Methods and limitations of evaluation and impact research

Reinhard Hujer, Marco Caliendo, Dubravko Radic

In:

Descy, P.; Tessaring, M. (eds)

The foundations of evaluation and impact research

Third report on vocational training research in Europe: background report.
Luxembourg: Office for Official Publications of the European Communities, 2004
(Cedefop Reference series, 58)

Reproduction is authorised provided the source is acknowledged

Additional information on Cedefop's research reports can be found on:
http://www.trainingvillage.gr/etv/Projects_Networks/ResearchLab/

For your information:

- the **background report** to the third report on vocational training research in Europe contains original contributions from researchers. They are regrouped in three volumes published separately in English only. A list of contents is on the next page.
- A **synthesis report** based on these contributions and with additional research findings is being published in English, French and German.

  Bibliographical reference of the English version:
  Descy, P.; Tessaring, M. *Evaluation and impact of education and training: the value of learning*. Third report on vocational training research in Europe: synthesis report. Luxembourg: Office for Official Publications of the European Communities (Cedefop Reference series)
- In addition, an **executive summary** in all EU languages will be available.

The background and synthesis reports will be available from national EU sales offices or from Cedefop.

For further information contact:

Cedefop, PO Box 22427, GR-55102 Thessaloniki
Tel.: (30)2310 490 111
Fax: (30)2310 490 102
E-mail: info@cedefop.eu.int
Homepage: www.cedefop.eu.int
Interactive website: www.trainingvillage.gr

Contributions to the background report of the third research report

Impact of education and training

Preface

The impact of human capital on economic growth: a review
Rob A. Wilson, Geoff Briscoe

Empirical analysis of human capital development and economic growth in European regions
Hiro Izushi, Robert Huggins

Non-material benefits of education, training and skills at a macro level
Andy Green, John Preston, Lars-Erik Malmberg

Macroeconometric evaluation of active labour-market policy – a case study for Germany
Reinhard Hujer, Marco Caliendo, Christopher Zeiss

Active policies and measures: impact on integration and reintegration in the labour market and social life
Kenneth Walsh and David J. Parsons

The impact of human capital and human capital investments on company performance Evidence from literature and European survey results
Bo Hansson, Ulf Johanson, Karl-Heinz Leitner

The benefits of education, training and skills from an individual life-course perspective with a particular focus on life-course and biographical research
Maren Heise, Wolfgang Meyer

The foundations of evaluation and impact research

Preface

Philosophies and types of evaluation research
Elliot Stern

Developing standards to evaluate vocational education and training programmes
Wolfgang Beywl; Sandra Speer

Methods and limitations of evaluation and impact research
Reinhard Hujer, Marco Caliendo, Dubravko Radic

From project to policy evaluation in vocational education and training – possible concepts and tools. Evidence from countries in transition.
Evelyn Viertel, Søren P. Nielsen, David L. Parkes, Søren Poulsen

Look, listen and learn: an international evaluation of adult learning
Beatriz Pont and Patrick Werquin

Measurement and evaluation of competence
Gerald A. Straka

An overarching conceptual framework for assessing key competences. Lessons from an interdisciplinary and policy-oriented approach
Dominique Simone Rychen

Evaluation of systems and programmes

Preface

Evaluating the impact of reforms of vocational education and training: examples of practice
Mike Coles

Evaluating systems' reform in vocational education and training. Learning from Danish and Dutch cases
Loek Nieuwenhuis, Hanne Shapiro

Evaluation of EU and international programmes and initiatives promoting mobility – selected case studies
Wolfgang Hellwig, Uwe Lauterbach, Hermann-Günter Hesse, Sabine Fabriz

Consultancy for free? Evaluation practice in the European Union and central and eastern Europe Findings from selected EU programmes
Bernd Baumgartl, Olga Strietska-Ilina, Gerhard Schaumberger

Quasi-market reforms in employment and training services: first experiences and evaluation results
Ludo Struyven, Geert Steurs

Evaluation activities in the European Commission
Josep Molsosa

Methods and limitations of evaluation and impact research

Reinhard Hujer, Marco Caliendo,
Dubravko Radić [1]

Abstract

The need to measure and judge the effects of social programmes, and the importance of evaluation studies in this context, is no longer questioned. Evaluation is a complex task and involves several steps. The following contribution mainly focuses on the methodological aspects of evaluation, particularly on econometric evaluation techniques. In consequence, we concentrate on evaluation studies conducted in the field of active labour-market policies (ALMP) and especially labour-market training (LMT) in Germany and Europe. The ideal evaluation process can be viewed as a series of three steps. First, the impacts of the programme on the individual should be estimated. Second, it should be determined whether the estimated impacts are large enough to yield net social gains. Finally, evaluation should question whether the best outcome have been achieved for the money spent (Fay, 1996). The focus of our paper is the first step, namely the microeconometric evaluation, although the other two steps, i.e. macroeconometric evaluation and cost-benefit analysis (CBA), are also discussed.

When discussing microeconometric evaluation, analysts such as LaLonde (1986) or Ashenfelter and Card (1985) view social experiments as the only valid evaluation method. A second group, including Heckman and Hotz (1989) and Lechner (1998), believe that it is possible to construct a comparison group using non-experimental data and econometric and statistical methods to solve the fundamental evaluation problem. This problem arises because we cannot observe individuals simultaneously with and without participation in a programme.

Since experimental data is scarce in Europe and evaluators often have to evaluate programmes already running, we focus on the techniques that deal with non-experimental data only. To do so we discuss the fundamental evaluation problem and different methods to solve this problem empirically. The methods presented include the before-after estimator (BAE), the cross-section estimator (CSE), the matching estimator, the difference-in-differences estimator (DID) and finally the duration model approach. Every estimator makes some generally untestable assumption to overcome the fundamental evaluation problem and therefore we also discuss the likelihood that these assumptions are met in practice. In addition we present a numerical example to show how the estimators are implemented technically and also discuss their data requirements.

Aggregated data can be used as well as individual data to evaluate the effects of ALMP programmes (including LMT). Instead of looking at the effect on individual performance, we would like to know if the ALMP represent a net gain to the whole economy. If the total number of jobs is not affected by labour-market policies, the effects will only be distributional. The need for a macroeconometric evaluation arises from the presence of dead-weight losses, displacement and substitution effects. Important

[1] Acknowledgements: the authors thank Björn Christensen, Pascaline Descy, Viktor Steiner, Manfred Tessaring, Stephan Thomsen and Christopher Zeiss for valuable comments and Yasemin Illgin, Paulo Rodrigues and Oliver Wünsche for research assistance.

methodological issues in a macroeconometric evaluation are the specification of the empirical model, which should always be based on an appropriate theoretical framework, and the simultaneity problem of ALMP which has to be solved.

Despite being able to reveal the impacts for the individuals and for the whole economy, micro- and macroeconometric evaluations do not cover the full range of effects associated with social programmes. Social programmes also involve various other, more qualitative, aims and objectives. Examples include distributional and equity goals, which are often hard to measure but nevertheless important. In order to capture these effects, a cost-benefit analysis should be conducted as a third step. Such an analysis widens the perspective of an impact analysis and contributes to better understanding of the complete effects of a social programme.

After having discussed the methodological issues related to the three evaluation steps, we present empirical findings from micro- and macroeconometric evaluations in Europe and, in particular, in Germany. We find that training programmes seem to have positive effects on the individuals in most of the studies and perform better than alternative labour-market programmes such as, for example, job creation schemes (JCS).

Finally, we use the results from our methodological discussion and the empirical findings to draw some implications for policy and evaluation practice. We focus on the choice of the appropriate estimation method, data requirements, the problem of heterogeneity in evaluation analysis, the successful design of training programmes and the transferability of our findings (for ALMP) to other social programmes.

Table of contents

1. Introduction 136

2. Methods of evaluation 139
- 2.1. Process of evaluation 139
  - 2.1.1. The evaluand and the time of the evaluation 139
  - 2.1.2. The evaluators 140
  - 2.1.3. The evaluation criteria 140
  - 2.1.4. The evaluation method 141
- 2.2. Microeconometric evaluation 142
  - 2.2.1. Fundamental evaluation problem 142
  - 2.2.2. Before-after estimator 144
  - 2.2.3. Cross-section estimator 146
  - 2.2.4. Matching estimator 146
  - 2.2.5. Difference-in-differences estimator 149
  - 2.2.6. Duration models 150
  - 2.2.7. An intuitive example 151
- 2.3. Macroeconometric evaluation 154
- 2.4. Cost-benefit analysis (CBA) 156
  - 2.4.1. Time of a CBA 158
  - 2.4.2. Selection of the alternatives 158
  - 2.4.3. Defining the reference group 158
  - 2.4.4. Enumerate and forecast all impacts 159
  - 2.4.5. Measuring and aggregating impacts 160
  - 2.4.6. Comparing costs and benefits 161
  - 2.4.7. Conducting sensitivity tests 161

3. ALMP in the EU and empirical findings 165
- 3.1. Recent developments in ALMP in the EU 165
- 3.2. Microeconometric evaluation studies 167
  - 3.2.1. Germany 167
  - 3.2.2. Other European countries 168
    - 3.2.2.1. Switzerland 168
    - 3.2.2.2. Austria 169
    - 3.2.2.3. Belgium 169
    - 3.2.2.4. France 169
    - 3.2.2.5. Denmark 169
    - 3.2.2.6. United Kingdom 169
    - 3.2.2.7. Ireland 169
    - 3.2.2.8. Norway 169
    - 3.2.2.9. Sweden 170
- 3.3. Macroeconometric evaluation studies 170
  - 3.3.1. Sweden 170
  - 3.3.2. Germany 170
- 3.4. Vocational training programme success in previous years 171

4. Policy implications: some guidance for evaluation and implementation practice 175
4.1. Selection problem and the choice of estimation method 175
4.2. Heterogeneity 176
4.3. Data requirements 176
4.4. Macroeconometric analysis and cost-benefit analysis 177
4.5. The design of training programmes: some suggestions 178
4.6. Transferability to other social programmes 179

5. Summary and conclusions 180

List of abbreviations 182

Annex: Tables 183

References 186

List of tables and figures

Tables

Table 1: Comparison of the different evaluation estimators 151
Table 2: Labour-market outcomes for participants and non-participants 152
Table 3: Estimates of the programme effects 154
Table 4: Hypothetical example before reallocation 162
Table 5: Hypothetical example after reallocation 162
Table 6: Costs and benefits of the Job Corps programme 163
Table 7: Expenditure on LMT as a percentage of total spending on ALMP 167
Table 8: Summary of the empirical findings from microeconometric studies 172
Table A1: Standardised unemployment rates in the EU (as a percentage of total labour force) 183
Table A2: Public expenditure on ALMP as a percentage of GDP 184
Table A3: Public expenditure on PLMP as a percentage of GDP 184
Table A4: Public expenditure on LMT as a percentage of GDP 185

Figures

Figure 1: Spending on ALMP, PLMP and LMT (EU average) and unemployment rates (EU average) from 1985 to 2000 137
Figure 2: Major aspects of an evaluation process 141
Figure 3: Ashenfelter's dip 145
Figure 4: Different stages of a CBA 157
Figure 5: Spending on ALMP and PLMP in the EU, 1985 166
Figure 6: Spending on ALMP and PLMP in the EU, 2000 166

1. Introduction

The need to measure and judge the effects of social programmes, and the importance of evaluation studies in this context, is no longer questioned. Evaluation is, however, a complex task and involves several steps. The range of topics is virtually unlimited and, since every topic requires a specific methodological approach, there exists no general evaluation strategy. To be successful, every evaluation must pre-specify a set of preliminary aspects. First it must be stated precisely which subject is to be evaluated. A second important question is 'who are the evaluators and what are their competences?'. Finally, and probably most important, the evaluation criteria and the chosen procedure for the evaluation must be fixed.

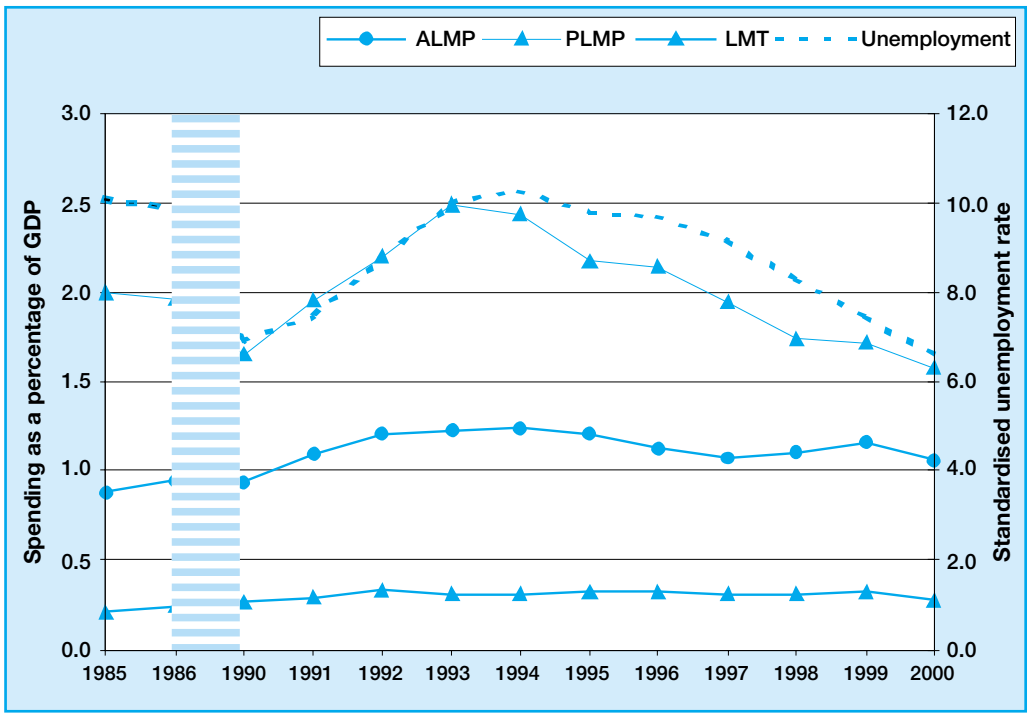
This contribution will focus mainly on the last aspect, i.e. offer some suggestions and advice on which econometric evaluation techniques should be used under different economic circumstances. In doing so, we will concentrate on evaluation studies conducted in the field of active labour-market policies (ALMP) and especially labour-market training (LMT). We regard them as fruitful object of analysis for several reasons. The increasing expenditure on ALMP in the last decades has raised a considerable interest in evaluation and motivated many studies in that field. Also the choice of the outcome variable that is of interest is more obvious than for other social programmes; the question of what should be defined as a success and how to measure it can be answered more easily. Whereas evaluation in the United States, for example, has mainly focused on earnings after participating in a certain programme, in Europe the focus is more on the employment prospects of the participants. We will address the issue of choosing an appropriate outcome variable later on. Finally, in contrast to health programmes, for example, there is a closer link between ALMP programmes, which are often short-term activities, and the outcome considered. Therefore, the evaluation of the effect of these programmes is often easier.

ALMP have been seen as one way to fight unemployment, rising in most European countries since the early 1970s. The growing interest in these measures is easy to understand in view of the disillusionment with more aggregate policies. Traditional demand stimulation has been discredited because it faces the risk of increasing inflation with only small effects on employment. Supply-side structural reforms, aimed at removing various labour-market rigidities, are difficult to implement or appear to produce results rather slowly. In this situation, as Calmfors (1994) notes, ALMP are regarded by many as the *deus ex machina* that will provide the solution to unemployment. Not only do they provide a more efficient outcome in the labour market, they also equip individuals with higher skills and therefore lower the risk of poverty. In this sense ALMP are capable of meeting efficiency and equity goals at the same time (OECD, 1991) ([2]). Another development that spurred interest in ALMP is the fact that it has become a common theme in the political debate that governments should shift the balance of public spending on labour-market policies away from passive income support towards more active measures designed to get the unemployed back into work. Anglo-Saxon policy-makers especially favour the idea of tying the right of welfare to the duty of work. Welfare then becomes workfare (Card, 2000).

This should manifest itself in a higher relative importance of ALMP. Figure 1 shows the average spending on active (including LMT) and passive labour-market policies (PLMP) as a percentage of the national gross domestic product (GDP) from 1985 to 2000 for the countries of the European Union (EU). The average spending on ALMP rose from 0.88 % of GDP in 1985 to 1.23 % in 1994. After that peak it went back to 1.06 % in 2000. The spending on PLMP, in contrast, peaked in 1993 at 2.49 % before it was reduced to 1.57 % in 2000. If one compares the relationship between the two measures in 1985 and 2000 the growing impor-

([2]) Several means by which ALMP might influence the labour markets, such as enhancing and adapting skills according to skill needs, improving individual employability and avoiding skill shortages, can be thought of.

Figure 1: **Spending on ALMP, PLMP and LMT (EU average) and unemployment rates (EU average) from 1985 to 2000**



tance of active measures becomes clearer. Whereas in 1985 the spending on PLMP was 126 % higher than the spending on ALMP, it was only 48 % higher in 2000. Even though this is a remarkable change in the structure of the spending, high amounts are still directed to passive measures.

One obvious reason for the limited success in switching resources into active measures is the fact that unemployment benefits are entitlement programmes. Rising unemployment automatically increases public spending on passive income support, whereas most of the active labour-market programmes are discretionary in nature and therefore easier to dispense with in a situation of tight budgets (Martin, 1998). This becomes clear from examination of Figure 1 which also shows standardised unemployment rates from 1985 to 2000. The EU average peaked at 10.1 % in 1985, 6.9 % in 1991 and 10.3 % in 1994. After that unemployment went down to 6.9 %. The movement of the standardised unemployment rates and spending on PLMP is nearly synchronous. LMT plays an important role as one measure of ALMP. The average spending on LMT in the EU rose from 0.21 % in 1985 to 0.28 % in 2000.

Clearly, the money spent on labour-market programmes is not available for other projects or private consumption. In an era of tight government budgets and a growing disbelief regarding the positive effects of ALMP ([3]), evaluation of these policies becomes imperative. Various evaluation methods have been developed for this. The ideal evaluation process can be viewed as a series of three steps. First, the impacts of the programme on the individual or groups of individuals should be estimated. Second, it should be

([3]) There have been several microeconometric studies which showed no positive effects of ALMP, especially for job creation schemes. See Hujer and Caliendo (2001) for an overview of the evaluation studies on vocational training and job creation schemes in West and East Germany.

determined whether the estimated impacts are large enough to yield net social gains. Finally, it should be decided if this is the best outcome that could have been achieved for the money spent (Fay, 1996) ([4]). The focus of our paper is the first step, namely the microeconometric evaluation, although the other two steps, i.e. macroeconometric evaluation and cost-benefit analysis (CBA), are discussed too.

Empirical microeconometric evaluation is conducted with individual data. The main question is whether the desired outcome variable for an individual is affected by participation in an ALMP programme. Relevant outcome variables could be the future employment probability or the future earnings. We would like to know the difference between the value of the participants' outcome in the actual situation and the value of the outcome if there had been no participation in the programme. The fundamental evaluation problem arises because we never observe both states (participation and non-participation) for the same individual at the same time. Therefore, finding an adequate control group is necessary to make a comparison possible. This is not easy because participants in programmes usually differ systematically from the non-participants in more aspects than just participation. Simply taking the difference between their outcomes after the programme will not reveal the true programme impact but will lead to a biased estimate. The literature on the solution to this problem is dominated by two points of view.

Analysts such as LaLonde (1986) or Ashenfelter and Card (1985) view social experiments as the only valid evaluation method. A second group, including Heckman and Hotz (1989) and Lechner (1998), believe that it is possible to construct a comparison group using non-experimental data and econometric and statistical methods to solve the fundamental evaluation problems. In non-experimental or observational studies, the data are not derived in a process that is completely under the control of the researcher. Instead, one has to rely on information about how individuals actually performed after the intervention; that is, we observe the outcome with treatment for participants and the outcome without treatment for non-participants.

The objective of observational studies is to use this information to restore the comparability of both groups by design. To do so, more-or-less plausible identification assumptions have to be imposed. There are several approaches differing with respect to the methods applied to this problem. Some studies control for observables as part of parametric evaluation models, others construct matched samples. Some authors think that conditioning on observables is not enough and that one has to take into account unobservables too. We present here several different estimation techniques, discuss the methodological concepts associated with them, highlight their (dis-) advantages and identify the environments under which they work best. While presenting the methodological concepts of each estimator, we also discuss the data requirements needed for their implementation. By doing this, we hope to give advice for the construction of datasets in the future.

The remainder of this paper is organised as follows:

(a) first, we present some methodological concepts of evaluation. We discuss microeconometric evaluation concepts in detail and also present some ideas on macroeconometric analysis and the cost-benefit approach;
(b) in the following chapter we present some standardised facts about the evolution of unemployment and labour-market policies in the EU, before we give an overview of the most relevant previous empirical findings regarding LMT in Germany and Europe;
(c) finally, we summarise the findings and give advice for the implementation and evaluation of training or other social programmes in the future.

