" subsample displays the same substantial male/female difference in the treatment effect that we have seen for the full sample.

Table 5.4 reports the same comparison between samples and various subdivisions for Training. Both treatment effect estimates from comparison samples (A) and (B) suggest an insignificantly negative effect of Training participation, while the estimate obtained from sample (C) indicates that Training raises the individual employment probability by 13.8%. This sudden switch of signs is in line with our observations drawn from Figure 5.9. Further looking at comparison sample (C), we conclude that in the case of Training a classification by gender does not seem to add any insights to the interpretation: Treatment effects for men and women are almost identical. While a categorization by gender and date of program entry shows contradictory results (upward for men, downward for women from one period to the other), the number of observations per subsample is in fact too small to draw any firm conclusions. Looking at a classification by labor market history, once more we find the "peaks" from Figure 5.5, indicating here that the share of "1111" sequences is almost as large as the that of "2222" sequences. Again, subsample sizes are quite small for interpretation purposes.

Table 5.4 Average post-treatment employment rate treatment effect for subsamples – Training

Subdivision by	Categories	treated units	matched comparison units	effect[a]	std.err.
Sample A	-	121	6751	-.027	.046
Sample B	-	114	983	-.048	.049
Sample C:	-	87	111	.138	.059
Gender	Men	36	39	.148	.092
	Women	51	72	.130	.070
Date of Program Entry	≤ June 1994	38	52	.212	.088
	≥ July 1994	39	59	.080	.064
Program Entry & Gender	≤ June 1994 Men	15	17	.056	.156
	≤ June 1994 Women	23	35	.313	.104
	≥ July 1994 Men	21	22	.214	.094
	≥ July 1994 Women	28	37	-.020	.086
Labor market history	1111	24	34	.071	.115
	2222	32	43	-.077	.103
Labor market history & Gender	1111 Men	11	12	.045	.194
	1111 Women	13	22	.092	.129
	2222 Men	11	12	-.046	.192
	2222 Women	21	31	.093	.116

[a] Average employment rate in the three post-treatment quarters.

From these calculations results the observation that an appropriate subdivision of a matched sample can substantially contribute to disentangling and identifying heterogeneous treatment effects. In particular, the example of a simple classification by gender for the Intervention Works sample is striking: The overall negative effect is almost exclusively due to the dismal post-treatment labor market performance of male participants. Thus, while the recognition of the principal idea that treatment effects are heterogeneous across the population has led to the development of sophisticated econometric methods for constructing convincing

counterfactuals, it is easy to forget the necessity to stratify the sample appropriately in order to interpret the results in economically meaningful terms. Thus, controlling for observable characteristics in establishing the statistical model does not seem to be sufficient – it appears to be good advice to re-consider the same observable characteristics (which we already controlled for) when analyzing the empirical results. This recommendation seems imperative if one wants to assess for example targeting issues: bad targeting of programs is often claimed to be one reason for disappointing treatment effects. In our particular application, Intervention Works has been uncovered as an extremely disappointing measure in the case of men – a result that would have remained hidden, had we not pursued an appropriate sample split.

Of course, these negative treatment effects could be explained by other factors than poor targeting. Stigma is often given as a reason why participants of an employment program like Intervention Works perform worse in the labor market than non-participants.[48] Prospective employers identify participants as "low productivity workers" and are not willing to accept them into regular jobs. Another explanation, which might have particular merit in the Polish case, is benefit churning. Workers with long unemployment spells who have difficulty finding regular employment are identified by labor bureau officials and might only be chosen for participation in an employment scheme so that they re-qualify for another round of benefit payment.

While the presented evidence cannot pinpoint precisely the cause underlying the poor labor market performance of males participating in Intervention Works, stigmatization seems to be the least likely cause. For if participation in the scheme was a bad signal to prospective employers, it is not clear why this would not be the case for female participants. It may be that those males – males are for the most part heads of households – are targeted by labor bureau officials who have especially poor prospects for regular employment. Once the publicly subsidized job comes to an end, so officials might reason, they at least qualify for another round of unemployment benefits, if they cannot find regular employment elsewhere or if their subsidized job is not transformed into a regular job. It is probably not a mere coincidence that the large majority of Intervention Works jobs lasts six months, the length of time one needs to work within the year preceding benefit receipt in order to qualify for unemployment benefits.

However, more work is needed to determine firmly the factor(s) that drive the poor labor market performance of males after their participation in Intervention Works comes to an

[48] A large part of the intervention works jobs are actually in the public domain, i.e. we can also think of this scheme as a public employment program.

end. For example, the fate of female participants after the end of the subsidized job needs to be more thoroughly analyzed. Specifically, one needs to ask whether female participants are more likely to be kept on by employers or whether they find regular jobs elsewhere more readily than men because their characteristics are better than those of men, i.e. because the targeting criteria are different for men and women. It could also be that women who participate in Intervention Works are selected into jobs that are more conducive to prolonged job matches because demand in these jobs is strong (e.g. nursing jobs).

Table 5.5 Counterfactual treatment effects for samples C

Treatment	Weights	Effect[a]	Std.Err.	Interpretation
Intervention Works	Intervention Works	-.126	.040	Factual IW treatment effect
Intervention Works	Training	-.048	.064	Counterfactual IW treatment effect
Training	Training	.138	.059	Factual Training treatment effect
Training	Intervention Works	.089	.083	Counterfactual Training treatment effect
Intervention Works – Training	Intervention Works	-.218	.093	Differential treatment effect Intervention Works vs. Training
Training – Intervention Works	Training	.185	.087	Differential treatment effect Training vs. Intervention Works

[a] Average employment rate in the three post-treatment quarters.