([4]) The results can also be used to provide a feedback for improvement of subsequent programmes.

2. Methods of evaluation

2.1. Process of evaluation

The need to measure and judge the effects of social programmes, and the importance of evaluation studies in this context, is no longer questioned. Founders of these programmes increasingly ask for hard evidence of the efficient use of investments and enquire whether the programmes were successful or not. Evaluation is, however, a complex task that involves several steps. The range of topics is virtually unlimited and, since every topic requires a specific methodological approach, there is no general evaluation strategy.

Evaluation is, according to a definition of Worthen et al. (1997), the determination of the worth or merit of an evaluation object, the evaluand. It comprises (Worthen et al., 1997, p. 5) the 'identification, clarification, and application of defensible criteria to determine an evaluation object's value, quality, utility, effectiveness, or significance in relation to those criteria.' Its outstanding attribute, therefore, is its scientific claim. Another feature of evaluation that distinguishes it from basic research is its practical access. Evaluation and research are similar in that they use empirical methods and techniques to discover new knowledge. However, whereas research is primarily interested in advancing knowledge for its own sake, evaluation of social programmes is concerned with practical utilisation and matters such as:

(a) did the programme achieve what it was intended to do?
(b) who benefited from the programme?
(c) could the programme be conducted better or more efficiently?
(d) what changes must be undertaken to improve particular aspects of the programme?

To be successful, every evaluation must pre-specify a set of preliminary aspects (Kromrey, 2001). First, it must be stated precisely which subject should be evaluated and when this evaluation should take place. A second important point is to consider who the evaluators are and what their competences are. Finally, and probably most important, the evaluation criteria and the chosen procedure for the evaluation must be fixed.

2.1.1. The evaluand and the time of the evaluation

Defining the evaluation object seems straightforward at first sight. It comprises a detailed description of the programme which should be evaluated. A broad variety of programmes exists: those already implemented; those at the implementation stage; regionally concentrated pilot-programmes; and well established national programmes. However, a detailed description of the programme is not enough. Since an evaluation of every aspect of the programme would not be feasible, it must be clearly specified whether only the implementation of the programme should be evaluated or certain impacts of it.

Monitoring the implementation of the programme consists of a complete enumeration of what has happened during the different stages of its execution. It gives a first impression of the programme under consideration, for example the number of participants, the average time the agencies conducting the programmes spent with each client, the expected costs of the programme, the completion rate, the employment status reached after the participation, etc. Such information can be a form of control over the agents implementing the programme. Although this monitoring gives a first impression about the success or failure of a programme, it does not give any explanations for it. Monitoring is thus a reasonable first step in an evaluation process and can be seen as a minimum requirement to check how large sums of public money are spent. Evaluation, on the other hand, goes a step further and aims at determining whether a programme is successful or not by defining certain criteria and assessing whether these criteria were met.

Another question which has to be answered before the evaluation takes place concerns the time when an object should be evaluated. In this context, a distinction between formative and summative evaluation seems to be useful. If the evaluation takes place during the development

stage of the programme, and if the results of the evaluation are used to provide some feedback on how to improve it, a formative evaluation is conducted. Such an open formative evaluation is of practical value because of its feedback feature. However, the feedback of evaluation results on the programme does not allow for interpretation of the results in terms of success or efficiency since the evaluation itself influences the success of the programme. In addition, most programmes yield benefits only in the medium the long term and hence cannot be used as feedback for current programmes but, at best, as a basis for future improvements of similar programmes.

A summative evaluation, on the other hand, is conducted after the programme has been finished. Thereby an immediate influence on the results on the programme is abandoned. The major task of a summative evaluation is to determine whether the programme should be continued or terminated after it was carried out. Scriven (1991) puts the difference between the two approaches using the following illustration: 'When the cook tastes the soup, that's formative evaluation; when the guest tastes it, that's summative evaluation.'

2.1.2. The evaluators

Another important point is the question of who is authorised to conduct an evaluation. In an internal evaluation, the employees working for the programme are in charge of it. The alternative would be to assign this task to external evaluators, for example independent consultants or researchers. The advantage of an internal process is that the evaluators are familiar to the programme and that they have trouble-free access to all necessary information. One potential problem is the desiderative professionalism and objectivity. If, for example, the institution in charge of the evaluation is also responsible for the implementation of the programme, an incentive to find results that correspond to the aims and objectives of the programme could arise. This hazard can be avoided by external evaluators who are also able to bring in new views and ideas. But, even if the evaluation is done by internal staff, the transparency and accountability can be increased if there is some kind of cooperation with external institutions in certain areas, or if the material underlying the evaluation, for example datasets, etc., are made available to the scientific community.

2.1.3. The evaluation criteria

The *a-priori* fixing of criteria ensures that the evaluation is not done in an ad hoc manner but transparently and comprehensibly. There is a broad spectrum of potential criteria, including various direct and secondary impacts of the programme. In the case of ALMP for example, direct impacts could be on future employment probability or wages, whereas secondary effects would include potential displacement and crowding-out-effects. We will address these issues shortly in more detail. Other criteria could be the efficiency in conducting the programme or even the legitimacy of the objectives themselves. In this context an important question is where these criteria come from, i.e. who sets them?

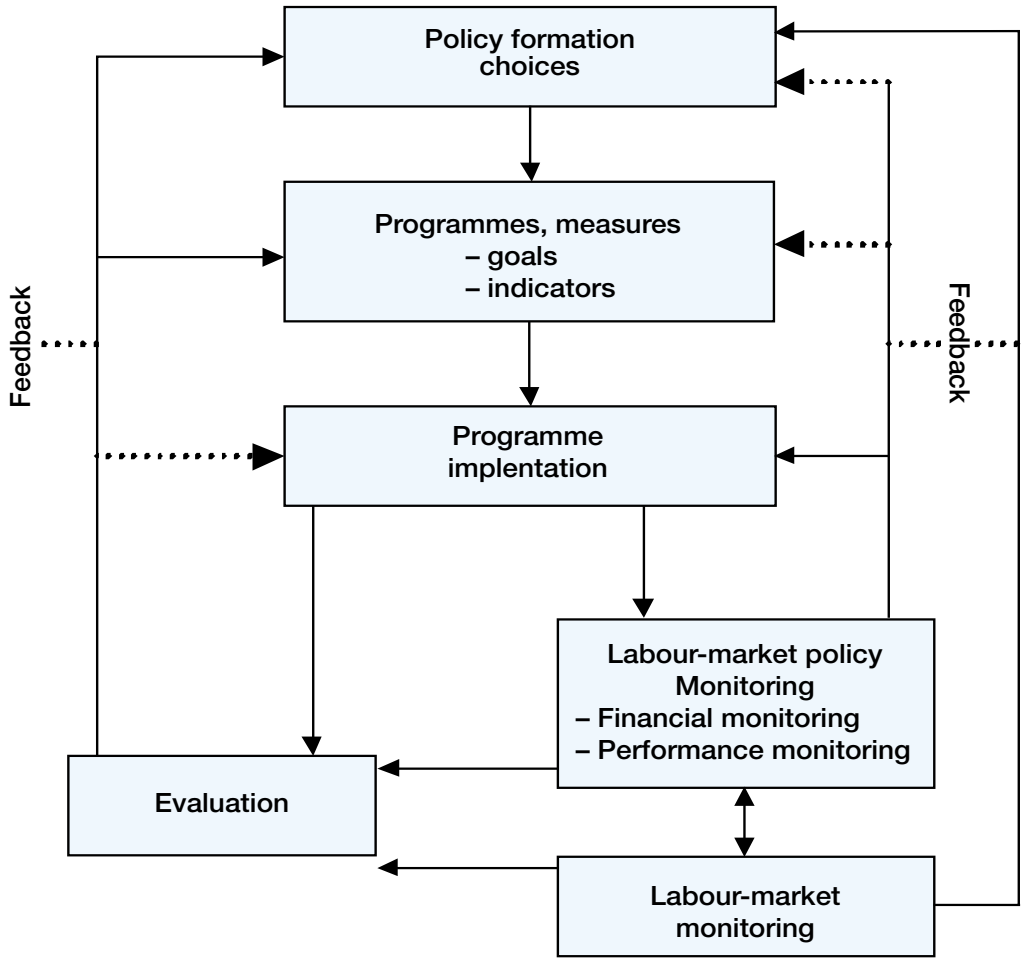
Typically, the criteria stem from the programme itself. If for example the programme to be evaluated aims at improving the re-employment probability of the disabled, then an obvious criterion would be the change in the re-employment probability attributable to this programme. The evaluation methods described in this contribution are mainly established on such quantitative measures. Whereas evaluation in the US, for example, has mainly focused on earnings after training participation, in Europe the training effect on employment plays a dominant role. This is not surprising, as unemployment in Europe is much higher and more persistent than in the US. But the same quantitative outcome, say improvement of employment situation, may plausibly be measured in several ways, for instance by hours per week in the new job or by a simple distinction between employment and unemployment. Schmidt (1999) notes, that there are some problems arising with the choice of an appropriate outcome measure. Outcomes may not be comparable across interventions. Therefore a policy-maker who has to decide which measure to implement will normally try to translate the gains of a programme into monetary terms or to carry out a so-called cost-utility analysis. Still this is not easy because new problems like time or group preferences emerge. We will pick up these problems in the following section. Trying to measure the gains of a social programme using monetary terms, however, does not mean that other more qualitative aims and objectives, such as the quality of life of the affected participants or fairness aspects, are of less importance. Other

important examples for more qualitative aims and objectives include equity goals, social inclusion, civic participation or reduction in crime. We will address these issues later on.

2.1.4. The evaluation method

Having defined the programme and the success criteria, evaluating the impact of a programme requires disentangling the effect of the programme from other exogenous factors. The programme can be seen as the independent variable, whereas the dependent variable is the success criterion. Various empirical strategies can be used to measure the impact of the independent variable. The most favoured empirical strategy is a natural experiment where individuals participating in social programmes and those who were excluded from participation are randomly selected so that differences in the outcome variable after the treatment, i.e. after the programme took place, are solely attributable to the programme. Although seen as the golden path in evaluation, experiments have their own drawbacks, the most severe being an ethical one. It is simply very hard to refuse help to some people who are supposed to be in need of it. Besides this ethical issue, experiments suffer also from other shortcomings, such as randomisation bias, Hawthorne effect, disruption and substitution bias. We will address these issues later on. Quasi-experimental strategies, on the other hand, rely on a non-randomly chosen group of non-participants and differ in the way the control group is constructed. Examples include the before-after-estimator or the matching-procedure. This paper will address issues concerned with these estimators.

Figure 2: **Major aspects of an evaluation process**



Source: Auer and Kruppe (1996)

Following Auer and Kruppe (1996) the major aspects which have to be taken into account when conducting an evaluation of social programmes are once again summarised in Figure 2. With the definition of the goals and objectives of the programme, appropriate indicators for measuring the success of the programme should be defined as well. These goals are then, once the programme has been implemented, confronted with the actual data obtained from a first monitoring. The following sections will mainly focus on the last aspect, i.e. derive some suggestions and advice on which econometric evaluation techniques should be used under different economic circumstances in order to conduct the evaluation.

2.2. Microeconometric evaluation

2.2.1. Fundamental evaluation problem

Inference about the impact of a treatment on the outcome of an individual involves speculation about how this individual would have performed in the labour market, had he or she not received the treatment ([5]). The framework serving as a guideline for the empirical analysis of this problem is the potential outcome approach, also known as the Roy-Rubin-model (Roy, 1951) (Rubin, 1974). In the basic model there are two potential outcomes (or responses), Y^1 and Y^0, for each individual, where Y^1 indicates an outcome with training and without. In the former case, the individual is in the treatment group and in the latter case it is in the comparison group. To complete the notation we define a binary assignment indicator D, indicating whether an individual actually participated in training (D=1) or not (D=0). The treatment effect for each individual is then defined as the difference between his/her potential outcomes:

$$\Delta = Y^1 - Y^0. \qquad (1)$$

The fundamental problem of evaluating this individual treatment effect arises because the observed outcome for each individual is given by:

$$Y = D \cdot Y^1 + (1 - D) \cdot Y^0. \qquad (2)$$

This means that for individuals who participated in training ($D = 1$) we observe Y^1 and for those who did not participate we observe Y^0. Unfortunately, we can never observe Y^1 and Y^0 for the same individual simultaneously and therefore we cannot estimate (1) directly. The unobservable component in (1) is called the counterfactual outcome, so that for an individual who participated in the training measure ($D = 1$), Y^0 is the counterfactual outcome, and for another one who did not participate it is Y^1.

The concentration on a single individual requires that the effect of the intervention on each individual is not affected by the participation decision of any other individual, i.e. the treatment effect Δ for each person is independent of the treatment of other individuals. In statistical literature (Rubin, 1980) this is referred to as the stable unit treatment value assumption (SUTVA) and guarantees that average treatment effects can be estimated independently of the size and composition of the treatment population ([6]). Note that there will never be an opportunity to estimate individual effects with confidence. Therefore we have to concentrate on the population average of gains from treatment. The most prominent evaluation parameter is the so-called mean effect of treatment on the treated:

$$E(\Delta \mid D = 1) = E(Y^1 \mid D = 1) - E(Y^0 \mid D = 1). \text{ ([7])}. \qquad (3)$$

The expected value of the treatment effect Δ is defined as the difference between the expected values of the outcome with and without training for those who actually participated in training. In the sense that this parameter focuses directly on actual training participants, it determines the realised gross gain from the training programme and can be compared with its costs, helping to decide whether the programme is a success or not (Heckman et al., 1997, 1998b; Heckman et al., 1999).

([5]) This is clearly different from asking whether there is an empirical association between training and the outcome (Lechner, 2000). See Holland (1986) for an extensive discussion of concepts of causality in statistics, econometrics and other fields.

([6]) Among other things SUTVA excludes cross-effects or general equilibrium effects. Its validity facilitates a manageable formal setup; nevertheless in practical applications it is frequently questionable whether it holds.

([7]) E is the expectation operator, in that case the expected value of the treatment effect Δ. $E(\Delta \mid D = 1)$ is the expected treatment for those who participated (| means conditional on)

Despite the fact that most evaluation research focuses on average outcomes, partly because most statistical techniques focus on mean effects, there is also a growing interest regarding effects of policy variables on distributional outcomes. Examples where distributional consequences matter for welfare analysis include subsidised training programmes (LaLonde, 1995) or minimum wages (DiNardo et al., 1996). Koenker and Bilias (2002) show that quantile regression methods can play a constructive role in the analysis of duration (survival) data too. They describe the link between quantile regression and the transformation model formulation of survival analysis, offering a more flexible analysis than conventional methods, for example if one is interested in the duration of employment/unemployment after the programme took place.

Nevertheless we will focus on the average treatment effect on the treated $E(\Delta \mid D = 1)$ in this paper. The second term on the right side in equation (3) is unobservable as it describes the hypothetical outcome without treatment for participants in a programme. If the condition:

$$E(Y^0 \mid D = 1) = E(Y^0 \mid D = 0) \tag{4}$$

holds, we can use the non-participants as an adequate control group. In other words we would take the mean outcome of non-participants as a proxy for the counterfactual outcome of participants. This identifying assumption is definitely valid in social experiments. The key concept here is the randomised assignment of individuals into treatment and control groups. Individuals who are eligible to participate, for example, in training are randomly assigned to a treatment group that participates in the programme and a control group that does not. This assignment mechanism is a process that is completely beyond employee or administrator control. If the sample size is sufficiently large, randomisation will generate a complete balancing of all relevant observable and unobservable characteristics across treatment and control groups. Therefore the comparability between experimental treatment and control groups is facilitated enormously. On average, the two groups do not systematically differ except for having participated in training. As a result, any observed difference in the outcomes of the groups after training is supposed to be solely induced by the programme itself, i.e. the impact of training is isolated and there should be no selection bias. Formally, random assignment ensures that the potential outcomes are independent of the assignment to the training programme. We write:

$$Y^1, Y^0 \perp\!\!\!\perp D \tag{5}$$

$\perp\!\!\!\perp$ denoting independence. When assignment to treatment is completely random it follows that:

$$E(Y^1 \mid D = 1) = E(Y^1 \mid D = 0),$$

and:

$$E(Y^0 \mid D = 1) = E(Y^0 \mid D = 0) \tag{6}$$

Therefore, treatment assignment becomes ignorable (Rubin, 1974) and we get an unbiased estimate of $E(\Delta)$, i.e. the randomly generated group of non-participants can be used as an adequate control group to estimate consistently the counterfactual term $E(Y^0 \mid D = 0)$ and thus the causal training effect $E(\Delta \mid D = 1)$. Although this approach seems to be very appealing in providing a simple solution to the fundamental evaluation problem, there are also problems associated with it. Besides relatively high costs and ethical issues concerning the use of experiments, in practice, a randomised experiment may suffer from similar problems that affect behavioural studies. Bijwaard and Ridder (2000) investigate the problem of non-compliance to the assigned intervention, that is when members of the treatment sample drop out of the programme and members of the control group participate. If the non-compliance is selective, i.e. correlated with the outcome variable, the difference of the average outcomes is a biased estimate of the effect of the intervention, and correction methods have to be applied too. Besides relatively high costs and ethical issues concerning the use of experiments, further methodological problems might arise, such as substitution or randomisation bias, which make the use of experiments questionable ([8]). For an extensive discussion of

([8]) A randomisation bias occurs when random assignment causes the types of persons participating in a programme to differ from the type that would participate in the programme as it normally operates, leading to an unrepresentative sample. We talk about a substitution bias when members of an experimental control group gain access to close substitutes for the experimental treatment (Heckman and Smith, 1995).

these topics the interested reader should refer to Burtless (1995), Burtless and Orr (1986) and Heckman and Smith (1995) ([9]).

More important for practical applications is the fact that, in most European countries, experiments are not conducted and researchers have to work with non-experimental data anyway. In non-experimental data, equation (4) will normally not hold:

$$E(Y^0 \mid D = 1) \neq E(Y^0 \mid D = 0), \tag{7}$$

The use of the non-participants as a control group might, therefore, lead to a selection bias. Heckman and Hotz (1989) point out that selection might occur on observable or unobservable characteristics. Good examples for observable characteristics are sociodemographic variables like qualification, age or gender of the individual, which are usually available in evaluation datasets. Unobservable characteristics might be the motivation or working habits of an individual. The aim of any observational evaluation approach is to ensure the comparability of treatment and control group by design; that is through a plausible identifying assumption. Taking account of observable factors might not be sufficient, if unobservable factors invalidate the comparison, for example when more motivated workers have a higher employment probability and are also more likely to participate in a training programme (Schmidt, 1999) ([10]).

In the following subsections we will present four different evaluation approaches. Each approach invokes different identifying assumptions to construct the required counterfactual outcome. Therefore each estimator is only consistent in a certain restrictive environment. As Heckman et al. (1999) note, all estimators would identify the same parameter only if there is no selection bias at all.

2.2.2. Before-after estimator

The most obvious and still widely used evaluation strategy is the before-after estimator (BAE). It compares the outcome of participants before training took place with their outcome after training. The basic idea is that the observable outcome in the pre-training period t' represents a valid description of the unobservable counterfactual outcome of the participants without training in the post-training period t. The central identifying assumption of the BAE can be stated as:

$$E(Y_{t'}^0 \mid D=1) = E(Y_t^0 \mid D=1). \tag{8}$$

Given the identifying assumption in (8), the following estimator of the mean treatment effect on the treated can be derived:

$$\Delta^{BAE} = E[(Y_t^1 \mid D=1) - (Y_{t'}^0 \mid D=1)]. \tag{9}$$

Heckman et al. (1999) note that conditioning on observable characteristic X makes it more likely that assumption (8) will hold. If the distribution of X characteristics is different between the treatment and the control group, conditioning on X may eliminate systematic differences in the outcomes ([11]).

The validity of (9) depends on a set of implicit assumptions. First, the pre-exposure potential outcome without training should not be affected by training. This may be invalid if individuals have to behave in a certain way in order to get into the programme or behave differently in anticipation of future training participation. Second, no time-variant effects should influence the potential outcomes from one period to the other. If there are changes in the overall state of the economy or changes in the lifecycle position of a cohort of participants, assumption (8) may be violated (Heckman et al., 1999).