In addition to displaying the treatment effects by sample and subdivision, Table 5.5 presents treatment effect estimates for comparison samples (C) obtained from a "counterfactual experiment". The first line reports the factual Intervention Works treatment effect estimate computed as shown in Table 5.2. This estimate tries to answer the question: "How much did Intervention Works participants benefit from participating in Intervention Works?" The second line reports a "counterfactual" Intervention Works treatment effect for Training

participants, i.e. it tries to answer the question: "How much would Training participants have benefited, if they had participated in Intervention Works?" The estimate is obtained by history-wise reweighting the Intervention Works sample using the fraction of the treated units in the Training sample as weights. Looking at Table 5.2 this is the same as if for each history the second column contained the corresponding number of observations from the Training sample. Apparently, this reweighting by labor market history implicitly assumes that there are no relevant changes in other elements of X.

The estimate in the second line of Table 5.5 shows that, while the Intervention Works effect on Training participants still displays a negative sign, the effect is insignificant, so that Training participants participating in Intervention Works would have done better than Intervention Works participants themselves. Looking at the effects of Training on Training participants and Intervention Works participants, respectively, we find the counterpart to this result: Intervention Works participants participating in Training instead would have not gained as much from the treatment as Training participants themselves. Thus, persons with better observable and unobservable characteristics seem to have been targeted for the Training program.

The last two lines in Table 5.5 report differential treatment effects of Intervention Works vs. Training. The estimates represent the difference between the difference of treated and comparison units in Intervention Works (second to last column, Table 5.2) and the difference of treated and comparison units in Training. Once more, differences are taken history-wise and weighted using either Intervention Works participants or Training participants sample weights. Both estimates clearly show that Training is the superior ALMP to Intervention Works.

The methods used in this chapter allow us to evaluate ALMP at the individual level. It thus tells us that those persons participating in Polish Training programs have better employment prospects than they would have had had they not participated and also that they have better employment prospects than those who take part in Intervention Works. The methodology does not address the issue whether Training improves the overall performance of the labor market, i.e., for example, whether it lowers the aggregate unemployment rate. Even if Training is beneficial at the individual level, substitution effects - Training participants just "jump the queue" of those in line for regular jobs - could neutralize its impact at the aggregate level. On the other hand, the finding that a program is not even effective at the individual level, like the Polish Intervention Works scheme, helps us to focus attention on

targeting issues and/or wrong incentive structures that distort the behavior of labor bureau officials and of the unemployed.

5.5 Conclusion

In this chapter we have analyzed treatment effects of two Polish measures of active labor market policy: Training and Intervention Works. The analysis was based on matched samples to overcome the inherent evaluation problem of constructing a credible counterfactual in a nonexperimental setting. We have seen how matching methods can solve this problem by balancing distributions of relevant covariates. Matching methods can be based on exact-covariate-matching, propensity score matching, or a combination of both (partial score). We have argued that on both theoretical and above all empirical grounds the decision for one approach or the other depends heavily on the data.

We have illustrated our own approach to the data by the construction of three different comparison samples using exact-matching-within-calipers, imposing increasingly stricter preconditions. Figures 5.1 to 5.5 have depicted how strong requirements, i.e. a more detailed match on observable characteristics substantially improve the balancing of covariates, and thus the quality of the match. As long as sample sizes do not decrease considerably, such a procedure appears promising. We have illustrated the balancing property of our exact matching approach using the estimated propensity score as a summary measure of balance.

The estimation of the treatment effect is based on a history-specific generalized difference-in-differences estimator. Our estimates suggest that, while Training seems to clearly enhance individual employment prospects, Intervention Works participants fare substantially worse than their comparisons. This is in line with previous findings (cf. Kluve, Lehmann and Schmidt 1999, Puhani 1998). However, we do point to the fact that appropriate subdivision of the matched sample can add considerable insight to the interpretation of results. In our study, for instance, we find that the overall negative treatment effect of Intervention Works is almost exclusively due to the dismal employment performance of male participants, while women do neither gain nor lose anything by participating. From an empirical point of view, we thus doubt that controlling for covariates in constructing the counterfactual is sufficient to account for the heterogeneity of treatment effects – appropriate subdivision of the matched sample may often add clarity to the economic interpretation.

5.6 Appendices

A. Categorizing labor market status sequences

Pre-treatment

Category	1	2	3	4	5	6	7	8	9	10	11
Histories	0000	0001	0012	0022	2201	2222	2220	2211	1102	1110	1111
		0010	0102	0202	2021		2202	2121	1012	1101	
		0100	1002	2002	0221		2022	1221	0112	1011	
		1000	0120	0220	2210		0222	2112	1120	0111	
		0002	1020	2020	2012		2221	1212	1021	1112	
		0020	0021	2200	0212		2212	1122	0121	1121	
		0200	1200		2120		2122		1210	1211	
		2000	0201		2102		1222		1201	2111	
			0210		0122				0211		
			2100		1220				2110		
			2010		1202				2101		
			2001		1022				2011		
			0110						0011		
			1010						0101		
			1100						1001		

Post-treatment

Category	1	2	3	4	5	6	7	8	9
Histories	000	001	210	220	222	221	012	110	111
		010	120	202		212	021	101	
		100	102	022		122	201	011	
		002						112	
		020						121	
		200						211	

B. Calculation of treatment effects and variances

The history-specific treatment effect estimator (5.9) is based on the differences in average employment rate outcomes between treatment and comparison units. One notable element of this estimator is that multiple comparison units matched to a single treated unit (due to the oversampling algorithm) are handled as if they were one single comparison unit. The variance for (5.9) is then composed of the sum of independent single variances of each of the employment rate averages entering (5.9) for "individual" treated and comparison units. This appendix illustrates the generic calculation of this individual variance, and how this yields variances for (5.9) and (5.10).

Within each stratum – defined by pre-treatment labor market history – employment success in the three post-treatment quarters is summarized by the average employment rate $\frac{\sum 1}{3}$. For the unrestricted multinomial model each of the $3^3=27$ possible outcomes is associated with a separate probability. For instance, conditional on the *k-th* history the probability to be employed in all subsequent quarters is $p(111|h_k)$, the probability to be employed in the first and unemployed in the following two quarters is $p(122|h_k)$, the probability to be unemployed in the first two and out-of-the-labor-force in the third quarter is $p(220|h_k)$ etc. Let us order the 27 probabilities in the following way

$\frac{\sum 1}{3}=0$	$\frac{\sum 1}{3}=\frac{1}{3}$	$\frac{\sum 1}{3}=\frac{2}{3}$	$\frac{\sum 1}{3}=1$
$p(000\|h_k)=p_1$	$p(001\|h_k)=p_9$	$p(011\|h_k)=p_{21}$	$p(111\|h_k)=p_{27}$
$p(002\|h_k)=p_2$	$p(021\|h_k)=p_{10}$	$p(211\|h_k)=p_{22}$	
$p(020\|h_k)=p_3$	$p(201\|h_k)=p_{11}$	$p(101\|h_k)=p_{23}$	
$p(200\|h_k)=p_4$	$p(221\|h_k)=p_{12}$	$p(121\|h_k)=p_{24}$	
$p(022\|h_k)=p_5$	$p(010\|h_k)=p_{13}$	$p(110\|h_k)=p_{25}$	
$p(202\|h_k)=p_6$	$p(012\|h_k)=p_{14}$	$p(112\|h_k)=p_{26}$	
$p(220\|h_k)=p_7$	$p(210\|h_k)=p_{15}$		
$p(222\|h_k)=p_8$	$p(212\|h_k)=p_{16}$		
	$p(100\|h_k)=p_{17}$		
	$p(102\|h_k)=p_{18}$		
	$p(120\|h_k)=p_{19}$		
	$p(122\|h_k)=p_{20}$		

where $p_{27} = 1 - \sum_{m=1}^{26} p_m$. Then, for each individual i with history k (suppressing the subscripts h_k for notational convenience)

$$
\begin{aligned}
E\left(\frac{\sum 1}{3}\right) &= E\left[\frac{1}{3}\sum_q \mathbf{1}(Y_{qi} = 1)\right] \\
&= 0(p_1 + ... + p_8) + \frac{1}{3}(p_9 + ... + p_{20}) + \frac{2}{3}(p_{21} + ... + p_{26}) + 1p_{27} \\
&= \frac{1}{3}(p_9 + ... + p_{20}) + \frac{2}{3}(p_{21} + ... + p_{26}) + (1 - \sum_{m=1}^{26} p_m) \\
&= m
\end{aligned}
\tag{B1}
$$

and

$$
\begin{aligned}
Var\left(\frac{\sum 1}{3}\right) &= (-m)^2(p_1 + ... + p_8) + \left(\frac{1}{3} - m\right)^2(p_9 + ... + p_{20}) \\
&\quad + \left(\frac{2}{3} - m\right)^2(p_{21} + ... + p_{26}) + (1 - m)^2(1 - \sum_{m=1}^{26} p_m) \\
&= s^2
\end{aligned}
\tag{B2}
$$

In practice, the p_i are estimated as sample fractions. For the n_h individuals with a common history follows

$$E\left(\frac{1}{n_h}\sum_i m\right) = m_h \quad \text{and} \tag{B3}$$

$$Var\left(\frac{1}{n_h}\sum_i \left(\frac{\sum 1}{3}\right)\right) = \frac{1}{n_h} s^2 = s_h^2 \tag{B4}$$

which yields the variance for both elements of the difference in (5.9). The variance of (5.9) then results from the sum of the two history-specific variances (B4) for treated and comparison units. Parallel to the derivation of the overall treatment effect (5.10) from the history-specific effect (5.9), the variance of (5.10) is a weighted sum (with squared weights) of the variance of (5.9).

Chapter 6

Conclusion

> *The most challenging empirical questions in economics involve "what if" statements about counterfactual outcomes.*
>
> *– J. Angrist and A. Krueger (1999) –*

At the end of this thesis, it probably does not come as a surprise that I entirely subscribe to this point of view expressed by Angrist and Krueger in their chapter in the 1999 Handbook of Labor Economics. Having departed from disillusioning diagnostics on the current state of European labor markets, mainly expressed in terms of high and persistent unemployment rates, we have seen that European policy makers engage in measures of Active Labor Market Policy to combat this phenomenon of widespread unemployment. The empirical question induced by this undertaking clearly is of causal nature: Do ALMP programs actually *cause* unemployment to decrease? This question is of immediate policy concern, and the answer to it of decisive interest for decision makers in European governments.

This thesis has assessed quite a number of pivotal aspects that are involved in answering this question. At the outset, Chapter 2 has presented the very foundations of causal inference in the empirical sciences. I have discussed different ways of modeling causation, and explicitly drawn the connection between the counterfactual theory of causation in philosophical logic and a specific statistical model for causal inference based on counterfactuals. This model is known as the Potential Outcome Model (POM). The representation has given new insight into the foundations of the POM and the way it is being applied. The model is of especial interest for answering the above question, as it is

predominantly used in evaluation research – the type of research aiming at identifying ALMP program effectiveness. In particular, chapter 2 has shown why we need counterfactuals to address causal questions, and which causally meaningful "what if" questions can be answered within the model.

What would have happened to unemployed individuals participating in an ALMP program in terms of their labor market success if they had not participated in the program? This is the counterfactual question that evaluation studies on European ALMP seek to answer in order to infer the causal effect of the program on some response variable indicating individual labor market fortune. Chapter 3 has shown how the pressure of rising unemployment in Europe led policy makers to react, and how for EU member states this resulted in a concerted action called the Luxembourg Process, a joint employment strategy developed and started in 1997. Since ALMP programs form a major part of this employment strategy, Chapter 3 has devoted much attention to current ALMP practice in Europe. As a result, we find that both the "policy side" and the "science side" have made considerable progress in recent years: Policy in the form of launching the Luxembourg Process, and science in further developing appropriate evaluation tools. However, the two seem to be largely disconnected from each other. Frequently programs are implemented without any deliberate evaluation effort, and scientists often conduct studies without any possibility to communicate their results to decision makers. The main lesson from this account is that only a tighter connection of those who ask the "what if" question – the decision makers – with those who know the answer – the researchers – can help combat European unemployment effectively.