A good example where this might be the case is Ashenfelter's dip. That is a situation where shortly

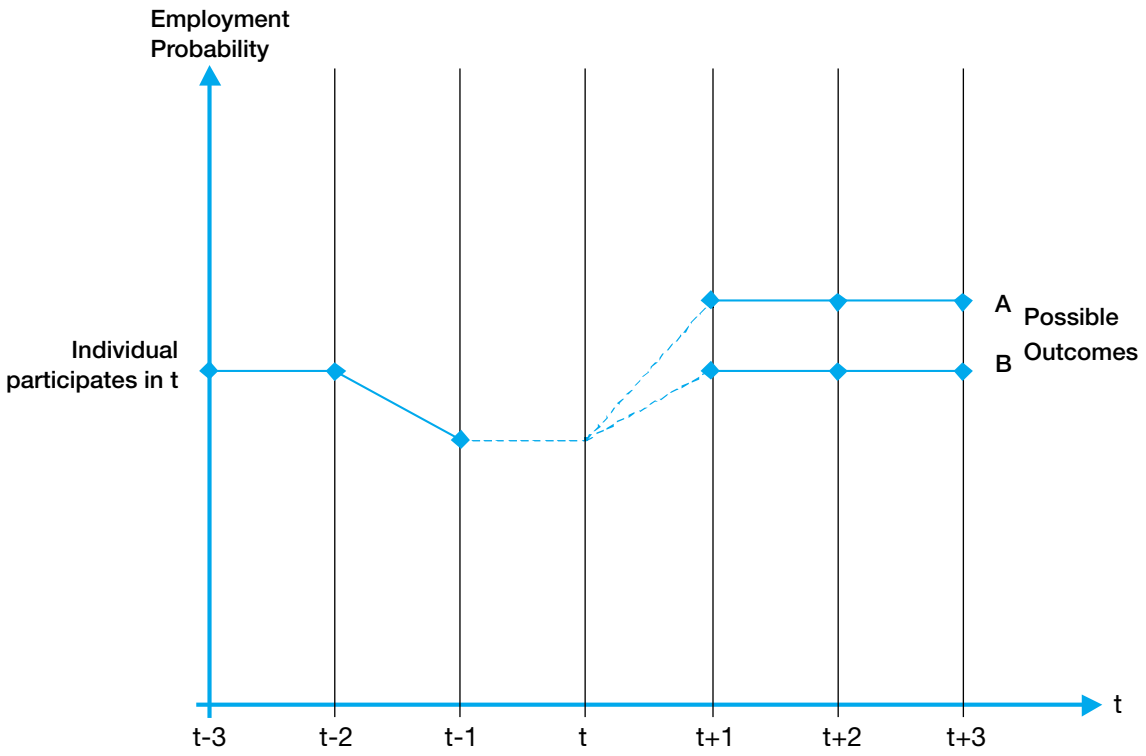
([9]) As Smith (2000) notes, social experiments have become the method of choice in the US. The most famous among them is the National Job Training Partnership Act which had a major influence regarding the view on non-experimental studies. In Europe, however, social experiments have not received similar acceptance, although recently some test experiments have been conducted. The most important one is the Restart experiment in Britain (Dolton and O'Neill, 1996).

([10]) The distinction between observable and observed factors depends on the dataset. Whereas some factors like the motivation of an individual are hard to measure and therefore usually not observable, other factors like education are in general observable but might not be observed in the dataset at hand.

([11]) On the other hand, if the difference in the treatment and the control group are due to unobservable characteristics, conditioning may accentuate rather than eliminate the differences in the no-programme state between the both groups (Heckman et al., 1999).

before the participation in an ALMP programme, the employment situation of the future participants deteriorates. Ashenfelter (1978) found this dip while evaluating the effects of treatment on earnings, but later research demonstrated that this dip can be observed on employment probabilities for participants too. If the dip is transitory and only experienced by participants, assumption (8) will not hold. By contrast, permanent dips are not problematic, as they affect employment probability before and after treatment in the same way. Figure 3 illustrates Ashenfelter's dip.

Figure 3: **Ashenfelter's dip**



We assume that an individual participates in a training programme in period t and experiences a transitory dip in the pre-training period $t-1$; for example the employment probability may be lowered as the individual is not actively seeking work because of the intended participation in a programme in the next period. There are no economy-wide, time-varying effects that influence the employment probability of the individual. After the programme takes place, employment probability might assume many values. For the sake of simplicity we consider two cases. Case A assumes that there has been a positive effect on employment probability (vertical difference between A and B). As the BAE compares the employment probability of the individual in period $t+1$ and $t-1$ (vertical difference between point A and C) it would overestimate the true treatment effect, because it would attribute the restoration of the transitory dip completely to the programme. In a second example, we assume that there has been no treatment effect at all (B). In period $t+1$ the employment probability is restored to its original value before the dip took place in $t-1$. Again the BAE attributes this restoration completely to the programme and would estimate a positive treatment effect (vertical difference between point B and C). Clearly the problem of Ashenfelter's dip could be avoided, if a time period is chosen as a reference level before the dip took place, for example period $t-2$. But as we will not know *a-priori* in an

empirical application when the dip starts, the question arises of which period to choose.

A major advantage of the BAE is that it does not require information on non-participants. All that is needed is longitudinal data on outcomes on participants before and after the programme took place ([12]). As the employment status of the participants is usually known in the programmes or even prerequisite for participation, the BAE does not impose any major problems regarding the data availability, which might explain why it is still widely used ([13]).

2.2.3. Cross-section estimator

The basic idea of the cross-section estimator is to compare the outcome of participants after the programme took place with the outcome of non-participants at the same time period. Instead of comparing participants at two different time periods, the cross-section estimator compares participants and non-participants at the same time (after the programme took place), i.e. the population average of the observed outcome of non-participants replaces the population average of the unobservable outcome of participants. This is useful if no longitudinal information on participants is available or macroeconomic conditions shift substantially over time (Schmidt, 1999). The identifying assumption of the cross-section estimator can be stated formally as:

$$E(Y_t^0 \mid D=1) = E(Y_t^0 \mid D=0). \tag{10}$$

That is, those who participate in the programme have, on average, the same no-treatment outcome as those who do not participate. If this assumption is valid, the following estimator of the mean true treatment effect can be derived:

$$\Delta^{CSE} = E[(Y_t^1 \mid D=1) - (Y_t^0 \mid D=0)]. \tag{11}$$

It is worth noting that conditioning on observable characteristics makes it more likely that this assumption will hold. If the distribution of X characteristics is different between the treatment and the control group, conditioning on X may eliminate systematic differences in the outcomes ([14]). To give a straightforward example, let us assume that X represents the qualification level of an individual. For the sake of simplicity, we assume that X might only take two values (1 for high-skilled and 0 for low-skilled workers). Therefore, conditioning on X results in estimating the treatment effects separately for both skill groups and is intuitively appealing. The identifying assumption is then:

$$E(Y_t^0 \mid X, D=1) = E(Y_t^0 \mid X, D=0). \tag{12}$$

The first approach in equation (11) can be seen as a 'naive' estimator, because it just compares the results for the whole group, whereas the second takes into account observed differences in individual's characteristics, such as different skill levels. The resulting estimator can be written as:

$$\Delta^{CSE|X} = E[(Y_t^1 \mid X, D=1) - (Y_t^0 \mid X, D=0)]. \tag{13}$$

Schmidt (1999) notes, that for assumption (12) to be valid, selection into treatment has to be statistically independent of its effects given X (exogenous selection), that is, no unobservable factor should lead individual workers to participate. A good example where this is violated might be the case if motivation plays a role in determining the desire to participate and the outcomes without treatment. Then we have, even in the absence of any treatment effect, a higher average outcome in the participating group compared to the non-participating group. Ashenfelter's dip is not problematic for the cross section estimator as we compare only participants and non-participants after the programme took place. Moreover, as long as economy-wide shocks and individual lifecycle patterns operate identically for the treatment and the control group, the cross-section estimator is not vulnerable to the problems that plague the BAE (Heckman et al., 1999).

2.2.4. Matching estimator

The matching approach originated in statistical literature and shows a close link to the experi-

([12]) The BAE might also work with repeated cross-sectional data from the same population, not necessarily containing information on the same individuals. See Heckman and Robb (1985) or Heckman et al. (1999) for details.

([13]) Note that the labour-market status (unemployed, part-time or low-skilled employed, etc.) is very often one of the entry conditions for ALMP programmes.

([14]) On the other hand, if the differences in the treatment and the control group are due to unobservable characteristics, conditioning may accentuate rather than eliminate the differences in the no-programme state between both groups (Heckman et al., 1999).

mental context ([15]). The basic idea underlying the matching approach is to find in a large group of non-participants those individuals who are similar to the participants in all relevant pre-training characteristics. That being done, the differences in the outcomes between the well selected and thus adequate control group and the trainees can then be attributed to the programme. Matching does not need to rely on functional form or distributional assumptions, as its nature is non-parametric (Augurzky, 2000).

Matching is first of all plagued by the same problem as all non-experimental estimators, which means that assumption (4) cannot be expected to hold when treatment assignment is not random. However, following Rubin (1977), treatment assignment may be randomly given a set of covariates. The construction of a valid control group via matching is based on the identifying assumption that conditional on all relevant pre-training covariates Z, the potential outcomes (Y^1, Y^0) are independent of the assignment to training ([16]). This so-called conditional independence assumption (CIA) can be written formally as:

$$Y^1, Y^0 \perp\!\!\!\perp D \mid (Z). \quad ([17]) \qquad (14)$$

If assumption (14) is fulfilled we get:

$$E(Y^0 \mid Z, D = 1) = E(Y^0 \mid Z, D = 0) = E(Y^0 \mid Z). \qquad (15)$$

Similar to randomisation in a classical experiment, the role of matching is to balance the distributions of all relevant pre-treatment characteristics in the treatment and control group, and thus to achieve independence between potential outcomes and the assignment to treatment, resulting in an unbiased estimate. The exact matching estimator can be written as:

$$\Delta^{MAT} = E[(Y_t^1 \mid Z, D=1) - (Y_t^0 \mid Z, D=0)]. \qquad (16)$$

Conditioning on all relevant covariates is, however, limited in case of a high dimensional vector Z. For instance, if Z contains n covariates which are all dichotomous, the number of possible matches will be 2^n. In this case cell matching, that is exact matching on Z, is not possible since an increase in the number of variables increases the number of matching cells exponentially. To deal with this dimensionality problem, Rosenbaum and Rubin (1983) suggest the use of balancing scores $b(Z)$, i.e. functions of the relevant observed covariates Z such that the conditional distribution of Z given $b(Z)$ is independent of the assignment to treatment, that is $Z \perp\!\!\!\perp D \mid b(Z)$ holds.

For trainees and non-trainees with the same balancing score, the distributions of the covariates Z are the same, i.e. they are balanced across the groups. Moreover Rosenbaum and Rubin (1983) show that if the treatment assignment is strongly ignorable ([18]) when Z is given, it is also strongly ignorable given any balancing score. The propensity score, i.e. the probability of participating in a programme is one possible balancing score. It summarises the information of the observed covariates into a single index function.

Rosenbaum and Rubin (1983) show how the conditional independence assumption extends to the use of the propensity score so that:

$$Y^0 \perp\!\!\!\perp D \mid P(Z) \qquad (17)$$

Therefore we get:

$$E(Y^0 \mid P(Z), D = 1) = E(Y^0 \mid P(Z), D = 0)\ E(Y^0 \mid P(Z)), \qquad (18)$$

which allows us to rewrite the crucial term in the average treatment effect (3) as:

$$E(Y^0 \mid D = 1) = E_{p(z)}[(Y^0 \mid P(Z), D = 0) \mid D = 1]. \qquad (19)$$

Hujer and Wellner (2000b) note that the outer expectation is taken over the distribution of the propensity score in the treated population. The major advantage of the identifying assumption (17) is that it transforms the estimation problem into a much easier task since one has to condition on a univariate scale, i.e. on the propensity

([15]) See Rubin (1974, 1977, 1979), Rosenbaum and Rubin (1983, 1985a, 1985b) or Lechner (1998).

([16]) If we say relevant we mean all those covariates that influence the assignment to treatment as well as the potential outcomes. In contrast to the cross-section estimator, the matching procedure can also use information from the pre-treatment period, such as employment status or other time-varying covariates. To make this difference clear, we denote the covariates by (Z).

([17]) For the purpose of estimating the mean effect of treatment on the treated, the assumption of conditional independence of Y^0 is sufficient because we like to infer estimates of Y^0 for persons with $D = 1$ from data on persons with $D = 0$ (Heckman et al., 1997).

([18]) Strongly ignorable means that assumption (16) holds and: $0<P(Z)\equiv P(D = 1 \mid Z)<1$. The latter ensures that there are no characteristics in Z for which the propensity score is zero or one. Proofs go beyond the scope of this work and can be found in Rosenbaum and Rubin (1983).

score, only. When $P(Z)$ is known, the problem of dimensionality can be eliminated. The evaluation of the counterfactual term via matching on the basis of the group of non-participants then only requires to pair participants with non-participants which have the same propensity score. This ensures a balanced distribution of Z across both groups. Unfortunately, $P(Z)$ will not be known *a-priori* so it has to be replaced by an estimate. This can be achieved by any number of standard probability models, for example a probit model.

The empirical power of matching to reduce the problem of selection bias relies crucially on the quality of the estimate of the propensity score, on the one hand, and on the existence of comparison persons that have equal propensity scores as the treated persons. If the latter is not ensured we run the risk of incomplete matching with biased estimates ([19]). Several procedures for matching on the propensity score have been suggested and will be discussed briefly; a good overview can be found in Heckman et al. (1998a) and Smith and Todd (2000). To introduce them a more general notation is needed. We estimate the effect of treatment for each observation in the treatment group, by contrasting his/her outcome with treatment with a weighted average of control group observations j in the following way:

$$Y_i^1 - \sum_{j\in\{D=0\}} W_{N_0N_1}(i,j)Y_j^0, \tag{20}$$

where N_0 is the number of observations in the control group and N_1 is the number of observations in the treatment group and is a weighting function which can take different forms. Matching estimators differ in the weights attached to the members of the comparison group (Heckman et al., 1998a). Nearest neighbour (NN) matching sets:

$$C(P_i) = \min_j \left\| P_i - P_j \right\|, j \in N_0. \tag{21}$$

Doing so, the non-participant with the value of that is closest to is selected as the match, therefore $W_{N_0\,N_1}(i,j)=1$ for this unit and $W_{N_0\,N_1}(i,j)=0$ otherwise ([20]). Several variants of nearest neighbour matching are proposed, for example nearest neighbour matching 'with' and 'without replacement'. In the former case, a non-participating individual can be used more than once as a match, whereas in the latter case it is considered only once. Use of more than one nearest neighbour (over-sampling) is also suggested. Nearest neighbour matching runs the risk of bad matches, if the closest neighbour is far away.

This can be avoided by imposing a tolerance on the maximum distance $\|P_i\text{-}P_j\|$ allowed. This form of matching, caliper matching (Cochrane and Rubin, 1973), imposes the condition:

$$\left\| P_i - P_j \right\| < \varepsilon, j \in N_0 \quad . \tag{22}$$

where is a pre-specified level of tolerance.

Kernel matching (KM) is a non-parametric matching estimator that uses all individuals in the control group to construct a match for each programme participant. KM defines the weighting function as:

$$W_{N_0}(i,j) = \frac{K_{ij}}{\sum_{k\in(D=0)} K_{ik}}, \tag{23}$$

where $K_{ik} = K((P_i\text{-}P_k)/h)$ is a kernel that down-weights distant observations from P_i and h is a bandwidth parameter (Heckman et al., 1998a).

The matching estimator is very data-demanding in the sense that we need information for participants and non-participants before and after the programme took place. In the case of exact matching, a 'rich' dataset is needed to ensure that we find comparable individuals in the control group for every combination of observable characteristics. Even if we do not use exact matching but matching over the propensity score, a rich dataset is needed. In that case the quality of the score depends on our ability to account for all relevant covariates that determine the participation decision. Problems arise, if a programme is compulsory, for example if all the unemployed have to participate in a training programme at a certain point in their unemployment spell. In this case it might get difficult to find a suitable control group,

([19]) Matching was much discussed in recent econometric literature. Heckman and his colleagues reconsidered and further developed the identifying assumptions of matching stated by Rubin (1977) and Rosenbaum and Rubin (1983). It turns out that the new identifying assumptions are weaker compared to the original statements which brings along some advantages. Presenting these ideas goes beyond the scope of this work. The interested reader should refer to Heckman et al. (1996, 1997, 1998a and b) and Heckman and Smith (1995).

([20]) Exact matching imposes an even stronger condition, where only non-participants with exactly the same propensity score or the same realisation of characteristics X are considered as matches.

because all the unemployed will be in the programme at some time. However, it is still possible to evaluate the optimal timing of a programme, for example after three, six or 12 months of unemployment.

2.2.5. Difference-in-differences estimator

It has been claimed that controlling for selection on observable characteristics may not be sufficient since remaining unobservable differences may still lead to a biased estimation of treatment effects. These differences may arise from differences in the benefits which individuals expect from participation in a treatment, which might influence their decision to participate. Furthermore, some groups might exhibit bad labour-market prospects or differences in motivation. These features are unobservable to a researcher and might cause a selection bias.

To account for selection on unobservables, Heckman et al. (1999) suggest econometric selection models and difference-in-differences (DID) estimators. The DID estimator requires access to longitudinal data and can be seen as an extension to the classical BAE. Whereas the BAE compares the outcomes of participants after they participate in the programme with their outcomes before they participate, the DID estimator eliminates common time trends by subtracting the before-after change in non-participant outcomes from the before-after change for participant outcomes. The simplest application of the method does not condition on X and forms simple averages over the group of participants and non-participants. Changes in the outcome variable Y for the treated individuals are contrasted with the corresponding changes for non-treated individuals (Heckman et al, 1998a):

$$\Delta^{DID} = [Y_t^1 - Y_{t'}^0 | D=1] - [Y_t^0 - Y_{t'}^0 | D=0]. \tag{24}$$

The DID estimator is based on the assumption of time-invariant linear selection effects. The critical identifying assumption of this method is that the biases are the same, on average, in different time periods before and after the period of participation in the programme, so that differencing the differences between participants and non-participants eliminates the bias (Heckman et al., 1998a).

To make this point clear, we assume that we want to evaluate a training programme that aims to improve the employment prospects of individuals. And let us further assume that the participants have a higher motivation and, owing to this higher motivation, the outcome for participants is on average 5% higher than for non-participants. If we additionally presume that this bias remains constant over time, we can estimate the effect of a programme by comparing the pre- and post-programme differences of the participants and non-participants. Let us say that the participants had an average outcome of 50% in the period before the programme took place, that the treatment effect is 10% and, owing to a cyclical upswing, the outcome is increased by 15%. Therefore the participants have an average outcome of 75% in the period after the programme took place. The non-participants have an average outcome of 45% in the first period (5% less than the participants because of the motivational selection effect) and 60% in the second period. Comparing the average outcomes for participants in period two (75%-60%=15%) is misleading since the selection effect due to unobservable characteristics is not taken into account. The DID estimator, in contrast, leads to the correct result of 10% ([21]). Through the differencing procedure, the cyclical upswing is correctly not attributed to the programme and neither is the motivational effect.

This example can be formally explained by denoting the outcome for an individual at time t as:

$$Y_{it} = \alpha_{it} + D_{it} \cdot Y_{it}^1 + (1 - D_{it}) \cdot Y_{it}^0, \tag{25}$$

where α_{it} captures the effects of selection on unobservables. The validity of the DID estimator relies crucially on the assumption:

$$\alpha_{it} = \alpha_{it'}. \tag{26}$$

Only if the selection effect is time-invariant will it be cancelled out and an unbiased estimate result. The differencing leads to:

$$\begin{aligned} Y_{it} - Y_{it'} = {} & [D_{it} \cdot Y_{it}^1 + (1 - D_{it}) \cdot Y_{it}^0] - \\ & [D_{it'} \cdot Y_{it'}^1 + (1 - D_{it'}) \cdot Y_{it'}^0) + \\ & [\alpha_{it} - \alpha_{it'}]. \end{aligned} \tag{27}$$

([21]) Difference for participants: (75%-50%) = 25%; difference for non-participants: (60%-45%) = 15%; DID estimator: 25%-15% = 10%.

If (26) is fulfilled, the last term in the expression can be cancelled out, leading to an unbiased estimate. Compared to the method of matching, the DID approach does not require that the bias vanishes for any 'matched' individuals, but only that it remains constant (Heckman et al., 1998a).

If we condition the DID approach on observable characteristics X, the new estimator is given by:

$$\Delta^{DID|X} = [Y_t^1 - Y_{t'}^0 \mid X, D=1] - [Y_t^0 - Y_{t'}^0 \mid X, D=0]. \quad (28)$$

The identifying assumption of this method is:

$$E(Y_t^1 - Y_{t'}^0 \mid X, D=1) = E(Y_t^0 - Y_{t'}^0 \mid X, D=0). \quad (29)$$

Ashenfelter's dip is definitely a problem for the DID estimator. If the dip is transitory and is eventually restored even in the absence of participation in the programme, the bias will not average out. Therefore Bergmann et al. (2000) apply a combination of the matching- and the DID estimator in a recent paper, by implementing a 'conditional DID estimator', where conditional means that treatment and control groups are already partly comparable regarding their observable characteristics. Kluve et al. (1999) suggest using the pre-treatment (labour market) histories of the individuals as an important variable in the matching process, so that only individuals with identical pre-treatment histories are compared. Basically the conditional DID estimator extends the one suggested from Heckman et al. (1998a) by adding a longitudinal dimension. The basic question in this case is deciding the appropriate time span to be considered.