Nonetheless, this thesis has extracted those lessons on European ALMP effectiveness that can be drawn from the available evidence. While chapter 3 has done so from a global European perspective, chapters 4 and 5 have pinpointed detailed program effects of ALMP in one particular European OECD country, Poland. Throughout chapters 4 and 5, the discussion has also centered upon methodical issues. These chapters delineate how matching methods – a variant of the POM for observational studies – can identify the causal effect of program participation on employment probability, and offer ample illustration for the claim that this is the appropriate empirical approach meticulously adapted to the Polish data.

For the purposes of summing up, I will be as explicit as possible with respect to the overall lessons that should be drawn for European ALMP from this thesis. Generally speaking, one should not expect too much from European Active Labor Market Policy.

Whereas the cross-country evidence does make out some effective programs, the number of disappointing results cannot be ignored. Regarding the broad ALMP classification given in chapter 1, training seems to be the most promising program, even though country studies do not unanimously report positive effects. Programs providing jobs in the public sector yield very disappointing results and should be largely abolished given the current evidence. Wage subsidy schemes to the private sector apparently do seem to work in quite a number of cases, although the incentive structure needs to be well-specified. Moreover, basic ALMP programs like job search assistance appear to be helpful.

It is certainly undisputed among economists that a practice of unconditional payment of unemployment benefits for an indefinite period of time is associated with high European unemployment (Nickell 1997). Some even argue for a causal relation between the two (Layard et al. 1991, Ljungqvist and Sargent 1998). Such unconfined benefit regulations also display strong distorting effects on ALMP programs, as we have seen that the renewal of benefit receipt eligibility conditional on program participation results in devastatingly negative program impacts. Across countries there is abundant evidence that a labor market program will not generate positive outcomes if individuals merely participate in order to re-enter unemployment with renewed benefit receipt afterwards. Still, many countries follow this practice. This is one of the most robust and most irritating results on ALMP in Europe.

Methodical implications from both the review in chapter 3 and the specific country studies in chapters 4 and 5 are clear: Empirical evaluation research in labor economics *is* capable of answering the relevant "what if" questions regarding ALMP effectiveness with confidence. The hope remains that European policy makers will aggrandize their inclination to listen to the answers.

References

Angrist, J.D. (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants", *Econometrica* 66, 249-288.

Angrist, J.D. and J. Hahn (1999), "When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects", *NBER Technical Working Paper* 241, Cambridge, MA.

Angrist, J.D., G.W. Imbens and D.B. Rubin (1996a), "Identification of Causal Effects Using Instrumental Variables" (with discussion), *Journal of the American Statistical Association* 91, 444-472.

Angrist, J.D., G.W. Imbens and D.B. Rubin (1996b), "Rejoinder", *Journal of the American Statistical Association* 91, 468-472.

Angrist, J.D. and A.B. Krueger (1999), "Empirical Strategies in Labor Economics", in O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics* 3, 1277-1368.

Ashenfelter, O. (1978), "Estimating the Effect of Training Programs on Earnings", *Review of Economics and Statistics* 60, 47-57.

Ashenfelter, O. and D. Card (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs", *Review of Economics and Statistics* 67, 648-660.

Augurzky, B. (2000), "Optimal Full Matching", *Dept. of Economics Discussion Paper* 310, University of Heidelberg.

Augurzky, B. and C.M. Schmidt (2000), "The Propensity Score: A Means to an End", *Dept. of Economics Discussion Paper* 334, University of Heidelberg.

Bassi, L.J. (1983), "The Effect of CETA on the Post-Program Earnings of Participants", *The Journal of Human Resources* 18, 539-556.

Bell, B., R. Blundell and J. van Reenen (1999), "Getting the Unemployed Back to Work: The Role of Targeted Wage Subsidies", *International Tax and Public Finance* 6, 339-360.

Bennett, J. (1993), "Event Causation: The Counterfactual Analysis", in E. Sosa and M. Tooley (eds), *Causation*, Oxford University Press: Oxford.

Björklund, A. and R. Moffitt (1987), "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models", *Review of Economics and Statistics* 69, 42-49.

Blanchard, O. (1997), *The Economics of Post-Communist Transition*, Clarendon Press: Oxford.

Boeri, T. (1994), " 'Transitional' unemployment", *Economics of Transition* 2, 1-25

Boeri, T. (1997), "Learning from Transition Economies: Assessing Labour Market Policies across Central and Eastern Europe", *Journal of Comparative Economics* 25, 366-384.

Bonnal, L., D. Fougère and A. Sérandon (1997), "Evaluating the Impact of French Employment Policies on Individual Labour Market Histories", *Review of Economic Studies* 64, 683-713.

Brodaty, T., B. Crépon and D. Fougère (2001), "Using Matching Estimators to Evaluate Alternative Youth Employment Programs: Evidence from France, 1986-1988", in M. Lechner and F. Pfeiffer (eds), *Econometric Evaluation of Labour Market Policies*, forthcoming.

Calmfors, L. (1994), "Active Labour Market Policy and Unemployment - A Framework for the Analysis of Crucial Design Features", *OECD Labour Market and Social Policy Occasional Papers* 15, OECD: Paris.

Calmfors, L. (1995), "What Can We Expect from Active Labour Market Policy?", *Konjunkturpolitik* 43, 1-30.

Card, D. and D. Sullivan (1988), "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment", *Econometrica* 56, 497-530.

Cochran, W.G. (1965), "The Planning of Observational Studies of Human Populations" (with discussion), *Journal of the Royal Statistical Society* Series A 128, 234-266.

Council of the European Union (2001), "Council Decision of 19 January 2001 on Guidelines for Member States' employment policies for the year 2001", *Official Journal of the European Communities* L22.