2.2.6. Duration models

An alternative approach to modelling treatment effects of ALMP is the application of duration models. The basic idea here is to include the duration of (un)employment and the duration of the programmes when estimating effects. In a bivariate duration model the durations T_u and T_p measure the duration of unemployment until entry into employment and the duration until entry into ALMP, respectively. T_u and T_p are random variables, t_u and t_p denote their realisations. We are interested in the causal effect of participation in ALMP on exit from unemployment, that is the effect of the realisation of T_p on the distribution of T_u. This implies that the causal effect is captured by the effect of t_p on the hazard rate for job finding: $\theta_u(t|T_p,x,v)$ for $t>t_p$.

The transition rate from unemployment to employment at time conditional on and can be specified by a mixed proportional hazard model as follows:

$$\theta_u(t \mid x, t_p, d_z, v_u) = \lambda_u(t) \cdot exp(x'\beta_u + \delta(t \mid t_p, x) \cdot \cdot I(t > t_p) + \mu_{uz} \cdot d_z + v_u), \quad (30)$$

where $I(\cdot)$ denotes the indicator function, which is 1 if its argument is true and 0 otherwise. The function $\lambda_u(t)$ is called the 'baseline hazard' individual duration dependence. δ measures the effect of the participation in ALMP on the transition rate from unemployment to employment, x is a vector of explanatory variables, μ_{uz} measures whether there is any benefit exhaustion effect and the term d_u represents unobserved heterogeneity. In a similar way, the hazard rate to programme p at time t conditional on x and d_z can be specified in the following equation:

$$\theta_p(t \mid x, d_z) = \lambda_u(t) \cdot \exp\,(x'\beta_p + \mu_{pz} d_z + v_p). \quad (31)$$

In the equation for θ_u and θ_p, unobserved heterogeneity is allowed to affect the transitions to both a job and to a programme.

'If the unobserved characteristics have a negative effect on the job-finding rate and a positive effect on the transition to a programme, then conditional on the observed characteristics and the elapsed duration of unemployment, the average quality of workers in a programme is lower than the average quality of workers who do not enter a programme. Then, if we would simply compare the transition rates to regular jobs of both groups we would compare workers with unfavourable characteristics and programme participation with workers with more favourable characteristics and non-participation. Therefore, we would underestimate the true effect of participating in a programme. The opposite effect is also possible. One could imagine that the people in control of the programmes want their programmes to be a success. Therefore they prefer workers with good characteristics to flow into their programme. This would imply that there is a positive correlation between unobserved heterogeneity components in both transition rates. Then we would overestimate the treatment effect of programmes.' (Lalive et al., 2000).

The authors of the empirical studies assume for the joint distribution of the unobserved characteristics $G(v_u, v_p)$ a multivariate discrete distribution using a multinomial logit specification (Heckman and Singer, 1984).

2.2.7. An intuitive example

Table 1 summarises the different evaluation estimators introduced in the previous subsections. It presents short descriptions of the various approaches and discusses their advantages and disadvantages using a simple numerical example ([22]). The goal of this example is not to show the (dis-) advantages of each evaluation estimator in detail, but to give some guidance on how they perform in a certain economic context. For the sake of simplicity it is assumed that only two types of workers (highly-skilled (X=1) and low-skilled (X=0)) exist. It is further assumed that the interest lies in evaluating a training programme that is intended to improve the skills of the workers and therefore enhance their employment prospects. To make the importance of heterogeneity clear it is furthermore assumed that the programme works better for highly-skilled workers. It can be thought of as a very specialised training measure to improve, for example, management skills and is therefore only taken up by a relatively small group of highly-skilled workers.

Table 1: **Comparison of the different evaluation estimators**

Estimator	Description	(Dis-) Advantages
BAE	Compares the outcome for participants before and after the programme took place.	+ easy to implement + low data requirements (no information for non-participants needed) - economic changes from the before to the after period might be falsely attributed to the programme - Ashenfelter's dip
CSE	Compares the outcome for participants in the period after the programme took place, with the outcome for non-participants in the same period.	+ economy-wide changes are not attributed to the programme + Ashenfelter's dip is not a problem - needs data for participants and non-participants for the period after the programme took place
Matching estimator	Compares the outcome of participants in the period after the programme took place, with the outcome of matched (statistical twins) non-participants in the same period.	+ takes account of selection on observable characteristics + Ashenfelter's dip is not a problem - needs data for participants and non-participants before and after the programme took place (e.g. labour-market history) - unobservable characteristics
DID estimator	Compares the before-after change for participant outcomes with the before-after change for non-participant outcomes.	+ takes account of selection on unobservable characteristics - needs data for participants and non-participants before and after the programme took place - Ashenfelter's dip
Duration models	The effects are measured by the coefficient of an explanatory dummy variable (participation yes/no) in a bivariate model framework.	+ takes into account spells (that is duration of (un)employment and policy measures) - selection problem is solved only implicitly

([22]) See Schmidt (1999) for a similar example.

Table 2 displays the labour-market outcomes of the workers who received (D=1) and did not receive (D=0) treatment. Columns 2 and 3 show the labour-market outcomes for non-participants before ($Y_{t'}$) and after (Y_t) the programme took place. Column 5 contains the pre-programme outcome of participants ($Y_{t'}$), whereas column 7 contains the post-programme outcome of participants (Y_t+Δ). All these outcomes are actually observed. Column 6 contains the counterfactual outcome under no treatment for participants. Clearly, this outcome is never observed in an empirical study but allows us to answer the question 'What would have happened to the participants if they had not participated?'. The programme impact can be calculated by comparing the actual outcome of the participants after the programme took place (Y_t+Δ) with their counterfactual outcome under no treatment (Y_t). Again it is important to note that a researcher will not be able to do so since column 6 is unobservable. This unobservable outcome is only introduced as a reference level to illustrate the functioning of the different estimators.

Table 2: **Labour-market outcomes for participants and non-participants**

Non-Participants (D = 0)			Participants (D = 1)			
Low-skilled workers (X = 0)						
Worker	**$Y_{t'}$**	**Y_t**	**Worker**	**$Y_{t'}$**	**Y_t**	**Y_t+Delta**
1	0	1	11	0	0	1
2	0	0	12	0	0	0
3	0	0	13	0	0	0
4	0	0	14	0	1	1
5	0	0	15	0	1	1
6	1	1	16	0	0	0
7	1	1	17	0	1	1
8	1	1	18	1	1	1
9	1	1	19	1	1	1
10	1	1	20	1	0	1
Mean	0.5	0.6		0.3	0.5	0.7
Highly-skilled workers (X = 1)						
Worker	**$Y_{t'}$**	**Y_t**	**Worker**	**$Y_{t'}$**	**Y_t**	**Y_t+Delta**
21	0	0	31	0	0	1
22	0	1	32	0	1	1
23	0	0	33	0	0	1
24	1	1	34	1	1	1
25	1	1	35	1	1	1
26	1	1				
27	1	1				
28	1	1				
29	1	1				
30	1	1				
Mean	0.7	0.8		0.4	0.6	1

The two skill groups will be discussed separately.

In the group of low-skilled workers there are 10 non-participants (workers 1-10) and 10 participants (workers 11-20). 50 % of the non-participants are employed in both periods and for one non-participant the labour-market situation improves over time (worker 1). In the group of participants only 30 % are employed in the first period. For three workers (workers 14, 15 and 17) the situation would have improved even in absence of the training programme (Y_t), whereas for one worker (worker 20) it would have worsened. The programme impact can be estimated by comparing columns 7 and 6. The programme improves the employment prospects for two workers (workers 11 and 20) and therefore the effect is 0.2. The following is how a researcher would calculate the programme impacts with some of the presented estimators ([23]).

The BAE invokes the identifying assumption that the pre-training outcome of the participants represents a valid description of their unobservable counterfactual outcome in the post-training period (equation 7). Therefore the estimator is calculated by comparing the outcome in column 7 (0.7) with the outcome in column 5 (0.3) and leads to a result of 0.4. Clearly, this estimator attributes all improvement in the employment situation between the two periods to the programme. But since, in our example, the situation for two workers would have improved anyway, for example due to an overall economic improvement, the BAE overestimates the real impact.

The implementation of the cross-section estimator requires that the population average of the observed outcome of the non-participants is equal to the population average of the unobservable outcome of participants (equation 10). Therefore the estimator is calculated by comparing the actual outcome for participants (workers 11-20) in period t (average: 0.7, column 7) with the actual outcome for non-participants (workers 1-10) in the same period (average: 0.6, column 3). The estimated impact would be 0.1. This is due to the fact that the constructed labour-market situation for non-participants in t (average: 0.6, column 3) is very similar to the counterfactual labour-market situation for participants (average: 0.5, column 6).

A matching estimator replaces the counterfactual outcome of the participants with the population average of a matched control group. In this example there is only one available matching variable, namely the labour-market situation before training. Estimation is then simply done by comparing the outcome in t of workers 11-17 (average: 0.57, column 7) with the outcome of workers 1-5 (average: 0.2, column 3) on the one hand and between workers 18-20 (average: 1.0, column 7) and 6-10 (average: 1.0, column 3) on the other. The estimated impact is then a weighted average (by the number of participants) between both groups. The effect in the first group (seven participants unemployed before training) is 0.37, the effect in the second group is 0 (three workers employed before training). The weighted average is, therefore, (0.37 x 7/10) = 0.26.

Finally, a DID estimator eliminates common time trends by subtracting the before-after change in non-participant outcomes from the before-after change for participant outcomes. In our example we form simple averages over the group of participants and non-participants and contrast changes in the labour-market situation for the treated individuals with the corresponding changes for non-treated individuals. The estimated impact in the example is then (0.7 – 0.3) – (0.6 – 0.5) = 0.3. Since the increase in potential non-treatment outcome is higher for participants (0.2 vs. 0.1), the DID overestimates the true treatment effect.

The same calculations can be done for the group of highly-skilled workers. This group contains 10 non-participants (workers 21-30) and five participants (workers 31-35). 60 % of the participants are unemployed before training and only one (worker 32) would have experienced an improvement in his labour-market situation without training. The programme effect in the group of highly-skilled workers is 0.4. The

([23]) For the sake of simplicity this numerical example has been kept on a basic methodological level. Therefore, more complex estimators like the conditional DID or the duration models could not be illustrated and the focus lies on the BAE, CSE, DID and the exact matching estimator.

evaluation estimators can be calculated analogously as described above and the results can be found in column 3 of Table 3. The matching estimator is very successful this time as it exactly calculates the true impact, whereas the BAE and DID overestimate the programme impact and the cross-section estimator underestimates it.

Table 3: **Estimates of the programme effects**

	Low-Skilled	Highly-Skilled	Average
Impact	0.20	0.40	0.27
BAE	0.40	0.60	0.47
CSE	0.10	0.20	0.13
Matching	0.26	0.40	0.31
DID	0.30	0.50	0.37

The last column is a weighted average of the programme effects in the two groups (weighted according to the number of participants). This is simply to show how important it is to account for heterogeneity in the estimation of programme impacts. Due to data limitations previous evaluation studies had to pool heterogeneous programmes and/or heterogeneous individuals and estimate one composite treatment effect. That this is a misleading approach becomes clear in the given example. The weighted impact for low- and highly-skilled workers is 0.27. Since the true impact for highly-skilled (low-skilled) workers is 0.4 (0.2), disregarding the heterogeneity of the worker leads to an underestimation (overestimation) of the true effect. Clearly this is an unnecessary source of bias which has to be avoided. Therefore, heterogeneity with respect to individual characteristics, regional aspects and, in particular, different programmes should be modelled in the evaluation.

2.3. Macroeconometric evaluation

In this chapter we will discuss the impact of ALMP, not on particular individuals but on aggregate economic variables. Instead of looking at the effect on individual performance we would like to know if the ALMP represent a net gain for the whole economy. If the total number of jobs is not affected by labour-market policies, the effects will be distributional only. This might be desirable, for example, if work is shifted from the old to the young, but can hardly justify the substantial fiscal costs of the ALMP.

The need for macroeconometric evaluation results from the need to estimate whether a positive effect on the microeconomic level is also positive on the macroeconomic level. This is a question of spillover effects, i.e. if the effects on the participants is counteracted or enforced by the effects on the non-participants. ALMP is often thought to have a positive effect on the participants but negative effects on the non-participants. Important effects in this context are the so-called dead-weight losses and substitution effects that have received substantial attention in the literature (Layard et al., 1991 or OECD, 1993), mainly in the context of job creation schemes. If the programme outcome is no different from that in its absence, we talk about a dead-weight loss. A common example is the hiring from the target group that would have occurred with or without the programme. If a firm hires a subsidised worker instead of an unsubsidised worker we talk about a substitution effect. The net short-term employment effect in this case is zero. Calmfors (1994) defines 'the substitution effect as the extent to which jobs created for a certain category of workers simply replace jobs for other categories, because relative wage costs are changed.' Such effects are likely in the case of subsidies for private-sector work. There is always a risk that employers hold back ordinary job creation in order to be able to take advantage of the subsidies. In order to minimise this danger, a principle of additionality may be imposed.

Another problem might be that active labour-market programmes may crowd out

regular employment. This can be seen as a generalisation of the so-called displacement effect. This effect typically refers to displacement in the product market; for example, firms with subsidised workers may increase output, while output may be reduced in firms that do not have subsidised workers. Clearly these effects have to be taken into account before making statements about the net effect of ALMP.

To derive the empirical model for a macroeconometric evaluation, the theoretical analysis of ALMP becomes crucial. This is because the specification of the empirical model matters in addition to the outcome (i.e. dependent) variable. Calmfors (1994) shows how a theoretical framework can be developed to allow an analysis of ALMP. He presents the 'revised Layard-Nickell model' as a basic framework for the analysis of the effects of ALMP on a number of economic variables and processes that influence aggregate employment and unemployment rates. Therefore, such a framework can be used to identify the different effects of ALMP on the whole economy. Following the discussion of Calmfors (1994), important effects of ALMP are considered below.

The first effects relate to the matching process. ALMP can improve matching between workers and jobs through several channels. First, ALMP can improve the active search behaviour of participants. Second, ALMP can speed up the matching process by adjusting the structure of labour supply to demand. Here we primarily think of retraining programmes that adapt the skills of the unemployed to the requirements of vacant jobs. Third, participation in an ALMP programme can serve as a substitute for work experience that reduces the employer's uncertainty about the employability of the job applicant.

If ALMP can improve the matching process, what are the effects on regular employment or wages? First, an improved matching process means that for a given stock of vacancies there is a greater inflow into employment ([24]). Furthermore, the improved matching process reduces the average duration that a vacancy remains unfilled. Since this reduces the costs of maintaining a vacancy, firms provide more vacancies, which is equivalent to an increase in labour demand. The same effect also improves the firm's position in a wage bargaining process, since the firm can expect to fill a vacancy much quicker if a worker was laid off. Therefore, improved matching also leads to a reduction in wages.

ALMP programmes are also expected to have negative effects on the matching process, i.e. so-called locking-in effects. If a participation in an ALMP programme is associated with full time engagement, there might be insufficient time for actively searching for a regular job. In this case, the search effectiveness of participants is lower than the search effectiveness of the unemployed (Holmlund and Linden, 1993). Since this locking-in effect vanishes the moment the programme expires, do the positive effects on the search effectiveness persist after the participation has ended?

A second potential effect of ALMP is reduced welfare losses for the unemployed. If an ALMP programme increases re-employment probability or if the compensation level is higher than the unemployment benefits, the ALMP programme increases the expected welfare of the unemployed. This is caused by the fact that an unemployed person faces a positive probability of being placed in a programme with a consequent rise in the expected income. In the context of a wage bargaining process this is the same as an increase in fallback income, i.e. the income that is obtained if the bargaining fails and the worker becomes unemployed (Layard et al., 1991). The rise in the fallback income leads to a higher outcome for wages, since the position of the workers in the bargaining process is improved. This effect of ALMP on wage pressure is not avoidable, since every improvement of the situation of the unemployed is connected with a reduction in the welfare losses.

There is also a *competition effect*, in that ALMP (especially training programmes) are expected to improve the skills of participants and so make them more competitive. This means that not only is there improved competition between the unemployed, but also improved competition between the employed and the unemployed. Additionally, ALMP can affect competition if it stimulates participants to search more actively (i.e. to counteract the discouraged worker effect) or if it helps to increase labour-force participation. In both cases there is a rise in the effective labour supply, which leads to a reduction in wages.

([24]) This is equivalent to an inward shift of the Beveridge Curve, i.e. a reduction of the unemployment rate for a given vacancy rate.

Finally, there are productivity effects. ALMP programmes that improve the skills of participants or serve as a substitute for work experience can be expected to improve or to maintain the productivity of participants.

The productivity effect refers to firms being able to produce more (or better quality) for given costs, particularly wages. Therefore, productivity is thought to be the marginal product of labour, i.e. the output of an additional hour of work. In an artificial model economy these measures would be equal to the wage rate. However, as most European economies face enormous wage rigidities, i.e. wages are not perfectly correlated with productivity. Considering a conventional labour-demand condition, a rise in productivity would lead to an increase in employment for a given wage rate. Calmfors (1994) notes that the rise in productivity is not self-evident, because there is also an opportunity to produce the same output with fewer, but more efficient, workers. Additionally, Calmfors et al. (2002) note that the rise in productivity of the participants may also have a wage-raising effect through a rise in the reservation wage of the participants.

Such a theoretical analysis should serve as a guideline for an empirical analysis of the impacts of ALMP. The impacts on matching efficiency, i.e. the estimation of aggregated matching functions, are the most frequent types of macroeconometric evaluations. Other types of empirical models are the reduced form relationships, in order to estimate the overall effect, i.e. the effect through the different channels, on employment or unemployment. Although such a model does not provide a differentiated picture of the effects of ALMP, it is able to quantify the net effect on the whole economy.

Another concern regarding macroeconometric evaluations is a serious simultaneity problem. In general, deciding how much money is spent on ALMP directly relates to the situation in the labour market. Spending on ALMP should, therefore, be determined by a policy reaction function. As a result, ALMP activity does not only determine the dependent variable in a macroeconometric evaluation; the dependent variable also determines ALMP activity. This classical simultaneity problem does not allow for identification of the parameters of ALMP measures. To solve this problem, instrumental variable (IV) estimators should be used for the estimation. Therefore, the main problem is to find valid and good instruments. The validity of the instruments refers to the requirement that the instrument should not be determined by the dependent variable. The requirement for a good instrument is that the set instruments should be able to explain significantly the ALMP activity. From a practical point of view, the problem of finding an appropriate set of instruments is clearly a problem of data availability. Therefore, in most cases where not enough data is available, a good strategy is to use lagged values for the ALMP measures as instruments.

2.4. Cost-benefit analysis (CBA)

The previous sections introduced a number of econometric methods which can be used to assess the impacts of social programmes. Microeconometric evaluation studies are able to disclose whether social programmes have an impact on the individual. In addition, macroeconometric studies can cope with external effects that have to be taken into account. They are a necessary first step in assessing the value and merits of social programmes. If, for example, a microeconometric evaluation reveals that a certain programme has no impact at individual level, it hardly makes sense to complement it with a CBA and to assess its financial effectiveness.

Conducting a CBA widens the perspective of an impact analysis. A CBA is a method that provides a consistent, explicit and transparent procedure to evaluate public projects in terms of their consequences, i.e. in terms of their costs and benefits. The aim of a CBA is an efficient allocation of scarce resources, allowing it to serve as a normative tool in social decision-making. It is similar to a commercial profitability calculation conducted by private establishments. There are, however, also differences in that a CBA considers additional aspects, such as equity and distributional aims, and takes into account all costs and benefits to society as a whole.

The growth of public spending for social programmes has increased interest in a systematic evaluation in terms of costs and benefits. Originally, CBA was applied to technical projects, such as water resources, or engineering projects such as highways. It was not easy to transfer these techniques to social programmes because such programmes often involve impacts which

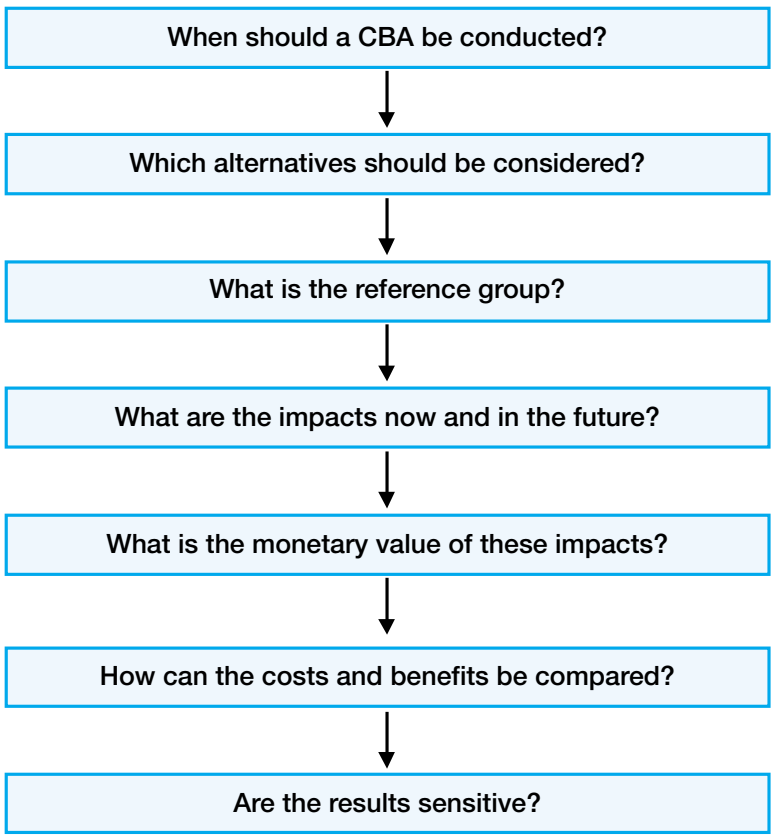
are hard to measure using market prices or even hard to assess. Additionally, social programmes often comprise distributional and equity aims which make some value judgements inevitable.

In the meantime, nearly all industrialised countries require such analysis for major social programmes (e.g. Boardman et al., 2001). The Unfunded Mandates Reform Act of 1995 requires that US agencies have to prepare an *ex ante* CBA for any regulation that may cost more than USD 100 million in any year. In Germany, *ex ante* CBA were established with the 1969 reform of the budget law which requires that CBAs have to be conducted for measures of considerable financial impact (Paragraph 7(2) *Bundeshaushaltsordnung*). *Ex post*, CBA are not explicitly regulated by law. Mandatory requirements for conducting *ex post* CBA do not exist in the US nor in Germany. However, in most cases, after programmes have been conducted officials find themselves having to justify these programmes. Thus, the intention to conduct similar programmes in the future depends on an *ex post* evaluation of the programmes.

Although CBA is a useful tool for assessing the efficiency of social programmes, it must be noted that its role in practice is rather limited, especially in the field of ALMP (e.g. Delander and Niklasson, 1996). One reason for this is the problems and difficulties inherent in the method. We will address them later. Another point is that the use and acceptance of CBA depends heavily on the institutional and political setting in which it operates. A major feature of CBA is its thinking in terms of alternatives. If, however, the political situation prohibits some alternatives which are potentially superior to the actual situation, then a CBA is also limited and restricted in its results. CBA can play a constructive role if politics bear the burden of providing a rationale for any governmental intervention.

CBA can be defined as a systematic, explicit and transparent method to assess the net present value (NPV) of all benefits less all costs, valued by a single monetary measure, which resulted from a certain social programme. In this sense CBA is a method that quantifies, in monetary terms, the consequences of political decisions. This general definition calls for specification and clarification. The following aspects have to be fixed which also form the different stages of a CBA (see also Figure 4).

Figure 4: **Different stages of a CBA**



2.4.1. Time of a CBA

In deciding when to conduct a CBA, one has to choose between *ex ante*, *ex post* and *in-medias-res*, see Boardman et al. (2001), who also introduced a fourth category, namely one which compares an *ex ante* CBA with an *ex post* CBA. An *ex ante* CBA is conducted before the project is actually implemented. Its major advantage is that there is still time to change certain aspects of the programme. The result of an *ex ante* CBA indicates whether a certain programme should be implemented or not. If more than one programme has been evaluated, it can help in deciding which one to choose. Its major drawback is the fact that most of the benefits or costs will arise in the future and thus have to be estimated with uncertainty. Although plagued with these uncertainties, *ex ante* CBA are useful for decision-making purposes because it is still possible to make different use of the resources.

In contrast, *ex post* CBA are conducted when the programme has already been implemented and all costs are irreversibly 'sunk'. They provide more accurate and detailed information about a social programme since they do not rely on estimates and can also be used to learn more about similar programmes which are still to be implemented. It should be mentioned, however, that they suffer from the same problems as *ex ante* CBA when looking at the medium/long-term benefits of social programmes, which again have to be predicted. *In-medias-res* CBA are conducted during the life course of a social programme. They share some advantages and disadvantages of *ex ante* and some of *ex post* CBA. If, for example, there are only low sunk costs, *in-medias-res* CBA can be used to shift resources to other more desirable projects. If the sunk costs are quite high, they only might, as was the case for *ex post* CBA, give a concluding assessment of the project.

2.4.2. Selection of the alternatives

Selection of the alternatives concerns the identification of the project options to be evaluated. Every social programme aims at one specific target, such as improving the labour-market prospects of disabled persons (e.g. Delander and Niklasson, 1996). Alternative policy instruments to achieve this aim could be for example wage subsidies or increasing public employment opportunities. Although research could make some proposals on potential alternatives, the available policy alternatives must be given by the policy-maker and are beyond the function of the analyst conducting a CBA.

This is important when interpreting the results of a CBA. If a CBA identifies one alternative as the optimal one, i.e. one with the highest present value of net-benefits, this only refers to the alternatives under consideration. There could be other alternatives with an even higher benefit which were not considered in the CBA.

2.4.3. Defining the reference group

Defining the reference group means deciding whose costs and benefits count, i.e. whose interests should be considered. Different perspectives are possible. One perspective is to differentiate between the target group, the non-target group and society as a whole. Additionally, one can include financiers as a separate group. A further alternative is the distinction between a local and global perspective. The choice of a certain perspective is given by the institutions who have ordered the analysis and is, therefore, not the function of the analyst. The perspective heavily influences which impacts have to be considered.

A profitability analysis in private establishments can be regarded as a CBA where only the private benefits, i.e. revenues, and private costs are considered. The net benefit is in this case identical to the firm's profits. A social CBA on the other hand, i.e. one which takes the perspective of the society as a whole, extends those private costs and benefits by including all impacts of the programmes whether they are private or social, tangible or intangible, direct or indirect.

The target group consists of those individuals who are eligible for the programme under consideration. If we consider a labour-market programme aimed at improving the labour-market prospects of disabled persons, the benefits for the target group consist of the effects of these programmes on the employment probability and wages. The costs, which have to be covered by this group, consist of earnings and transfers from which the participants have to abstain during the participation, i.e. opportunity costs of foregone income.

If we look at the group of non-participants, additional benefits and costs arise. Benefits which have to be taken into account in this perspective would be the additional output

attributable to the participants after they have finished the programme, the increasing tax payments, and more intangible impacts such as reduced delinquency. Costs which arise for this group are the expenditures for operating the programmes and the opportunity costs for the participants during participation. Aggregating these two views provides the social perspective which yields the net-benefit effect of the programme. This perspective yields the most complete, exhaustive and most complex view, since it requires the consideration of all possible impacts of a programme.

2.4.4. Enumerate and forecast all impacts

Probably the most crucial and important step in conducting a CBA consists of a complete enumeration of all impacts of a programme as costs and benefits, with a forecast of these impacts over the lifetime of the project. A condition for the identification of the impacts of a programme is the existence of a model which gives us the cause-and-effect relationship between the programme under consideration and the costs and benefits perceived by the target group. For some impacts this relationship is obvious, for example there is no doubt that measures for the disabled will influence their labour-market prospects or that the construction of a highway will reduce travel costs. The effects social programmes might have on more intangible factors, for example decreasing delinquency or more equity, are much harder to assess.

Impacts can be classified into real and monetary, direct and indirect, tangible or intangible, final or intermediate and, finally, internal or external effects. Real impacts mean the final impact on the social welfare, i.e. the final utility gain or loss of a programme, whereas monetary impacts only change relative prices while having no real welfare effects. Direct impacts refer to the intrinsic project goal while indirect effects cover those not intended by the programme. Examples of indirect effects in the context of labour-market programmes include the displacement effect or dead-weight losses. Final impacts occur at the level of the consumer while intermediate impacts occur at the producer level. The distinction between internal and external impacts is again closely related to the perspective of the CBA. Internal impacts are those which emerge within a pre-specified target group and which could also be made up of external effects. These could occur, for example, if a programme which should bring the long-term unemployed back into work not only increases their employment prospects but also has other positive externalities such as crime reduction. Spillover effects and externalities are also responsible for some effects outside this group or area. Popular examples for external effects are air and noise pollution but could also contain positive externalities like those mentioned before.

The impacts to be considered depends heavily on the perspective chosen. In the example of the labour-market programme for the long-term unemployed, the target group benefits could be increasing employment probability after finishing the programme, therefore increasing income and life quality, reduced alcohol and drug abuse and reduced criminal activity (see for a practical application Long et al., 1981). Costs for the target group include forgone income while in the programme. In respect of society as a whole, additional external benefits, such as avoided costs for alternative services and additional costs, such as programme expenditure, have to be taken into account.

Other aims and objectives, which arise at the level of the society as a whole and which differentiate a CBA from private profitability calculations, are distributional and equity effects. The present value of a social programme only contains information on whether the benefits exceed the costs, i.e. it answers the question of whether the programme is efficient. Another important issue which has to be considered, however, is how these costs and benefits are distributed within society.

At this point a fundamental problem arises. In most cases, evaluating distributional and equity goals proves very difficult. Every publicly financed measure, for example in the area of ALMP, implies redistribution in that it transfers income from the non-participants to the participants of the programmes. Usually one Euro taken from non-participants is bestowed the same weight as one Euro given to the participants. One could, however, argue in this context that the marginal utility of one Euro is higher for the non-participants, who are assumed to have a lower income, than for the participants and that therefore the weights attached to the participants should be

higher. But since there are no objective and transparent methods to measure these weights accurately, they must be set within a political context and are therefore arbitrary and vulnerable to criticism.

2.4.5. Measuring and aggregating impacts

Having defined the impacts which have to be taken into account, the next step is to measure them in monetary terms. Measuring the impacts, i.e. 'monetising' them, is the heart of a CBA and also, in most cases, the most difficult step. For some impacts, for example the direct tangible effects of a programme, one can rely on market prices, whereas other impacts, such as saved lives, quality of life or equity are much harder to assess.

The simplest proceeding is an accounting of the monetary flows caused by a certain programme, for example expenditure on salaries or saved payments for social welfare. Another straightforward possibility, which is especially applicable in the case of tangible effects, is the direct observation of market prices. Although it is useful to start with market prices, very often it is necessary to adjust them in order to include external effects. If, for example, the social costs of labour input into a social programme should be evaluated, and if there is a large amount of involuntary unemployment, wages may have to be adjusted downwards in order to account for this idle input factor. This adjustment ensures that market prices correspond to the real net impact on welfare, i.e. to their shadow prices. Only in the case where there are no market distortions, i.e. when there is perfect competition and no external effects, will market prices correspond to their shadow prices. Market prices can also be used to assess the value of more indirect effects, such as declining delinquency. One possibility in this context would be to observe how prices for real estate evolve and to use the increase in these prices as an approximation for the utility of declining delinquency.

For intangible goods, like the value of a life saved or the increasing contentment of people who participate in a labour-market programme and afterwards find a job, referring to market prices is not feasible. In this case other methods have to be applied. One way is to directly ask for the amount of money individuals are willing to pay for a certain measure, for example an improvement of the air quality or the like. To this end, a number of questions have been developed. By interpreting the results one has, however, to be aware of the hypothetical feature of these questions and therefore their limited applicability. Another way to measure intangible goods is the observation of political preferences. By observing, for example, that in the past a life-saving programme had been established that cost EUR 100 000 and was able to save 10 lives, one could infer that the value of a life amounts to EUR 10 000. Again such indirect inferences have to be made with care.

When a social perspective is chosen, one additional problem that arises is the aggregation of individual costs and benefits. We have addressed this issue already and also mentioned the problems and difficulties arising in this context. Aggregating the individual costs and benefits presupposes attaching weights to each individual. Due to the absence of other convincing weighting schemes, equal weights are usually used. One justification for this simplification is the following: when the net benefit of a social programme is greater than zero, the sum of all individual benefits exceeds the sum of all individual costs. In other words, the size of the 'common cake' increases (e.g. Delander and Niklasson, 1996). If this is the case, there is room for redistribution in the sense that the losers of the programme are compensated by the winners. We should note, however, that this indemnification view is only theoretical and one has to consider that with such redistribution policies, additional costs might arise which then will influence the net benefit of the programme.

If it turns out that the problem of measuring the utilities of a social programme cannot be resolved satisfactorily, then at least a cost-effectiveness analysis (CEA) can be conducted. Cost-effectiveness analysis evaluates programmes by comparing the costs associated with these programmes with a single quantified but not monetised effectiveness measure. It may be that in evaluating a health care programme the lives saved by alternative programmes have to be assessed. Conducting a CBA would require the valuation of the lives saved using monetary units. Instead, the analyst could also assess the different programmes by relating their costs to the lives saved, i.e. by constructing cost-efficiency

ratios. Another example of a cost-effectiveness analysis will be given later on.

2.4.6. Comparing costs and benefits

Once all costs and benefits are itemised and measured, another problem which has to be resolved stems from the fact that costs and benefits do not occur at one point in time but rather follow a dynamic path. In order to compare costs and benefits which occur in the future with those which occur now, it is necessary to define a discount rate r. Once a discount rate is found, the flows of costs Ct and benefits Bt can be discounted to give the net present value NPV of a social project according to:

$$NPV = \sum_t \frac{B_t - C_t}{1+r}. \qquad (32)$$

A simple decision rule is to implement a project if it has a net present value which is larger than zero and to reject it otherwise. If more than one project is evaluated, the one with the largest NPV should be implemented.

The choice of an appropriate discount rate, however, is contentious. One approach could be to use some market interest rate as an approximation to the social discount rate. The market interest rate, however, does not correspond to the social rate. This is because capital markets are not perfect and thus the time preferences of individuals do not correspond to market interest rates. Other points responsible for this departure are, for example, uncertainty about future inflation and the existence of tax effects which will be reflected in market rates.

In most programmes, costs occur during implementation while the benefits are spread over the future. Thus small changes in the discount rate might have a great impact on the net present value of the project. The choice of an appropriate discount rate is crucial in conducting a CBA and therefore an eligible candidate for a sensitivity analysis.

An alternative decision rule, which does not rely on a certain discount rate, is to determine the internal rate of return of the social programme. The internal rate of return r of a social programme is equal to the discount rate for which the net present value of the social programme becomes zero:

$$NPV = \sum_t \frac{B_t - C_t}{1+\rho} \overset{!}{=} 0. \qquad (33)$$

When more than one project is evaluated, the one with the highest internal rate of return is chosen. When only one project is considered, again a reference value is needed. If the internal rate of return is above this reference value, the project should be realised. If not, it should be rejected.

This method, however, has some drawbacks. If the flow of net-benefits alters between positive and negative values during the lifetime of the project, more than one value for the internal rate of return might fulfil the above equation. Additionally, decisions which rely on this rule might differ from decisions derived by using the net present value of the projects.

2.4.7. Conducting sensitivity tests

While presenting the necessary steps for conducting a CBA, there are various caveats and drawbacks which might heavily influence the results and recommendations. CBA relies on a number of assumptions and uncertain predictions, for example, with regard to the potential impacts, the way to measure them and at which interest rate to discount them. All these topics are eligible for a sensitivity analysis which examines how sensitive the NPV results are to different assumptions about those key parameters. Careful scenarios, which vary the most important assumptions, are one way to protect the results against indicated doubts.

In order to illustrate the above, one hypothetical and one real case study in the field of ALMP will be used. Let us assume that in a hypothetical situation an economy tries to fight its rising unemployment by running different measures of ALMP ([25]). More specifically, four different instruments will be considered: LMT, youth measures, subsidised employment and measures for the disabled. The following table contains hypothetical figures for expenditure on these programmes and for the number of participants in the different measures. Total expenditure amounts to EUR 33 million, with participation by 15 753 individuals.

([25]) Another assumption is that we are only interested in a reduction of overall unemployment. Other programme objectives are not included in this simple example.

Table 4: **Hypothetical example before reallocation**

Programme	Expenditures (Million EUR)	Participants	Per capita expenditures (EUR)	Treatment effect (%)	Employed participants
LMT	5	2 772	1 804	80	2 218
Youth measures	10	5 188	1 928	20	1 038
Subsidised employment	8	5 273	1 517	5	264
Measures for disabled	10	2 520	3 968	5	126
Total	33	15 753	2 095	23	3 645

Let us now assume that a microeconometric evaluation study has been conducted and that this study revealed that the various measures had different impacts on individual employment probability after finishing the various programmes. LMT measures were most successful in fighting unemployment with a treatment effect of 80 %. This means that participants in LMT measures had an 80 % chance of finding a new job after finishing the programme compared with an appropriate control group of non-participants. The treatment effects of the other programmes can be interpreted in a similar way.

Our hypothetical example indicates that the EUR 33 million spent enable 23 % of the 15 753 participants to find new employment. The cost efficiency ratio before an efficient reallocation is thus 23 %. The findings of the microeconometric analysis can now be used to re-allocate expenditure more efficiently. Let us consider that, on the basis of these findings, the government has decided to expand the measures of LMT and youth measures by 100 % and 50 % respectively, and at the same time to decrease expenditure on the other two labour-market programmes by 75 % respectively. This reallocation could be achieved by a more generous or a more restrictive interpretation of the regulatory laws. The following table contains the expenditure and the number of employed participants after such a re-allocation ([26]).

Table 5: **Hypothetical example after reallocation**

Programme	Expenditures (Million EUR)	Participants	Per capita expenditures (EUR)	Treatment effect (%)	Employed participants
LMT	10	5 544	1 804	80	4 435
Youth measures	15	7 782	1 928	20	1 556
Subsidised employment	2	1 318	1 517	5	66
Measures for disabled	2.5	630	3 968	5	32
Total	29.5	15 274	1 931	40	6 089

([26]) This allocation is not simply a replacement of more expensive with cheaper measures but a reallocation between less and more efficient ones. Even though youth measures, for example, are more expensive than subsidising employment (both in the total amount and per capita spending), the first measure is the more efficient one and thus should be expanded.

Even with a smaller amount of money, more participants were put back into the workforce and the average treatment effect of all measures has nearly doubled. Since an econometric evaluation also delivers other valuable information, for example for which subgroups of participants which programmes are especially successful, this gives a first impression of the value and the usefulness of econometric evaluation studies.

Another real example of a CBA (Long et al., 1981 and Delander and Niklasson, 1996) aimed to assess the effects of a US federal social programme for economically disadvantaged youths, the Job Corps programme. This programme is a comprehensive set of services for disadvantaged youths, such as vocational skill training, basic education and health care. The aim of the programme was to increase the employability of the participants.

In order to take account of distributional effects, three different perspectives have been considered: the group of participants, the group of non-participants and an aggregated view of society as a whole. Focusing on the perspective of society as a whole addresses the issue of efficiency while hints about the distributional consequences of the programme can be obtained by looking at the two groups separately. A wide range of potential benefits and costs have been taken into account as summarised in the following table:

Table 6: **Costs and benefits of the Job Corps programme**

Benefits	**Examples**
Output produced by Corps members	In-programme output, increased post-programme output, increased tax payments, increased utility due to preferences for work over welfare.
Reduced dependence on transfer programs	Reduced transfer payments, reduced administrative costs.
Reduced criminal activity	Reduced criminal justices system costs, reduced personal injury and property damage.
Reduced drug/alcohol abuse	Reduced treatment costs, increased utility from reduced drug/alcohol dependence.
Reduced utilisation of alternative services	Reduced costs of other programmes than the Job Corps.
Other benefits	Increased utility from redistribution, increased utility from improved well-being of Corps member.
Costs	**Examples**
Programme operating expenditures	Centre operating expenditures, excluding transfers to Corps members, central administrative costs.
Opportunity costs of Corps-member labour	Forgone output, forgone tax payments.

The effects of the programme were estimated using data from a survey conducted among the group of participants and a comparison group who were never enrolled in this programme. Since the programme was not constructed as a social experiment, multiple regression techniques, which account for both observable and unobservable effects, had to be employed to estimate the treatment effect of the programmes.

The other costs and benefits were valued so that they reflect the resources saved, consumed or produced as a result of the Job Corp programme. The benefits arise primarily from two factors. First, Corps members made less use of other social programmes and committed fewer crimes. Second, corps members improved their long-term employment prospects and increased their contribution to the gross national product (GNP) and tax payments. The effects of these changes were valued by multiplying the estimated change in the behaviour of the participants due to the programme by estimated dollar values.

The contribution of the participants to GNP, for example, was estimated by using the difference

between the earnings of the Job Corps members and the non-participants. The implicit assumption in this procedure is that labour markets are competitive, so that the wages are equal to the marginal contribution to the GNP, and that there are no displacement effects due to the programme. Using this method, the authors estimated that the total discounted value of the increased output over the first two years after the end of the programme is USD 925 per participant.

The benefits stemming from reduced delinquency were estimated by multiplying the estimated changes in arrests due to the programme by the shadow prices, which are equal to the cost savings per avoided arrest. The shadow prices were calculated by considering the costs caused by arrested persons, for example police custody, arraignment, detention, etc. Costs of the project mainly comprised operating expenditures for the programme centres and central administration costs. They represent costs to non-participants. Other categories which do not appear in the financial accounts are forgone income of the participants and forgone tax payments.