Cox, D.R. (1958), *Planning of Experiments*, Wiley: New York.

Dawid, A.P. (1979), "Conditional Independence in Statistical Theory", *Journal of the Royal Statistical Society* Series B 41, 1-31.

Dawid, A.P. (2000), "Causal Inference Without Counterfactuals" (with discussion), *Journal of the American Statistical Association* 95, 407-448.

Dehejia, R.H., and S. Wahba (1998), "Propensity Score Matching Methods for Non-Experimental Causal Studies", *NBER Working Paper* 6829.

Einstein, A., B. Podolsky, and N. Rosen (1935), "Can the Quantum Mechanical Description of Reality Be Considered Complete?", *Physical Review* 47, 777-780.

European Commission (1997a), *The European Jobs Summit – Frequently Asked Questions*, http://europa.eu.int/comm/employment_social/elm/summit/en/comments/faqs.htm

European Commission (1997b), *The 1998 Employment Guidelines*, http://europa.eu.int/comm/employment_social/elm/summit/en/papers/guide2.htm

European Commission (1998), *Employment Policies in the EU and in the Member States – Joint Report 1998*, European Commission DG V: Brussels.

European Commission (2000), *Joint Employment Report 2000*, European Commission DG V: Brussels.

Fisher, R.A. (1935), *The Design of Experiments*, Oliver & Boyd: Edinburgh.

Fougère, D., F. Kramarz, and T. Magnac (2000), "Youth employment policies in France", *European Economic Review* 44, Papers & Proceedings, 928-942.

Fraker, T. and R. Maynard (1987), "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs", *The Journal of Human Resources* 22, 194-227.

Galles, D. and J. Pearl (1998), "An Axiomatic Characterization of Causal Counterfactuals", *Foundations of Science* 3, 151-182.

Gerfin, M. and M. Lechner (2000), "Microeconometric Evaluation of the Active Labour Market Policy in Switzerland", *IZA Disc. paper* No. 154, IZA: Bonn.

Glymour, C. (1986), "Statistics and Metaphysics", comment on Holland (1986), *Journal of the American Statistical Association* 81, 964-966.

Goldberger, A. (1972), "Structural Equation Methods in the Social Sciences", *Econometrica* 40, 979-1001.

Good, I.J. (1961, 1962, 1963), "A Causal Calculus I and II", *British Journal for the Philosophy of Science* 44, 305-318, and 45, 43-51, and "Errata and Corrigenda", ibid. 46, 88.

Góra, M. and C.M. Schmidt (1998), "Long-term unemployment, unemployment benefits and social assistance: The Polish experience", *Empirical Economics* 23, 55-85.

Góra, M., H. Lehmann, M. Socha and U. Sztanderska (1996), "Labour Market Policies in Poland", *OECD Proceedings - Lessons from Labour Market Policies in the Transition Countries*, 151-176, OECD: Paris.

Granger, C. (1969), "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods", *Econometrica* 37, 424-438.

Granger, C. (1986), "Comment on 'Statistics and Causal Inference' by P.W. Holland", *Journal of the American Statistical Association* 81, 967-968.

Greenland, S. (2000), "Causal Analysis in the Health Sciences", *Journal of the American Statistical Association* 95, 286-289.

Gu, X.S. and P.R. Rosenbaum (1993), "Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms", *Journal of Computational and Graphical Statistics* 2, 405-420.

Haavelmo, T. (1943), "The Statistical Implications of a System of Simultaneous Equations", *Econometrica* 11, 1-12.

Haavelmo, T. (1944), "The Probability Approach in Econometrics", *Econometrica* 12, supplement.

Hahn, J. (1998), "On the Role of the Propensity Score in the Efficient Semi-parametric Estimation of Average Treatment Effects", *Econometrica* 66, 315-332.

Ham, J.C. and R.J. LaLonde (1996), "The Effect of Sample Selection and Initial Conditions in Duration Models: Evidence from Experimental Data on Training", *Econometrica* 64, 175-205.

Heckman, J.J. (1992), "Randomization and Social Policy Evaluation", in C. Manski and I. Garfinkel (eds), *Evaluating Welfare and Training Programs*, Harvard University Press: Cambridge, MA, 201-230.

Heckman, J.J. (1996a), "Randomization as an Instrumental Variable." *Review of Economics and Statistics* 78, 336-341.

Heckman J.J. (1996b), "Comment on 'Identification of Causal Effects Using Instrumental Variables'" by J.D. Angrist, G.W. Imbens, and D.B. Rubin, *Journal of the American Statistical Association* 91, 459-462.

Heckman, J.J. (2000), "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective", *Quarterly Journal of Economics* 115, 45-97.

Heckman, J.J. and V.J. Hotz (1989), "Rejoinder to Comments on 'Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training'", *Journal of the American Statistical Association* 84, 878-880.

Heckman, J.J., H. Ishimura and P.E. Todd (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *Review of Economic Studies* 64, 605-654.

Heckman, J.J., H. Ichimura and P.E. Todd (1998), "Matching as an Econometric Evaluation Estimator", *Review of Economic Studies* 65, 261-294.

Heckman, J.J., R.J. LaLonde and J.A. Smith (1999), "The Economics and Econometrics of Active Labour Market Programs", in O. Ashenfelter and D. Card (eds), *Handbook of Labour Economics*, vol. III, North-Holland: Amsterdam et al.

Heckman, J.J. and J.A. Smith (1999), "The Pre-programme Earnings Dip and the Determinants of Participation in a Social Programme: Implications for Simple Programme Evaluation Strategies", *The Economic Journal* 109, 313-348.

Hitchcock, C. (1997), "Probabilistic Causation", *The Stanford Encyclopedia of Philosophy* (Spring 2001 Edition), Edward N. Zalta (ed.), URL= http://plato.stanford.edu/entries/causation~probabilistic/

Holland, P.W. (1986), "Statistics and Causal Inference" (with discussion), *Journal of the American Statistical Association* 81, 945-970.