Making a set of assumptions, for example setting the discount rate to 5 %, assuming the effects of the programme to diminish at a rate of 50 % every 5 years, the present value of the net benefits can be estimated. The results suggest that the programme yields a net benefit to society and also to the target group. The group of non-participants experiences a slightly negative value from the programme. The programme was estimated to represent a socially efficient use of the resources and additionally entails redistribution from the group of non-participants to the group of participants.

3. ALMP in the EU and empirical findings

3.1. Recent developments in ALMP in the EU

ALMP have been seen as one way to fight the rising unemployment rates and the disequilibrium (skill shortage, etc.) in the labour markets which have arisen in most European countries since the early 1970s. Table A.1 in the annex shows standardised unemployment rates in the EU from 1985 to 2000. In 1985 the EU average was 10.8%. After declining in the following years, it came back to this value in 1994. In subsequent years unemployment eased, bringing the rate down to 6.9% in the year 2000 ([27]). The average reduction from 1985 to 2000 was 36%. It is quite interesting to compare this with the picture in the different countries. Sweden and Finland had to deal with extraordinary increases in unemployment during this period. The rate rose by 108% in Sweden (from 2.8% to 5.9%) and 94% in Finland (from 5.0% to 9.7%). In France there was a slight increase from 8.3% to 9.3% (+12%), whereas in all the other countries unemployment has fallen since 1985. The biggest decreases can be found in Ireland (-76%, from 17.7% in 1985 to 4.2% in 2000) and the Netherlands (-82%, from 13.1% to 2.8%). The remaining countries had decreases between 15% (Germany) and 52% (Portugal).

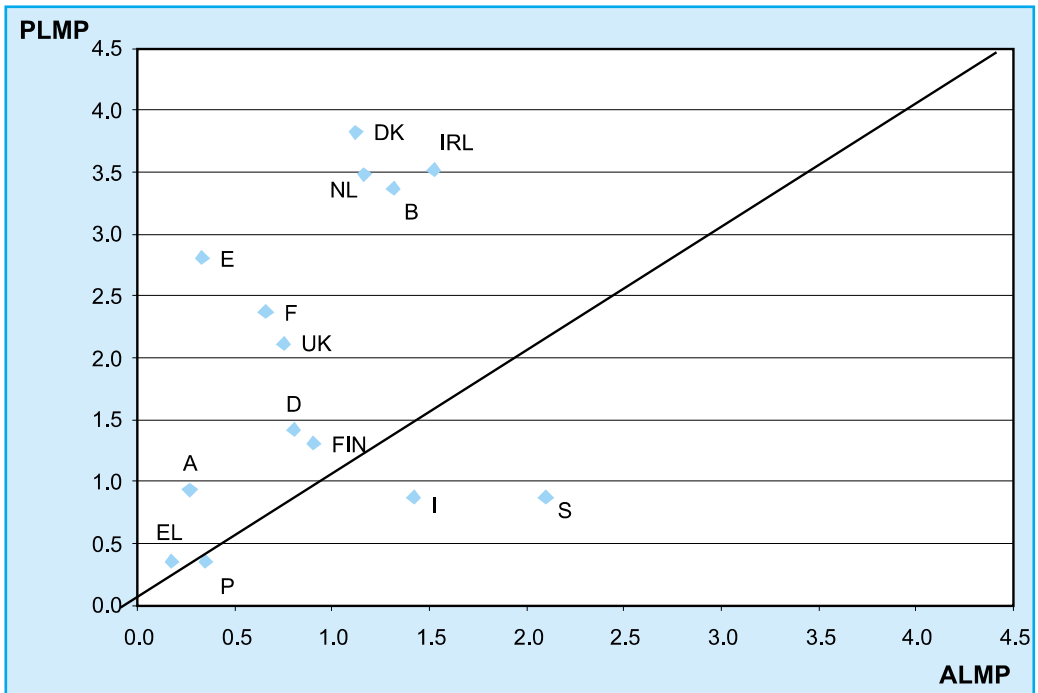
Whether these reductions are due to ALMP or other factors has to be determined by evaluation studies, which will be reviewed in the next chapter. We have pointed out already, that the importance of ALMP has been growing enormously from 1985 to 2000. The ratio between ALMP and PLMP has risen from 0.44 in 1985 to 0.68 in 2000 (EU average). Even though there is an EU-wide increase in ALMP this is not true for all of the countries. Therefore we need to take a closer look at the evolution in specific countries. Tables A.2 and A.3 in the annex show the spending on ALMP and PLMP for the years 1985 to 2000. If we look at Table A.3 we see that spending on PLMP peaked in the mid-1990s in line with the peak in unemployment. The obvious reason is that unemployment benefits are entitlement programmes, i.e. rising unemployment automatically increases public spending on passive income support. Active labour-market programmes, on the other hand, are discretionary in nature and therefore more easily disposed with in a situation of tight budgets. Despite that, ALMP spending reached its highest level in the mid-1990s; a clear indication that ALMP were seen as a suitable measure against the bad labour-market situation.

Figures 5 and 6 compare spending on ALMP and PLMP in the years 1985 and 2000. The bisecting line in the figures indicates a balanced relationship between both measures, whereby countries in the left/right half direct more money into passive/active measures respectively. It is interesting to note that the differences between the countries were more pronounced in 1985. Italy and Sweden spent more money on active measures, spending on PLMP by Belgium, Denmark, Ireland and the Netherlands exceeded spending on ALMP by more than two percentage points of GDP. In 2000 the situation is much more balanced and there is a tendency to an equal spending on both measures. Italy still spends more money on active measures, Greece and Sweden spend equally for ALMP and PLMP. The other countries all spend more money on PLMP but they are moving closer to an equal spending. Denmark still spends most money on PLMP (3% of the GDP), but the difference between PLMP and ALMP was only 1.46 percentage points in 2000, compared to 2.7 percentage points in 1985.

Having discussed the importance of ALMP in contrast to PLMP in general, we look now at the importance of one special programme, LMT. Table A.4 in the Annex shows the spending on LMT as a percentage of GDP for the years 1985 to 2000. We see that Denmark (0.72%) and Sweden (0.64%) direct the highest amounts of their GDP to LMT. Table 7 demonstrates the relative importance of LMT, by showing the spending on LMT as a

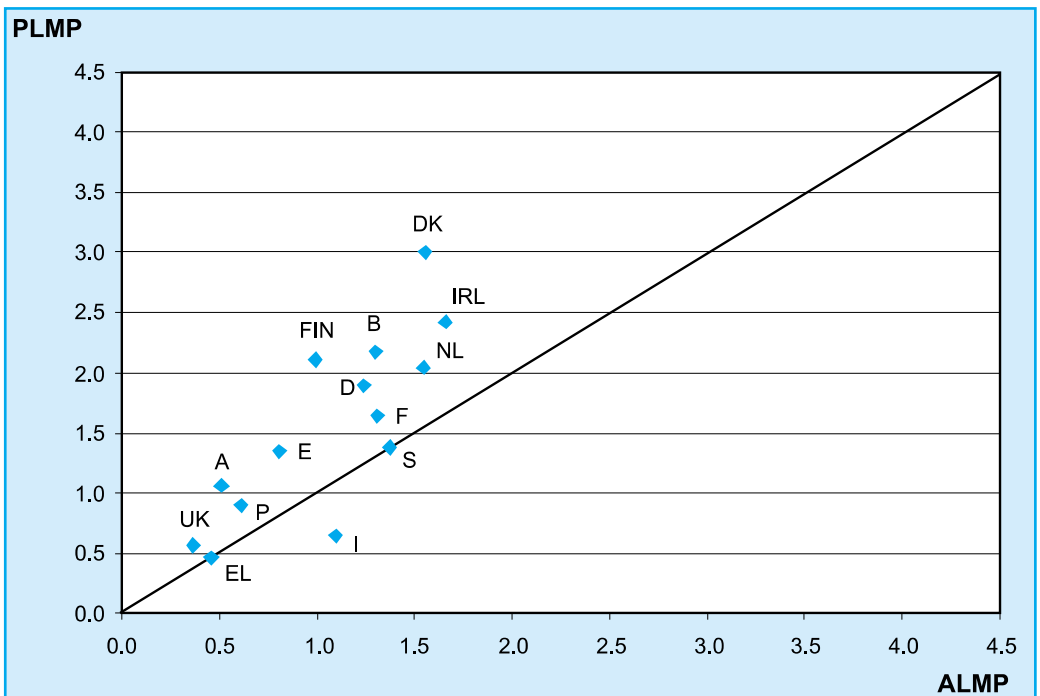
([27]) Unfortunately the OECD does not provide standardised unemployment rates for Austria and Greece until 1993.

Figure 5: **Spending on ALMP and PLMP in the EU, 1985**



Data for Italy from 1991, Denmark 1986, Portugal 1986
Source: OECD, several issues and own calculations

Figure 6: **Spending on ALMP and PLMP in the EU, 2000**



Data for Italy from 1991, Denmark 1986, Portugal 1986
Source: OECD, several issues and own calculations

percentage of the total spending on ALMP. Denmark directs 43 % of its spending on ALMP to LMT, in contrast to Italy's 5 %. The EU average spend on LMT is 0.3 % of GDP and this corresponds to 26 % of the total spending on ALMP. The justification for this expenditure has to be seen from the empirical evaluation of these programme studies and this will be done in the next chapter.

Table 7: **Expenditure on LMT as a percentage of total spending on ALMP**

(%)

	1985	1986	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	Average
Austria	30	34	31	33	29	32	31	33	38	39	33	36	33	33
Belgium	15	15	17	19	19	22	21	20	19	21	18	18	18	19
Denmark	. .	38	24	28	27	27	40	52	60	56	58	56	54	43
Finland	29	29	25	24	26	28	28	28	33	35	32	33	30	29
France	39	36	41	38	36	35	32	29	26	25	23	21	19	31
Germany	25	26	37	35	38	35	30	28	31	29	27	26	27	30
Greece	12	14	44	52	32	26	18	29	19	18	62	. .	. .	30
Ireland	41	36	32	19	. .	. .	15	13	13	. .	. .	. .	. .	24
Italy	. .	. .	. .	1	1	1	1	1	1	12	13	11	. .	5
Netherlands	11	16	21	22	25	23	27	26	23	24	20	21	19	21
Portugal	. .	51	20	24	31	30	30	29	33	35	38	36	25	32
Spain	6	11	19	22	18	23	36	38	50	31	24	17	19	24
Sweden	24	25	31	41	35	26	26	24		20	23	27	23	27
UK	9	8	34	26	27	26	25	22	21	19	18	15	214	20
Means	17	24	28	27	27	25	26	27	29	29	29	28	27	26

..=Missing values
Source: OECD Employment Outlook, 2002.

3.2. Microeconometric evaluation studies

To see how successful labour training programmes have been in the recent years, we have reviewed 45 studies from 10 European countries. The results from microeconometric evaluation studies in Germany and Europe will be presented before those of macroeconometric studies. The results of the studies, the outcome variables used, the programmes evaluated and the methods applied are also summarised in Table 8 at the end of this chapter.

3.2.1. Germany

All evaluation studies of vocational training for West Germany are based on the German Socioeconomic Panel (GSOEP)-West. With respect to the design of the programmes, measures on-the-job and off-the-job are considered as well as measures with and without income maintenance. The applied evaluation methods include discrete hazard rate models, matching, instrumental-variable estimation and simultaneous dynamic models. Besides unemployment duration and the re-employment probability, hourly wages and employment stability are considered as outcome variables. Hujer et al. (1998, 1999a) find that participation in vocational training has significant effect in reducing unemployment duration. In further studies, Hujer et al. (1999b), and Hujer and Wellner (2000b) discover positive effects only for short courses (< 6 months), whereas long courses do not have (significant) positive effects. Hujer et al. (1999c) found that on-the-job training has no significant effect on unemployment duration, whereas off-the-job training reduces it in the short-term and has no significant effects in the long-term. This finding corresponds to Pannenberg (1995, 1996), who observes that participation in off-the-job training increases re-employment probability in the short term. In contrast to this positive finding, the following studies produce rather negative results. Prey (1997) examines vocational training with income maintenance and finds negative (no) effects for men (women) on the employment probability. In a further study, Prey (1999) finds negative (no) effects on employment probability for measures with (without) income maintenance and no effects on the wages. Staat (1997) studies public sector training with income maintenance. His results indicate positive effects on the search duration only for subgroups, no significant effects on the employment probability but positive effects on wages.

The studies for East Germany are either based on the East German Labour Market Monitor (EGLMM), the GSOEP-East or the Labour Market Monitor Saxony-Anhalt, a panel based on the population of the state of Saxony-Anhalt. As with the studies for West Germany, training is analysed on- and off-the-job, as well as with and without income maintenance. The period under consideration ranges from 1989 to 1998. There is examination of several outcome variables; (un)employment duration, (stable and unstable) employment (probabilities), job search, working time and wages. Fitzenberger and Prey (1997) find that training outside the firm shows considerable positive effects on employment probabilities, whereas training inside has negative effects. Hübler (1998) found that on-the-job training increases job security, whereas off-the-job training leads to higher earnings. But this is true only for privately financed training. Publicly financed training has only positive effects in the short-term. Fitzenberger and Prey (2000) question training supported by public income maintenance outside a firm, and discover positive effects for employment and earnings (but only a few significant in the long-term). Pannenberg (1995, 1996) ascertains positive effects on the re-employment probability and the income for vocational training on- and off-the-job. Lechner (1999b) examines enterprise-related continuous vocational training and finds positive income effects too, but no effects on unemployment probability. Kraus et al. (1998) observe, for a sub-period of their study, positive effects on (stable) employment for on- and off-the-job training. Hübler (1994) examines on-the-job qualifying measures and finds that training induces search activities and reduces the effective hours of work.

Many studies do not find any clear positive or negative effects. Bergemann et al. (2000) examine multiple participation in further training. They discover no positive effects for first training programmes, and the additional effects of a second participation are, on average, not different from zero. Hujer and Wellner (2000a) find no significant effects for public-sector sponsored vocational training on unemployment duration, but very weak hints that short courses seem to be more effective in reducing the duration. Lechner (1999c) investigates off-the-job training (publicly financed and enterprise related). He cannot establish robust positive effects on either employment probability or the earnings. Staat (1997) examines public sector training with income maintenance and finds no effects either on the search duration, or on the employment stability or the level of hourly wages.

The results of the following studies tend to be negative. Fitzenberger and Prey (1998) examine the effects of training within and outside the firm and with and without public income maintenance on employment and wages. Mostly they get negative or no effects, differing with respect to the specification. Lechner (2000) finds no positive long-term effects on employment probabilities or earnings for public-sector-sponsored vocational training. He gets negative results regarding the risk of unemployment in the short-term.

3.2.2. Other European countries

This section presents an overview of microeconometric studies in Europe. Due to numerous microeconometric evaluations in Europe, the following section will focus only on recent studies that, in addition, deal with training programmes. Further overviews of European studies on the microeconometric evaluation of ALMP are given, for example, by Steiner and Hagen (2002) and Heckman et al. (1999).

3.2.2.1. *Switzerland*

Lalive et al. (2000) analysed vocational training and job creation programmes in Switzerland. They found that both programmes have negative effects on employment during participation. Furthermore, they found that these negative locking-in effects seem to dominate the positive effect on employment after the programme has expired. This problem seems to be particularly severe for job creation schemes for men, whereas for women it seems to be less important.

Another evaluation for Switzerland was performed by Gerfin and Lechner (2000). They also analysed different vocational training and job creation programmes, including training to improve basic skills, language and informatics abilities as well as further training. Their results indicate that the informatics courses and further training programmes have no effect on employment and that the basic skills and the language courses seem to have a negative effect. Furthermore, they found that unemployment duration prior to the programme participation seems to have an influence on the programme effect, i.e. programmes are

more effective for long-term unemployment (LTU). For the job creation programmes they found negative effects on employment, whereas positive effects were found for women.

3.2.2.2. *Austria*

Zweimüller and Winter-Ebmer (1996) analysed the effects of Austrian vocational training on employment stability. Their results indicate that training programmes significantly reduce unemployment risk for the participants.

Winter-Ebmer (2001) analysed the employment effects of job creation and training programmes by the *Österreichischen Stahlstiftung* on employment. These programmes are thought to counteract staff reduction within the steel industry, so only steelworkers are eligible for the participation. The results of Winter-Ebmer (2001) indicate that there are positive employment effects primarily for unemployed older than 27 years, whereas for younger people there are no effects.

3.2.2.3. *Belgium*

In a study for Belgium, Cockx et al. (1998) compared the effects of training programmes and subsidised employment programmes on employment duration. They analysed in-house and external training programmes. Whereas they found no effect from subsidised employment and external training programmes, they found a positive effect from in-house training programmes. As internal training programmes are an investment in firms' human capital, they should reduce the risk of being laid off.

In a subsequent study, Cockx and Bardoulat (2000) analysed the effect of vocational training programmes on exit from unemployment evaluating training programmes not organised by firms. They found a negative locking-in effect during the programme and a positive effect after the programme had expired. The positive effects subsequent to the programme, compensate for the locking-in effect and thus the overall effect was positive.

3.2.2.4. *France*

French vocational training and job creation programmes were analysed by Bonnal, Fougère and Sérandon (1997). In their analysis they distinguished between unemployed people with vocational education and those without, finding a positive effect from training programmes on exit from unemployment for the unemployed without vocational education. The found no positive results from job creation programmes. Brodaty et al. (2001) obtained similar results in their study where they used the same dataset but different evaluation methods.

3.2.2.5. *Denmark*

Jensen et al. (1999) analysed a youth unemployment programme that was established in Denmark in 1996. Unemployed youths without vocational education are obliged to participate in a vocational training programme to remain entitled to unemployment benefits. Jensen et al. (1999) analysed the effects of this programme on entry into regular employment and into a regular vocational education. While they find no significant effect on entry into regular employment, the effects on entry into regular vocational education are positive.

3.2.2.6. *United Kingdom*

Firth et al. (1999) analysed the effects of Employment Training (a vocational training programme) and Employment Action (a job creation programme) on the separation rate into employment in the United Kingdom. Employment Training offered a significant positive effect whereas for Employment Action no significant effects could be found. The same results were also obtained by the prior study from Payne et al. (1996) that analysed the effects of both programmes on employment.

3.2.2.7. *Ireland*

For Ireland, O'Connell and McGinnity (1997) analysed the effects of classroom training, on-the-job training and employment subsidy programmes on employment. Their results indicate that both training measures significantly increase employment, whereas for the employment subsidy programme no significant effect could be found.

3.2.2.8. *Norway*

Aakvik, Heckman and Vytlacil (2000) analysed the effects of subsidised employment and training programmes on employment for women in Norway with long-term diseases. All programmes recorded a negative effect. The authors note proposed that the negative effect may have been caused by the selection of the participants. It seems that those placed into the programmes

were mostly unemployed women who had good employment opportunities anyway.

3.2.2.9. *Sweden*

Compared to other European countries, Sweden has a long tradition of ALMP. Consequently, the Swedish active policy measures were evaluated more frequently than other European countries. Presented here are only three of the latest microeconometric evaluations for Sweden. A complete overview can be found in Calmfors et al. (2002).

Johansson and Martinson (2000) compared traditional vocational programmes (organised by the labour administration) with a special training programme to qualify the unemployed for jobs in the information technology sector (Sweden Information Technology – SwIT). This training programme is carried out in cooperation with industry in order to ensure a sufficient supply of qualified workers. Analysing the effects of both programmes, Johansson and Martinson (2000) find that the Sweden Information Technology programme increases employment probability more than the traditional training programmes. The authors see this as evidence that a more focused contact with firms can increase the efficiency of training programmes.

Larsson (2000) analysed the effects of job creation and training programmes for young people. The study finds no evidence that there is a positive effect from either set of programmes on regular employment or on regular vocational education.

Sianesi (2001) analysed the effects of ALMP programmes without a differentiation between the different programme types, finding a positive impact on the registered unemployed. A serious problem seems to be the fact that participation serves as a requirement to renew entitlement for unemployment benefits. Finally Richardson and van den Berg (2001) analysed (in an empirical study) the impact of employment training in Sweden targeted at unemployed individuals as well as employed persons who are at risk of becoming unemployed. The results show highly positive effects, with a doubling of individual re-employment rates after completion of training. Regarding the time spent within the programme, the individual net effect on unemployment duration reduces to zero.

3.3. Macroeconometric evaluation studies

Macroeconometric evaluations are rare in Europe, i.e. for most European countries there are no studies. As a single time series at national level usually does not provide enough observations, most studies rely on pooled cross-section time series data. Primarily these are studies using regional data to evaluate ALMP for one specific country. Besides these studies there are also some using a cross-country dataset for the OECD countries (e.g. Jackman et al., 1990 or OECD, 1993). Since this data, and hence the studies, are not restricted to Europe, statements for Europe alone cannot be made. Furthermore, cross-country studies suffer from the problem that they are supposed to analyse heterogeneous policy measures, which is a major drawback of these studies.

3.3.1. Sweden

Sweden provides the most macroeconometric evaluations using regional data. An overview can again be found in Calmfors et al. (2002). A prominent study for Sweden was conducted by Calmfors and Skedinger (1995). They used a reduced form relationship to analyse the effects of job creation and training programmes on the total rate of job seekers (i.e. openly unemployed and programme participants relative to the labour force). It is worth noting that they have looked specifically for the simultaneity problem of ALMP. As instruments they used not only labour-market indicators but also political factors such as the proportion of seats in the parliament assigned to left-wing parties. Their results indicate that job creation schemes tend to crowd out regular employment and that the results for vocational training, although they are unstable, are more favourable than the results for job creation schemes.

3.3.2. Germany

In addition to the studies for Sweden, there are some studies for Germany that were conducted with regional data. Büttner and Prey (1998) evaluated training programmes and public sector job creation for West Germany. They find that job creation schemes reduce mismatch, whereas training programmes do not have any significant effects. Prey (1999) extends this work by addi-

tionally controlling for the regional age structure and recipients of social assistance and estimating separately for men and women. She finds that vocational training increases (decreases) the mismatch for women (men), whereas job creation scheme decreases the mismatch for men.

Pannenberg and Schwarze (1998) use the data from 35 local labour office districts to evaluate training programmes in East Germany. They find that the programmes have negative effects on regional wages. Schmid et al. (2000) estimate the effects of further training, retraining, public sector job creation and wage subsidies on long-term unemployment and exit from unemployment. They find that job creation schemes reduce only 'short' long-term unemployment, vocational training reduces long-term unemployment and wage subsidies help only the very long-term unemployed. Steiner et al. (1998) examine the effects of vocational training on the labour-market mismatch in East Germany. They observe only very small effects on the matching efficiency which disappear in the long term. Hagen and Steiner (2000) evaluate vocational training, job creation schemes and social assistance measures (SAM) for East and West Germany using the data from local labour office districts. The estimated net-effects are not very promising as all measures increase unemployment in West Germany. Only social assistance measures reduce unemployment slightly in East Germany, whereas job creation schemes and vocational training increase it too.

3.4. Vocational training programme success in previous years

Table 8 summarises empirical microeconometric studies for Germany and Europe. The first thing to note is that there seem to be more favourable effects of training programmes in general compared to other types of programmes. Since training programmes are one of the most important measures of ALMP in Europe they are often compared to other measures such as job creation schemes or subsidised employment programmes. Whereas we find positive effects for vocational training programmes in most cases, the effects of other measures are positive only in one case. Furthermore we find that training programmes that are organised by firms seem to be more effective compared to publicly organised programmes. This may be reasoned by a closer contact with the firms which may increase the efficiency in adjusting the skills of the participants to the demands of the firm.

We find immense differences between East and West Germany. In West Germany the results for vocational training programmes look promising, especially evaluations with respect to the duration of (un)employment. Unfortunately this positive effect does not remain stable if the evaluation is based on other outcome variables. Studies using alternative outcome variables, such as re-employment probability or wages, show a very different picture. An interesting finding is the result from Hujer et al. (1999b), and Hujer and Wellner (2000b) where short (< 6 months) training programmes were found to be more effective than long ones. This may be caused by a negative selection effect, i.e. the unemployed that need comprehensive training to become eligible for a regular job are *a priori* disadvantaged in order to find a regular job. If such a negative selection effect is present it is clearly important to differentiate between different lengths of training programmes in an evaluation. Turning to East Germany we find a positive effect in most studies of vocational training, whereas there are hardly any studies that indicate a negative effect.

In evaluation it should be noted that the choice of the outcome variable seems to be an important issue. Outcome variables like earnings or wages may be questionable if we are evaluating European programmes. This is due to the fact that the welfare state and minimum wage regulations are responsible for distortions between the employment status and the earnings. Therefore, outcome variables that are directly associated with employment status, such as employment probability or the unemployment duration, are preferable for evaluations in Europe.

In this context it might be useful not only to regulate the evaluation by law but also to give some recommendations with respect to the outcome variable which should be used ([28]).

([28]) It should be noted that in Germany regulatory law as in the Social Code III, provides the mandatory use of outcome variables, for example re-employment.

Table 8: **Summary of the empirical findings from microeconometric studies**

Country/Author	Outcome variable	Programme	Result*	Applied method
Austria				
Winte-ebmer (2001)	Employment	Vocational training	(+)	Tobit / IV
Switzerland				
Lalive et al. (2000)	Inflows into employment	Vocational training Job creation schemes	(-) (-) men / (0) women	Duration analysis
Gerfin and Lechner (2000)	Employment rate	Training programmes - basic training programmes - language training - informatics courses - vocational training Job creation schemes	 (-) (-) (0) (0) (-) men / (+) women	 Matching
Belgium				
Cockx et al. (1998)	Employment duration	Training programmes: - in-firm - external Subsidised employment programmes	 (+) (0) (0)	Duration analysis
Cockx and Bardoulat (2000)	Outflow rate from unemployment	Vocational training programmes	(+)	Minimum Chi Squares / IV-estimation
France				
Bonnal et al. (1997)	Outflow rate from unemployment	Vocational training Job creation schemes	(+) (0)	Duration analysis
Brodaty et al. (2001)	Outflow rate from unemployment	Vocational training Job creation schemes	(+) (0)	Duration analysis
Denmark				
Jensen et al. (1999)	Inflows into regular employment	Youth unemployment programme (vocational training)	(0)	Duration analysis
	Inflows into regular vocational education		(+)	
United Kingdom				
Firth et al. (1999)	Inflows into regular employment	Employment training Employment action	(+) (0)	Duration analysis
Payne et al. (1996)	Employment rate	Employment training Employment action	(+) (0)	Matching / selection
Ireland				
O'Connell and McGinnity (1997)	Employment rate	Classroom training On-the-job training Employment subsidies programmes	(+) (+) (0)	Probit

Norway				
Aakvik et al. (2000)	Employment	Training programmes Subsidised employment programmes	(-) (-)	Discrete choice
Sweden				
Johansson and Martinson (2000)	Employment rate	a) Vocational programmes b) Information technology training programmes	(+) b>a	Matching
Larsson (2000)	Employment	Vocational training Job creation schemes	(0) (0)	Matching
Sianesi (2001)	Unemployment	ALMP in general	(+)	Matching
Richardson and van den Berg (2001)	Employment rate	Employment training programme	(+)	Duration analysis
Germany (West)				
Hujer et al. (1998, 1999a)	Unemployment duration	Vocational training	(-)	Duration analysis
Hujer et al. (1999b), Hujer and Wellner (2000b)	Unemployment duration	Vocational training - short courses (< 6 months) - long courses	 (-) (0)	Duration analysis
Hujer et al. (1999c)	Unemployment duration	Off-the-job training Off-the-job training	(0) (-)	Duration analysis
Pannenberg (1995, 1996)	Re-employment probability	Off-the-job training	(+)	Discrete hazard rate, logit and probit estimates
Prey (1997)	Employment probability	Vocational training: - men - women	 (-) (0)	Simultaneous probit
Prey (1999)	Employment probability / wages	Vocational training	(-) / (0)	Simultaneous probit
Staat (1997)	Search duration / Employment probability / wages	Public sector training	(+) /(0) / (+)	Ordered probit, IV-estimation
Germany (East)				
Fitzenberger and Prey (1997)	Employment probability	Training outside the firm Training inside the firm	(+) (-)	Simultaneous probit
Hübler (1998)	Job security / Earnings	Off-the-job training Off-the-job training	(+) (+)	Comparison of: matching / OLS /Logit-ML
Fitzenberger and Prey (2000)	Employment / Earnings	Public financed training	(0) - (+)	Simultaneous probit
Pannenberg (1995, 1996)	Re-employment probability	Off-the-job training Off-the-job training	(+) (+)	Duration analysis
Lechner (1999b)	Income / Unemployment probability	Vocational training	(+) /(0)	Matching

Hübler (1994)	Search activity / Hours of work	Off-the-job training	(+)/(-)	Simultaneous probit
Kraus et al. (1998)	(Stable) Employment	Off-the-job training Off-the-job training	(+) (+)	Multinomial Logit
Bergemann et al. (2000)	Employment	Further training	(0)	Matching
Hujer and Wellner (2000a)	Unemployment duration	Public-sector sponsored vocational training	(0)	Duration analysis
Lechner (1999a)	Employment probability Earnings	Off-the-job training	(0) (0)	Matching
Fitzenberger and Prey (1998)	Employment and wages	Training in the firm Training outside the firm	(-)-(0) (-)-(0)	Random-effects estimation, Matching
Lechner (2000)	Employment probabilities	Public-sector sponsored vocational training	(0)	Matching

* Effect on the outcome variable: (+) positive significant, (-) negative significant, 0 no significant effect.

4. Policy implications: some guidance for evaluation and implementation practice

The last two chapters have presented several different evaluation methods and previous empirical findings for the evaluation of LMT in Germany and the EU. The three evaluation steps discussed (micro- and macroeconometric evaluation as well as CBA) should be seen as additional ingredients to a complete evaluation. Clearly, the first step of every evaluation, and therefore also the most dominant in existing literature, is based at individual level. This is easy to understand if we bring to mind that the most relevant question to be answered after the introduction of the new programme is whether the programme has the desired effects for the participating individual. Other effects, indirect ones on the non-participants or macroeconomic effects, have to be taken into account too, but they are usually assessed after the microeconometric evaluation. This chapter summarises and reviews the findings and aims to give guidance to policy-makers on how to evaluate and implement labour-market programmes. Therefore, focus is on the microeconometric approach, the choice of the estimation method, the problem of heterogeneity, data requirements, the importance of additional macroeconometric and CBA, the design of training programmes and the transferability of the findings to other social programmes.

4.1. Selection problem and the choice of estimation method

The fundamental evaluation problem and the risk of a selection bias are the first things to worry about in the microeconometric evaluation context. We would like to compare the outcome of the participating individual after the programme took place with the hypothetical outcome if they have not participated. Since we cannot observe the same individual simultaneously in both states (participation and non-participation), the fundamental evaluation problem arises and, in order to estimate the true treatment effect, some identifying assumptions are required. We have presented several microeconometric evaluation estimators and the assumptions they impose in Section 2.2. We have also discussed how likely these assumptions are met in reality and showed how these estimators are implemented with a basic numerical example (Section 2.2.7). Of all the approaches presented, the matching estimator seems to be most favourable. The basic idea is to compare individuals who have the same characteristics. Ideally we have statistical twins and the only difference between them is participation in the programme. Therefore, we can interpret the difference in their outcomes as the average treatment effect of the programme. The implementation of the matching estimator relies on the assumption that all observable characteristics that influence the participation decision and the outcome variable can be controlled. Conditional to these characteristics, participation is independent of the outcomes and no selection bias should occur (conditional independence assumption). Thus it can be concluded that the matching estimator fully accounts for selection on observable characteristics. However, to implement it an informative dataset is needed which contains the variables from which they might influence the participation decision and the outcome variable ([29]). If one suspects that unobservable characteristics (like the motivation of an individual) or unobserved variables (some relevant variables, like education or labour-market history are not in the dataset) drive the selection bias, an additional DID procedure might be useful. Even though it is hard to give a general recommendation in this context, it can be concluded that the more informative the dataset the more likely it is that the conditions for the matching estimator are met.

([29]) In Section 4.3 we will discuss which variables that might be.

4.2. Heterogeneity

Another important issue for an evaluation is the problem of heterogeneity. Disregarding heterogeneity might lead to severe problems and biases for the estimation and interpretation of the results. Basically heterogeneity might arise on three levels (individual, programme, regional) which will be discussed later. First, an in-depth estimation of the effects of a training programme on the participating individual should regard the heterogeneity of the participants. For example, it makes a difference if one participant is long-term unemployed and another is unemployed only for a short time. This is because the effects for heterogeneous participants may differ significantly. To proceed with the example we assume that the effect of the programme is positive for the long-term unemployed but negative for the short-term unemployed. Taking the average over both groups would, therefore, lead to a bias estimation of the treatment effect. The positive effect for the long-term unemployed is underestimated and the negative effect for the short-term unemployed is overestimated. Clearly, this is a significant bias which has to be avoided by estimating the effects separately for the heterogeneous groups.

The second type of heterogeneity concerns the programmes. Programmes might differ for example in their duration or content. It is understandable that different programmes might have different effects. Comparing a two-week computer course with a six-month language course may be very misleading. Estimating the effect by taking the average over both programmes leads again to a biased estimate of the treatment effect. Finally, there is also a heterogeneity regarding the geographical distribution of the participants. The importance of local labour markets has been stressed in recent years. Different conditions in these local labour markets might affect the effectiveness of the programmes and the outcomes of participants and non-participants. In the context of the matching estimator, this means that comparing a participant from a booming region with a non-participant from a depressed region is not appropriate. But this might also extend to the programmes since a programme might work in a certain region but not in another one.

Bearing these considerations in mind, it should be emphasised that accounting for all three types of heterogeneity is crucial to obtaining a reliable and differentiated picture of the effects of LMT programmes. To do this certain data requirements have to be fulfilled and this is the topic of our next section.

4.3. Data requirements

The previous discussion has clarified the crucial role that datasets play in an evaluation study. To implement the matching estimator and to take account of the heterogeneity problem an informative dataset is needed. First individual information for participants and non-participants is required. To estimate the effects in heterogeneous subgroups of the treatment population (e.g. old vs. young participants) a sufficient number of treated individuals is needed. To find comparable control group members, the number of non-participants has to exceed the number of participants. Unfortunately there is no rule of thumb on how many control group members are needed to do this. Clearly the more comparable both groups are, the more likely it is that good matches will exist. For example, if the programme is addressed to the long-term unemployed and the control group is drawn out of long-term unemployed, the number of control units does not have to be much higher. On the other hand, if the control group is drawn randomly from the whole population (including short-term unemployed, employed, etc.) good matches are less likely and therefore the number of controls has to be higher. The type of explanatory variables which are needed depend on the programmes evaluated and should be based on a theoretical model (which variables influence the participation decision and the outcome variable?). The suggestions in this context are manifold. Sociodemographic variables like age, gender, education, children, etc., are essential as well as variables of the labour-market status (current job, industrial sector, function, high-skilled/low-skilled occupation). The importance of labour-market history as an explanatory variable for labour-market success today has been stressed in recent years. Therefore 'historical' information is used in most of the studies and is especially important if one suspects that there has been an Ashenfelter's dip. Finally, a good and reliable

outcome measure needs to be available too. The outcome measure should correspond to the aims of the programme (e.g. if the programme is intended to increase the employment prospects of individuals we need the employment status after the programme). It should be traced a sufficient period of time after the programme ends to draw conclusions for the short-, medium- and long-term effects of the programme. The reliability of the outcome measure is very important since the estimation of the effects depends directly on it.

One well-known problem for evaluators is that they are not included in the creation of the datasets used later on for the programme evaluation. In this situation one can either rely on already existing datasets or build up new ones. Using existing datasets, for example surveys like the German Socioeconomic Panel or administrative records collected by governmental agencies, has the major advantage of low costs. There are, however, also several drawbacks which make the interpretation of the results problematic (e.g. Heckman et al., 1999). First, such datasets are not focused on participants in certain programmes, i.e. the number of observations in the treatment group will probably be quite small. Additionally, even if information on treated individuals is available, it is very often difficult to construct an appropriate control group mirroring the treatment group in all characteristics. It may also be difficult to implement the matching estimator if not enough explanatory variables are available. Another possible factor of distortion is given when the outcome variable of interest is not covered in the dataset and has to be approximated. The major disadvantage of building up new datasets (either by collecting a new survey or merging the information from administrative records, etc.) is the high costs. Besides that, the advantages are numerous. The researcher has complete control over the information collected (important variables, outcome variable, etc.) in the group of treated and control individuals. The decisive objective of the programme can be taken into account (e.g. not only the employment status but also the hours worked and/or the wage) and the conditions for the implementation of the appropriate econometric method are more likely to hold. The ideal evaluation dataset is a panel dataset for participants and non-participants that starts some time before the programme begins and that traces the individuals for a certain time period. The necessary length of this time period depends on the views of the policy-makers on how long the effects last.

4.4. Macroeconometric analysis and cost-benefit analysis

Evaluating the impacts of labour-market programmes at an individual level is an important topic but can only serve as a first step in a comprehensive evaluation. This microeconometric approach has to be complemented by two additional steps. The first consists of an assessment of the indirect effects which can emerge at a macroeconomic level. Examples of such indirect spillover effects include dead-weight losses, substitution or displacement effects. Since these effects can counteract the actual goals and objectives of a programme, it is crucial to take them into account. The appropriate method in this context is dynamic panel models as they allow for persistent patterns in the labour market. The data requirements for their implementation are usually substantially lower compared to the microeconometric analysis. Aggregate (either on national or regional) data is required and variables of interest include: (un)employed, programme participants, spending on ALMP and PLMP, vacancies, etc. Particularly interesting in this context are aggregated flow data (from programmes into (un)employment and vice versa), for example to estimate the relationship between the number of unemployed which are put back into work through programmes. An additional methodological problem which has to be taken into account in this context is a potential simultaneity between the spending on labour-market programmes and unemployment. Clearly, in macroeconometric analysis it is harder to take account of the afore-mentioned heterogeneity, since individual characteristics are usually not available.

Once all micro- and macro-economic effects have been evaluated, confrontation between these estimated effects and associated costs is imperative, i.e. the next logical step is to conduct a CBA. We sketched the proceeding for a CBA in one of the previous sections. One of the major problems was a complete gathering of all benefits of labour-market programmes. Surely the effects

which are estimated typically by microeconometric methods, for example improvement in labour-market prospects, are not exhaustive. Instead various additional indirect effects have to be taken into account. Typical examples in this context include the reduction in criminal activities or equity aims. Once this enumeration is finished the next, and even more difficult, step is measuring these effects in monetary terms. If there is no clear guidance from doing so, at least a cost-effectiveness analysis should be conducted. We have given a numerical example for such a cost-effectiveness analysis in Section 2.4.7.

4.5. The design of training programmes: some suggestions

Since the importance of heterogeneity has been stressed in the previous sections, can it be concluded from empirical findings which programme designs are most effective? Naturally this question can only be answered with evaluation studies that compare different types of vocational training programmes. Since macroeconometric evaluations do not distinguish between different types of vocational training programmes, the following statements will rely solely on the review of microeconometric studies. It should be noted, however, that the suggestions we make may not be applicable across all occurrences. The success of every programme also depends on the (labour-market) situation in the implementing regions and therefore a rule of thumb is not available.

The major objective of LMT programmes is to reintegrate the unemployed into regular employment. In order to accomplish this objective the choice between on-the-job and off-the-job training is crucial. A straightforward argument would be that on-the-job training is more favourable, since such a programme does not only provide training but also imparts work experience. Another point might be that the immediate application of the learned knowledge should make it easier for participants to transfer the learned knowledge into their work. Considering the empirical results in respect of this context, no clear picture emerges. The study from Hujer et al. (1999c) suggests a more favourable effect for off-the-job training programmes; that on-the-job training programmes are too specific and only of use in the same firm. In contrast to this finding the majority of studies find that on-the-job training is more effective. The study from Johansson and Martinson (2000) suggests that a narrow contact between participants and firms is beneficial. The subject of the programme should also be considered in this context. General courses, like language or basic computer courses, may work better as off-the job training programmes, but if a training programme is aimed at a specific skill that is narrowly associated with practical work, on-the-job training seems to be more appropriate.

Another question might be how long the duration of a training programme should be in order to be most efficient. As a first guess it could be argued that a longer programme means more comprehensive training. However, participation in a training programme is mostly associated with reduced search activity and therefore the longer programme duration may thwart the effects. The locking-in effect in particular has become a major issue in empirical studies. The problem here is that programmes associated with a full-time engagement do not allow any time for active job search, the benefit of new knowledge is opposed to reduced search activity. In order to avoid this negative effect, vocational training programmes should be designed in a way that enough time remains for active search . Additionally, active search should be actively promoted. Compensation for the participants is relevant in this context too. If the compensation is too high, the incentive to search for a regular job is weakened. Therefore, compensation should lie significantly under the earnings associated with a regular employment.

A further concern associated with training programmes seems to be that, although these programmes are designed for problem groups, an often found phenomenon is that the organising institution selects the participants in a way that the success of the programme is artificially improved. In order to avoid this ‘creaming’ of participants, it could be useful to define the eligibility criteria for a programme participation in a fairly explicit way; the length of the unemployment spell, as well as the basic skill level, should serve as a major criterion. As the Winter-Ebmer study (2001) has shown, vocational training

programmes that are targeted to an explicit group can be quite successful. An additional advantage of explicit targeting is that the content of the training can be adjusted to the needs of the participants in a more efficient way.

A final issue is that, in some European countries, participation in a training programme is essential to preserve entitlement to unemployment benefits. Although this helps to avoid a misuse of the unemployment benefit system, it may be contradictive for the success of the training programme, since training cannot be effective if the participants are forced to participate. If there is the objective to inhibit misuse of the benefit system, job creation schemes seem to be more convenient, since they are of value also for the whole society.