Holland, P.W. (1988a), "Causal Inference, Path Analysis, and Recursive Structural Equation Models", *Sociological Methodology* 18, 449-484.

Holland, P.W. (1988b), "Causal Mechanism or Causal Effect: Which Is Best for Statistical Science?", Comment on "Employment Discrimination and Statistical Science" by A. P. Dempster, *Statistical Science* 3, 186-188.

Holland, P.W. (1989), "It's Very Clear", comment on 'Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training' by J.J. Heckman and V.J. Hotz, *Journal of the American Statistical Association* 84, 875-877.

Holland, P.W. and D.B. Rubin (1983), "On Lord's Paradox", in H. Wainer and S.Messick (eds), *Principals of Modern Psychological Measurement*, L. Erlbaum, Hillsdale, NJ.

Holland, P.W. and D.B. Rubin (1988), "Causal Inference in Retrospective Studies", *Evaluation Review* 12, 203-231.

Horwich, P. (1993), "Lewis's Programme", in E. Sosa and M. Tooley (eds), *Causation*, Oxford University Press: Oxford.

Hume, D. (1740b [1993]), "An Abstract of A Treatise of Human Nature", in *An Enquiry Concerning Human Understanding*, as reprinted by Hackett Publishing Co.: Indianapolis, IN, Eric Steinberg (ed).

Hume, D. (1740a [1992]), *Treatise of Human Nature*, as reprinted by Prometheus Books: Buffalo, NY.

Hume, D. (1748 [1993]), *An Enquiry Concerning Human Understanding*, as reprinted by Hackett Publishing Co.: Indianapolis, IN, Eric Steinberg (ed.), based on the 1777 posthumous edition.

Imbens, G.W. (2000), "The Role of Propensity Score in Estimating Dose-Response Functions", *Biometrika* 87, 706-710.

Imbens, G.W., and D.B. Rubin (1995), "Discussion of 'Causal Diagrams for empirical research' by J. Pearl, *Biometrika* 82, 694-695.

Jensen, P. (1999), "The Danish Youth Unemployment Programme", *mimeo*, Center for Labour Market and Social Research: Aarhus.

Jensen, P., M. Svarer Nielsen and M. Rosholm (2000), "The Effects of Benefits, Incentives, and Sanctions on Youth Unemployment", revised version of *Working Paper* 99-05, Center for Labour Market and Social Research: Aarhus.

Kluve, J. (1998), "The Evaluation of Active Labour Market Policies in Poland – An Application of Matching Estimators", *MLitt.Thesis*, University of Dublin, Trinity College.

Kluve, J., H. Lehmann and C.M. Schmidt (1999), "Active Labour Market Policies in Poland: Human Capital Enhancement, Stigmatization, or Benefit Churning", *Journal of Comparative Economics* 27, 61-89.

Kluve, J., H. Lehmann, C.M. Schmidt and M. Góra (1998), "Active Labor Market Policies in Poland: A Matching Approach" *mimeo*, Heidelberg and Leuven.

Koopmans, T. and W. Hood (1953), "The Estimation of Simultaneous Linear Economic Relationships", in W. Hood and T. Koopmans (eds), *Studies in Econometric Method*, Chapman & Hall: New York.

Lalive, R., J. Zweimüller and J.C. van Ours (2000), "The Impact of Active Labour Market Programs and Benefit Entitlement Rules on the Duration of Unemployment", University of Zurich, *IEW Working Paper* No. 41.

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

Larsson, L. (2000), "Evaluation of Swedish youth labour market programmes", Uppsala University, Dept. of Economics *Working paper* 2000-6.

Layard, R., S. Nickell and R. Jackman (1991), *Unemployment: Macroeconomic Performance and the Labour Market*, Oxford University Press: Oxford.

Leamer, E.E. (1983), "Let's Take the Con Out of Econometrics", *American Economic Review* 73, 31-43.

Lechner, M. (1997), "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany After Unification", *mimeo*, University of Mannheim.

Lechner, M. (2000), "An Evaluation of Public-Sector-Sponsored Continuous Vocational Training Programs in East Germany", *The Journal of Human Resources* 35, 347-375.

Lechner, M. (2001a), "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption", in M. Lechner and F. Pfeiffer (eds), *Econometric Evaluation of Labour Market Policies*, forthcoming.

Lechner, M. (2001b), "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies", *Review of Economics and Statistics*, forthcoming.

Lehmann, H. (1998), "Active Labor Market Policies in Central Europe: First Lessons", in R.T. Riphahn, D.J. Snower and K.F. Zimmermann (eds), *Employment Policy in the Transition: Lessons from German Integration*, Springer: Berlin and London, forthcoming.

Lewis, D. (1973a), *Counterfactuals*, Blackwell: Oxford.

Lewis, D. (1973b), "Causation", *Journal of Philosophy* 70, 556-567.

Lewis, D. (1979), "Counterfactual Dependence and Time's Arrow", *NOÛS* 13, 455-476.

Lewis, D. (1986), *Philosophical Papers: Volume II*, Oxford University Press: Oxford.

Lewis, D. (2000), "Causation as Influence", *Journal of Philosophy* 98, 182-197.

Ljungqvist, L. and T.J. Sargent (1998), "The European Unemployment Dilemma", *Journal of Political Economy* 106, 514-550.

Machin, S. and A. Manning (1999), "The Causes and Consequences of Longterm Unemployment in Europe", in O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics* 3, 3085-3139.

Magnac, T. (2000), "Subsidised Training and Youth Employment: Distinguishing Unobserved Heterogeneity from State Dependence in Labour Market Histories", *Economic Journal* 110, 805-837.

Marshall, A. (1890 [1965]), *Principles of Economics*, 8th ed., repr., Macmillan: London.

Martin, J.P. (2000), "What works among Active Labour Market Policies: Evidence from OECD Countries' experience", *OECD Economic Studies* 30.

Menzies, P. (2001a), "Counterfactual Theories of Causation", *The Stanford Encyclopedia of Philosophy* (Spring 2001 Edition), Edward N. Zalta (ed.), URL= http://plato.stanford.edu/entries/causation~counterfactual/

Menzies, P. (2001b), "Difference-Making in Context", in J. Collins, N. Hall and L. Paul (eds), *Causation and Counterfactuals*, MIT Press, forthcoming.

Morgan, M. (1990), *The History of Econometric Ideas*, Cambridge University Press: Cambridge.

Neyman, J. (1923 [1990]), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.", translated and edited by D.M. Sabrowska and T.P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczych Tom X (1923), 1-51 (Annals of Agriculture), *Statistical Science* 5, 465-472.

Neyman, J. (1935), with co-operation by K. Iwaszkiewicz, and S. Kolodziejczyk, "Statistical Problems in Agricultural Experimentation" (with discussion), *Supplement to the Journal of the Royal Statistical Society* 2, 107-180.

Nickell, S. (1997), "Unemployment and Labor Market Rigidities: Europe versus North America", *Journal of Economic Perspectives*, 11, 55-74.

OECD (1990), *Labour Market Policies for the 1990s*, OECD: Paris.

OECD (1991), *Labour Force Statistics 1969-1989*, OECD: Paris

OECD (1993), *Employment Outlook*, July, OECD: Paris.

OECD (1994), *The OECD Jobs Study – Facts, Analysis, Strategies*, OECD: Paris.

OECD (1998), *Labour Force Statistics 1977-1997*, OECD: Paris.

OECD (2000a), *Social Expenditure Database 1980-1997*, 2nd ed., CD-ROM, OECD: Paris.

OECD (2000b), *Employment Outlook*, June, OECD: Paris

OECD (2000c), *Labour Force Statistics 1979-1999*, October, OECD: Paris.

OECD (2000d), *Economic Outlook*, No. 68, December, OECD: Paris.

O'Leary, C.J. (1998), "Evaluating the Effectiveness of Active Labor Programs in Poland" W.I.E. Upjohn Institute for Employment Research, Kalamazoo, Michigan.

Pearl, J. (1995), "Causal diagrams for empirical research" (with discussion), *Biometrika* 82, 669-710.

Pearl, J. (1997), "The New Challenge: From a Century of Statistics to an Age of Causation", *Computing Science and Statistics* 29, 415-423.

Pearl, J. (1998), "Graphs, Causality, and Structural Equation Models", *Sociological Methods and Research* 27, 226-284.

Pearl, J. (2000a), *Causality: Models, Reasoning, and Inference*, Cambridge University Press: Cambridge.

Pearl, J. (2000b), Comment on "Causal Inference Without Counterfactuals" by A.P. Dawid, *Journal of the American Statistical Association* 95, 428-431.

Pearl, J. (2001), "Causal Inference in the Health Sciences: A Conceptual Introduction", UCLA Computer Science Dept., *Cognitive Systems Laboratory Technical Report* R-282.

Pratt, J.W. and R. Schlaiffer (1988), "On the Interpretation and Observation of Laws", *Journal of Econometrics* 39, 23-52.

Profit, S. and R. Tschernig (1998), "Germany's Labour Market Problems: What to Do and What Not to Do? A Survey Among Experts", *ifo Studien* 44, 307-325.

Puhani, P.A. (1998), "Advantage through Training? A Microeconometric Evaluation of the Employment Effects of Active Labour Market Programmes in Poland", *ZEW Disc. Paper 98-25*, Mannheim.

Puhani, P.A. and V. Steiner (1997), "The Effectiveness and Efficiency of Active Labour Market Programmes in Poland", *Empirica* 24, 209-231.

Quandt, R.E. (1958), "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes", *Journal of the American Statistical Association* 53, 873-880.

Quandt, R.E. (1972), "A New Approach to Estimating Switching Regressions", *Journal of the American Statistical Association* 67, 306-310.

Raaum, O. and H. Torp (2001), "Labour Market Training in Norway – Effect on Earnings", *Labour Economics*, forthcoming.

Regnér, H. (2001), "A nonexperimental evaluation of training programs for the unemployed in Sweden", *Labour Economics*, forthcoming.

Reichenbach, H. (1956), *The Direction of Time*, University of California Press: Berkeley and Los Angeles.

Robins, J.M. (1986), "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period – Application to Control of the Healthy Worker Survivor Effect", *Mathematical Modelling* 7, 1393-1512.

Robins, J.M. (1987), "Addendum to ' A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period – Application to Control of the Healthy Worker Survivor Effect'", *Comput. Math. Applic.* 14, 923-945.

Robins, J.M. (1995), "Discussion of 'Causal Diagrams for empirical research' by J. Pearl", *Biometrika* 82, 695-698.

Robins, J.M. and S. Greenland (2000), Comment on "Causal Inference Without Counterfactuals" by A.P. Dawid, *Journal of the American Statistical Association* 95, 431-435.

Rosenbaum, P.R. (1984), "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment", *Journal of the Royal Statistical Society* Series A 147, 656-666.

Rosenbaum, P.R. (1995a), *Observational Studies*, Springer: New York.

Rosenbaum, P.R. (1995b), "Discussion of 'Causal Diagrams for empirical research' by J. Pearl", *Biometrika* 82, 698-699.

Rosenbaum, P.R. and D.B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika* 70, 41-55.

Rosenbaum, P.R. and D.B. Rubin (1984a), "Estimating the Effects Caused by Treatments", Comment on "On the Nature and Discovery of Structure" by J.W. Pratt and R. Schlaifer, *Journal of the American Statistical Association* 79, 26-28.

Rosenbaum, P.R. and D.B. Rubin (1984b), "Reducing Bias in Observational Studies using Subclassification on the Propensity Score", *Journal of the American Statistical Association* 79, 516-524.

Rosholm, M. (1999), "Evaluating Subsidized Employment Programmes in the Private and Public Sector", *mimeo*, Center for Labour Market and Social Research: Aarhus.

Roy, A.D. (1951), "Some Thoughts on The Distribution of Earnings", *Oxford Economic Papers* 3, 135-146.

Rubin, D.B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology* 66, 688-701.

Rubin, D.B. (1977), "Assignment to Treatment Group on the Basis of a Covariate", *Journal of Educational Statistics* 2, 1-26.

Rubin, D.B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization", Annals of Statistics 6, 34-58.

Rubin, D.B. (1980), Comment on "Randomization Analysis of Experimental Data: The Fisher Randomization Test" by D. Basu, Journal of the American Statistical Association 75, 591-593.

Rubin, D.B. (1986), "Which Ifs Have Causal Answers", Comment on Holland (1986), Journal of the American Statistical Association 81, 961-962.

Rubin, D.B. (1990), "Neyman (1923) and Causal Inference in Experiments and Observational Studies", comment on Neyman (1923), *Statistical Science* 5, 472-480.

Salmon, W. (1980), "Probabilistic Causality", *Pacific Philosophical Quarterly* 61, 50-74.

Salmon, W. (1998), *Causality and Explanation*,Oxford University Press: New York and Oxford.

Schmidt, C.M. (1999), "Do we need Social Experiments? Potential and Limits of Non-experimental Project Evaluation", *mimeo*, UC Berkeley.

Schmidt, C.M. (1999), "Knowing What Works – The Case for Rigorous Program Evaluation", *IZA Discussion Paper* 77, IZA: Bonn.

Sianesi, B. (2001), "An Evaluation of the Active Labour Market Programmes in Sweden", *mimeo*, University College London.

Simon, H.A. and N. Rescher (1966), "Cause and Counterfactual", *Philosophy of Science* 33, 323-340.

Skyrms, B. (1984), "EPR: Lessons for Metaphysics", in P. French, T. Uehling, H. Wettstein (eds), *Midwest Studies in Philosophy IX – Causation and Causal Theories*, University of Minnesota Press: Minneapolis.

Skyrms, B. (1988), "Probability and Causation", *Journal of Econometrics* 39, 53-68.

Sobel, M.E. (1995), "Causal Inference in the Social and Behavioral Sciences", in G. Arminger, C. C. Clogg, and M. E. Sobel (eds), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Plenum Press, New York.

Sosa, E. and M. Tooley (1993a) (eds), *Causation*, Oxford University Press: Oxford.

Sosa, E. and M. Tooley (1993b), "Introduction", in E. Sosa and M. Tooley (eds), *Causation*, Oxford University Press: Oxford.

Speed, T. J. (1990), "Introductory Remarks on Neyman (1923)", *Statistical Science* 5, 463-464.

Spirtes, P., C. Glymour and R. Scheines (2000), *Causation, Prediction, and Search*, 2nd ed., Springer: New York.

Stalnaker, R. (1984), *Inquiry*, Boston, MA: Bradford Books.

Stanley, M., L. Katz and A. Krueger (1999), "Impacts of Employment and Training Programs: The American Experience", *mimeo*, Harvard University.

Suppes, P. (1970), *A Probabilistic Theory of Causality*, North Holland: Amsterdam.

Suppes, P. (1984), "Conflicting Intuitions about Causality", in P. French, T. Uehling, H. Wettstein (eds), *Midwest Studies in Philosophy IX – Causation and Causal Theories*, University of Minnesota Press: Minneapolis.

Van den Berg, G.J. (2000), "Duration Models: Specification, Identification, and Multiple Durations", in: Heckman, J.J. and E. Leamer (eds.), *Handbook of Econometrics*, Volume V, North-Holland: Amsterdam, forthcoming.

Van den Berg, G.J. and B. van der Klaauw (2000), "Counseling and Monitoring of Unemployed Workers: Theory and Evidence from a Social Experiment", *mimeo*, Free University of Amsterdam.

Van Ours, J.C. (2000), "Do Active Labour Market Policies Help Unemployed Workers to Find and Keep Regular Jobs?", Tilburg University, *CentER Working Paper* 00-10.

Wright, S. (1921), "Correlation and Causation", *Journal of Agricultural Research* 20, 557-585.

Wright, S. (1934), "The Method of Path Coefficients", *Annals of Mathematical Statistics* 5, 161-215.

Zweimüller, J. and R. Winter-Ebmer (1996), "Manpower Training Programmes and Employment Stability", *Economica* 63, 113-130.

Acknowledgements

I am delighted to express my sincerest gratitude to the many people that have played a role in the evolution of my work during these PhD years. First and foremost, I want to thank Christoph M. Schmidt for his excellent supervision, his stimulating advice, his unflinching guidance, and his enduring patience. I am also deeply indebted to Hartmut Lehmann for his continuing and continuous advice and complementary guidance, and for letting me learn more about the countries of Eastern Europe. This thesis has benefited from valuable comments by Boris Augurzky, Giuseppe Bertola, Mark Blaug, David Card, Michael Fertig, Ruth Miquel, Bas van der Klaauw, and Klaus F. Zimmermann. Moreover, I am grateful to comments by participants of a Volkswagen Foundation/Phare Ace workshop at Trinity College Dublin, a joint IZA/WDI conference at IZA Bonn, ESPE 2000 Bonn, and an IZA workshop on policy evaluation, as well as seminar participants at the Tinbergen Institute Amsterdam and the University of Heidelberg. I am grateful to David Card for his hospitality at the Center for Labor Economics, UC Berkeley, and to Gerard van den Berg for inviting me to an inspiring research visit at the Tinbergen Institute Amsterdam. Personally, I want to thank María Eugenia Sarasqueta Orte for being there for me, and my parents for constant support and keeping the faith. Last, not least, I am grateful to the Cusanuswerk for providing the financial support that rendered this dissertation possible.