The last point shows that the effectiveness of vocational training programmes does not only depend on the design and implementation of the programmes itself, but also on the general framework of labour-market policy. There is a strong interdependence between the unemployment benefit system, i.e. the passive labour-market policy, and vocational training, i.e. ALMP. Therefore, general recommendations are hard to make and the design has to be adjusted to the local situation.

4.6. Transferability to other social programmes

An important question which remains unanswered regards the transferability of the methods presented to other social programmes. Even though, in principle, transfer is possible, certain restrictions and possible problems will be considered in this section. Although able to reveal the impacts at individual and macroeconomic levels, micro- and macroeconometric evaluation methods do not cover the whole scope of effects associated with social programmes. The advantage of econometric methods lies in assessing the quantitative effects of social programmes. In evaluating LMT, for example, they are able to uncover whether the employment chances of the participants or the (un)employment situation in the economy are affected. Social programmes, however, might also involve various other more qualitative aims and objectives. These goals (e.g. well-being, health, criminal records) are often hard to measure and therefore hard to evaluate in the presented framework. The time horizon of the programme might also be a problem. The matching estimator identifies the true treatment effect by controlling for other characteristics which influence the outcome variable. Naturally, the longer the time horizon of the programme or the evaluation itself, the harder it is to control for all influences. After 10 years, for example, it is reasonable to assume that other factors which have nothing to do with the treatment are influencing the outcome variable. Therefore, the applicability of the estimator is questionable in this situation. Furthermore, all presented estimators have some data requirements (Section 4.3) which have to be met for their implementation. If these requirements are not fulfilled, their application fails. Another point concerns the choice of the control group. In the methodological discussion we introduced the control group as a group of non-participants. This is problematic if no non-participants are available, for example if all the unemployed have to participate in a measure after a certain period of unemployment or if comparison of the effects of different vocational training programmes is required. This problem is not so problematic and the framework can still be applied. The important difference is the different interpretation of the results. For example, if we are interested in comparing two training programmes we can use the participants of the second training group as controls. So, instead of comparing participation in a programme with non-participation we compare the effect of programme A relative to the effect of programme B. If we only have one identical programme in which all individuals have to participate, it might be interesting to compare individuals who participate after three months of unemployment with those who participate after two years of unemployment. This could lead to conclusions on when it is best to direct the unemployed into programmes. To emphasise this point: the control group does not necessarily have to be a group of non-participants but might also be a group of participants in other programmes or a group within the same programme but with other entry characteristics. Attention has to be given, in this case, to interpretation of the effects, since they are no longer measured relative to non-participation.

5. Summary and conclusions

The recent growth in public spending on social programmes, together with tight government budgets, has increased the demand for evaluation of these programmes. In consequence, several evaluation methods and procedures have been developed. With this contribution we have tried to give an overview of the necessary steps in such analysis. The ideal evaluation process can be viewed as a series of three steps. First, the impacts of the programme on the individual should be estimated. Second, it should be examined whether or not the estimated impacts are large enough to yield net social gains. Finally, it should be determined if this is the best outcome that could have been achieved for the money spent (Fay, 1996). For these reasons, we concentrated on evaluation of economic goals with econometric techniques.

The first step of an evaluation is concerned with the individual level, i.e. the researcher has to analyse if the programme under consideration has any effect at all on the individual. This means answering the question of how the participating individual would have behaved under the hypothetical situation of non-participation. Has the individuals who participated in this programme improved in contrast to a situation without participation? The fundamental evaluation problem which arises in this context is due to the fact that we can never observe an individual in both states, i.e. we have a counterfactual situation. We have presented several econometric estimators which simulate this hypothetical outcome. These estimators rest on a number of generally untestable assumptions which were discussed critically to determine their major advantages and disadvantages. We also assessed how likely it is that these assumptions are met in practice. Furthermore, we implemented a numerical example to show how the estimators are implemented technically and also discussed their data requirements.

This microeconometric evaluation, is only a first step towards an overall assessment of the social programme. In this contribution we have also sketched how the next steps should look in an evaluation study. We pointed out the possible existence of indirect and secondary effects which might counteract the actual aims and objectives of the programme. In this context the most important effects are dead-weight losses, substitution and displacement effects. These effects cannot be evaluated at an individual level, making an additional assessment on an aggregate level necessary. We have proposed macroeconometric methods, like augmented matching functions which can be used to take such indirect effects into account. Instead of looking at the effect on individual performance, we would like to know if the ALMP represent a net gain to the whole economy. Important methodological issues in a macroeconometric evaluation are the specification of the empirical model, which should always be based on an appropriate theoretical framework and the simultaneity problem of ALMP which has to be solved. Having identified all potential quantitative effects of a programme, whether they are direct or indirect, the final step consists of augmenting these effects with additional, more qualitative impacts of the programme. These are harder to measure but nevertheless important for an overall assessment. The defined benefits of the programme can be contrasted in a CBA with the costs caused by the programme in order to assess the overall net benefit.

Having discussed the methodological issues related to the three evaluation steps, we presented empirical findings form micro- and macroeconometric evaluations in Europe, with a particular focus on Germany. Training programmes seem to have positive effects on the individuals in most of the studies and perform better than alternative labour-market programmes, such as job creation schemes. Finally, we used the results from our methodological discussion and the empirical findings to draw some implications for policy and evaluation practice. We focused on the choice of the appropriate estimation method, data requirements, the problem of heterogeneity in evaluation

analysis, the successful design of training programmes and the transferability of our findings (for ALMP) to other social programmes. It was our aim to provide the interested reader with the methodological foundations of econometric evaluation analysis and present some easily understandable numerical examples. In doing so, we wanted to give some guidance for the evaluation and implementation practice of ALMP and, to a certain extent, also other social programmes.

List of abbreviations

ALMP	Active labour-market policies
BAE	Before-after estimator
CBA	Cost-benefit analysis
CEA	Cost-effectiveness analysis
CIA	Conditional independence assumption
CSE	Cross-section estimator
DID	Difference-in-differences
EGLMM	East German Labour Market Monitor
EU	European Union
GDP	Gross domestic product
GNP	Gross national product
GSOEP	German socioeconomic panel
JCS	Job creation scheme
KM	Kernel matching
LMT	Labour-market training
LTU	Long-term unemployment
NN	Nearest neighbour
NPV	Net present value
PLMP	Passive labour-market policies
SAM	Social assistance measure
SUTVA	Stable unit treatment value assumption
SwIT	Sweden Information Technology

Annex: Tables

Table A1: **Standardised unemployment rates in the EU (as a percentage of total labour force)**

(%)

	1985[b]	1986[b]	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	Avg.	Rel. Change 1985-2000
Austria	. .	. .	. .	. .	. .	4.0	3.8	3.9	4.4	4.4	4.5	4.0	3.7	4.1	. .
Belgium	12.0	11.3	6.6	6.4	7.1	8.6	9.8	9.7	9.5	9.2	9.3	8.6	6.9	8.9	-43
Denmark	9.0	7.9	7.2	7.9	8.6	9.6	7.7	6.8	6.3	5.3	4.9	4.8	4.4	7.0	-51
Finland	5.0	5.4	3.2	6.6	11.6	16.4	16.7	15.2	14.5	12.6	11.4	10.2	9.7	10.7	94
France	8.3	8.0	8.6	9.1	10.0	11.3	11.8	11.4	11.9	11.8	11.4	10.7	9.3	10.3	12
Germany[a]	9.3	9.0	4.8	4.2	6.6	7.9	8.4	8.2	8.9	9.9	9.3	8.6	7.9	7.9	-15
Greece	. .	. .	. .	. .	. .	10.5	10.9	10.5	10.6	10.4	9.8	9.0	8.1	10.0	. .
Ireland	17.7	18.2	13.4	14.7	15.4	15.6	14.3	12.3	11.7	9.9	7.5	5.6	4.2	12.3	-76
Italy	10.3	11.1	8.9	8.5	8.7	10.1	11.0	11.5	11.5	11.6	11.7	11.2	10.4	10.5	1
Netherlands	13.1	12.1	5.9	5.5	5.3	6.2	6.8	6.6	6.0	4.9	3.8	3.2	2.8	6.3	-79
Portugal	8.6	8.6	4.8	4.2	4.3	5.6	6.9	7.3	7.3	6.8	5.2	4.5	4.1	6.0	-52
Spain	21.9	21.5	16.1	16.2	18.3	22.5	23.9	22.7	22.0	20.6	18.6	15.8	14.0	19.5	-36
Sweden	2.8	2.2	1.7	3.1	5.6	9.1	9.4	8.8	9.6	9.9	8.3	7.2	5.9	6.4	108
UK	11.2	11.2	6.9	8.6	9.8	10.2	9.4	8.5	8.0	6.9	6.2	5.9	5.4	8.3	-52
Average	10.8	10.5	7.3	7.9	9.3	10.5	10.8	10.2	10.2	9.6	8.7	7.8	6.9	9.2	-36

. . Missing values

a) Up to and including 1992, western Germany; subsequent data concern the whole of Germany

b) Data taken from: OECD Economic Survey 86/87 for Italy; OECD Economic Survey 87/88 for Belgium, Denmark, Finland, France, Ireland, Portugal, Spain; OECD Economic Survey 88/89 for Austria, United Kingdom; OECD Economic Survey 89/90 for Greece, Germany, Netherlands; OECD Economic Survey 90/91 for Sweden.

Source: OECD Employment Outlook, 2002

Table A2: **Public expenditure on ALMP as a percentage of GDP**

(%)

	1985	1986	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	Avg.	Rel. Change 1985-2000
Austria	0.27	0.32	0.30	0.34	0.29	0.32	0.35	0.36	0.39	0.45	0.44	0.52	0.51	0.37	89
Belgium	1.31	1.41	1.21	1.18	1.19	1.24	1.34	1.38	1.47	1.23	1.42	1.35	1.30	1.31	-1
Denmark	. .	1.12	1.09	1.27	1.43	1.74	1.74	1.88	1.78	1.66	1.68	1.78	1.56	1.56	39
Finland	0.90	0.92	0.99	1.36	1.77	1.69	1.64	1.54	1.69	1.54	1.39	1.22	0.99	1.36	10
France	0.66	0.74	0.81	0.92	1.04	1.25	1.27	1.29	1.34	1.35	1.33	1.36	1.31	1.13	98
Germany	0.80	0.91	1.04	1.29	1.65	1.58	1.34	1.34	1.43	1.23	1.27	1.31	1.24	1.26	55
Greece	0.17	0.21	0.36	0.40	0.37	0.31	0.30	0.45	0.44	0.35	0.34	. .	. .	0.34	100
Ireland	1.52	1.58	1.44	1.30	. .	. .	1.58	1.68	1.66	. .	. .	. .	. .	1.54	9
Italy	. .	. .	. .	1.42	1.37	1.36	1.35	1.12	1.07	1.00	1.12	1.10	. .	1.21	-23
Netherlands	1.16	1.21	1.24	1.33	1.55	1.59	1.55	1.47	1.51	1.47	1.58	1.62	1.55	1.45	34
Portugal	. .	0.35	0.61	0.70	0.84	0.84	0.67	0.79	0.85	0.77	0.77	0.81	0.61	0.72	74
Spain	0.33	0.62	0.83	0.79	0.55	0.50	0.60	0.80	0.66	0.49	0.70	0.70	0.81	0.64	145
Sweden	2.10	2.01	1.68	2.46	3.07	2.97	2.99	2.36	. .	2.04	1.96	1.81	1.37	2.23	-35
UK	0.75	0.89	0.61	0.56	0.58	0.57	0.53	0.45	0.41	0.37	0.39	0.34	0.37	0.53	-51
Average	0.91	0.95	0.94	1.09	1.21	1.23	1.23	1.21	1.13	1.07	1.11	1.16	1.06	1.12	17

. . Missing values
Source: Employment Outlook (OECD, 2002); Martin, 1998

Table A3: **Public expenditure on PLMP as a percentage of GDP**

(%)

	1985	1986	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	Avg.	Rel. Change 1985-2000
Austria	0.93	0.98	0.95	1.05	1.13	1.41	1.53	1.41	1.39	1.28	1.27	1.19	1.03	1.20	14
Belgium	3.37	3.25	2.59	2.68	2.81	3.00	2.87	2.74	2.77	2.66	2.46	2.45	2.32	2.77	-31
Denmark	. .	3.82	4.26	4.57	4.81	5.33	4.93	4.42	4.15	3.83	3.41	3.13	3.00	4.14	-21
Finland	1.31	1.53	1.13	2.20	3.82	4.88	4.58	3.90	3.61	3.14	2.56	2.33	2.11	2.86	61
France	2.37	2.26	1.84	1.91	1.98	2.07	1.92	1.77	1.79	1.84	1.80	1.75	1.65	1.92	-30
Germany	1.42	1.33	1.10	1.75	1.91	2.52	2.47	2.33	2.49	2.52	2.28	2.13	1.90	2.01	34
Greece	0.35	0.42	0.46	0.50	0.43	0.41	0.43	0.43	0.44	0.49	0.47	. .	. .	0.44	34
Ireland	3.52	3.54	2.68	2.85	. .	. .	2.93	2.71	2.42	. .	. .	. .	. .	2.95	-31
Italy	1.33	1.24	0.84	0.87	1.02	1.15	1.11	0.86	0.87	0.86	0.76	0.68	0.63	0.94	-22
Netherlands	3.49	3.29	2.62	2.62	2.71	3.02	3.26	3.09	3.98	3.03	2.52	2.27	2.05	2.92	-41
Portugal	. .	0.35	0.36	0.44	0.59	0.90	1.06	0.91	0.89	0.83	0.80	0.81	0.90	0.74	157
Spain	2.81	2.65	2.48	2.81	3.05	3.33	3.01	2.31	2.03	1.80	1.56	1.40	1.34	2.35	-52
Sweden	0.87	0.88	0.88	1.65	2.71	2.76	2.53	2.26	. .	2.11	1.96	1.81	1.37	1.82	57
UK	2.12	2.01	0.94	1.36	1.62	1.59	1.39	1.24	1.03	0.82	0.78	0.63	0.56	1.24	-74
Average	1.99	1.97	1.65	1.95	2.20	2.49	2.43	2.17	2.14	1.94	1.74	1.72	1.57	2.02	-21

. . Missing values
Source: Employment Outlook (OECD, 2002); Martin, 1998

Table A4: **Public expenditure on LMT as a percentage of GDP**

(%)

	1985	1986	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	Avg.	Rel. Change 1985-2000
Austria	0.08	0.11	0.10	0.11	0.08	0.10	0.11	0.12	0.15	0.17	0.15	0.19	0.17	0.13	113%
Belgium	0.19	0.21	0.21	0.22	0.23	0.27	0.28	0.28	0.28	0.26	0.25	0.24	0.24	0.25	26%
Denmark	. .	0.43	0.27	0.35	0.39	0.47	0.69	0.98	1.07	0.93	0.97	0.99	0.85	0.72	98%
Finland	0.26	0.27	0.25	0.33	0.46	0.47	0.46	0.44	0.55	0.53	0.44	0.40	0.30	0.42	15%
France	0.26	0.27	0.33	0.35	0.38	0.43	0.41	0.38	0.36	0.34	0.31	0.28	0.25	0.35	-4%
Germany	0.2	0.24	0.38	0.45	0.63	0.55	0.41	0.38	0.45	0.35	0.34	0.35	0.34	0.42	70%
Greece	0.02	0.03	0.16	0.21	0.12	0.08	0.06	0.13	0.09	0.06	0.21	. .	. .	0.12	950%
Ireland	0.63	0.57	0.47	0.24	. .	. .	0.23	0.22	0.21	. .	. .	. .	. .	0.27	-66%
Italy	. .	. .	. .	0.02	0.02	0.01	0.01	0.01	0.01	0.12	0.15	0.12	. .	0.05	583%
Netherlands	0.13	0.19	0.27	0.29	0.39	0.37	0.42	0.39	0.34	0.35	0.31	0.34	0.30	0.34	131%
Portugal	. .	0.18	0.12	0.17	0.26	0.25	0.20	0.23	0.29	0.27	0.29	0.29	0.15	0.23	-17%
Spain	0.02	0.07	0.16	0.17	0.10	0.11	0.22	0.31	0.33	0.15	0.17	0.12	0.15	0.18	650%
Sweden	0.5	0.5	0.53	1.01	1.09	0.76	0.77	0.55	. .	0.42	0.45	0.48	0.31	0.64	-38%
UK	0.07	0.07	0.21	0.15	0.16	0.15	0.14	0.10	0.09	0.07	0.07	0.05	0.05	0.11	-29%
Average	0.21	0.24	0.26	0.29	0.33	0.31	0.32	0.32	0.32	0.31	0.32	0.32	0.28	0.30	32%

. . Missing values
Source: Employment Outlook (OECD, 2002); Martin, 1998

References

Aakvik, A.; Heckman, J. J.; Vytlacil, E. J. *Treatment effects for discrete outcomes when responses to treatment vary among observationally identical person: an application to Norwegian vocational rehabilitation programmes*. Cambridge: National Bureau of Economic Research – NBER, 2000 (Technical Working Paper, 262).

Ashenfelter, O. Estimating the effects of training programmes on earnings. In: *Review of Economics and Statistics*, 1978, Vol. 60, p. 47-57.

Ashenfelter, O.; Card, D. Using the longitudinal structure of earnings to estimate the effects of training programmes. In: *Review of Economics and Statistics*, 1985, Vol. 66, p. 648-660.

Auer, P.; Kruppe, T. *Monitoring of labour market policy in the EU member states*. Berlin: Reihe, 1996 (Serie: Wissenschaftszentrum Berlin für Sozialforschung. Discussion paper FS 1 No 96-202).

Augurzky, B. *Optimal full matching: an application using the NLSY79*. Heidelberg: University of Heidelberg, Department of Economics, 2000. (Discussion Paper, No 310).

Bergemann, A. et al. *Multiple active labor market policy participation in East Germany: an assessment of outcomes*. Halle: Institute of Economic Research, 2000 (Working Paper).

Bijwaard, G.; Ridder, G. *Correcting for selective compliance in a reemployment bonus experiment*. Baltimore: John Hopkins University, 2000 (Working Paper).

Boardman, A. E. et al. *Cost-benefit analysis: concepts and practice*. New Jersey: Prentice Hall, 2001.

Bonnal, L.; Fougère, D.; Sérandon, A. Evaluating the impact of French employment policies on individual labour market histories. In: *Review of Economic Studies*, 1997, p. 683-713.

Brent, R. J. *Applied cost-benefit analysis*. Vermont: Edward Elgar Publishing Limited, 1996.

Brodaty, T.; Crepon, B.; Fougère, D. Using matching estimators to evaluate alternative youth employment programmes: evidence from France, 1986-1988. In: Lechner, M.; Pfeiffer, F. (eds) *Econometric evaluation of labour market policies*. Heidelberg: Physika Verlag, 2001 (ZEW Economic Studies, Vol. 13).

Burtless, G. The case for randomized field trials in economic and policy research. In: *Journal of Economic Perspectives*, 1995, Vol. 9, p. 63-84.

Burtless, G.; Orr, L. Are classical experiments needed for manpower policy? In: *Journal of Human Resources*, 1986, Vol. 21, p. 606-640.

Büttner, T.; Prey, H. Does active labour-market policy affect structural unemployment? An empirical investigation for West Germany regions, 1986-1993. In: *Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 1998, Vol. 188, p. 389-413.

Calmfors, L. Active labour market policy and unemployment: a framework for the analysis of crucial design features. In: *OECD Economic Studies*, 1994, Vol. 22, p. 7-47.

Calmfors, L.; Forslund, A.; Hemström, M. *Does active labour market policy work? Lessons from the Swedish experiences*. Munich: CESifo, 2002 (Working Paper series, No 675).

Calmfors, L.; Skedinger, P. Does active labour market policy increase employment? Theoretical considerations and some empirical evidence from Sweden. In: *Oxford Review of Economic Policy*, 1995, No 11.

Card, D. *Reforming the financial incentives of the welfare system*. Bonn: IZA, 2000 (Discussion Paper, No 172).

Cochrane, W.; Rubin, D. Controlling bias in observational studies. In: *Sankyha*, 1973, Vol. 35, p. 417-446.

Cockx, B.; Bardoulat, I. *Vocational training: does it speed up the transition rate out of unemployment*? Amsterdam: Tinbergen Institute, 2000 (Discussion Paper No 9932).

Cockx, B.; Van Der Linden, B.; Karaa, A. Active labour market policy and job tenure. In: *Oxford Economic Papers*, 1998, Vol. 50, p. 685-708.

Delander, L.; Niklasson, H. Cost-benefit analysis. In: Schmid, G.; O'Reilly, J.; Schönemann, K. (eds) *Handbook of international labour market policy and economics*. Cheltenham: Edward Elgar, 1996.

DiNardo, J.; Fortin, N.; Lemieux, T. Labor market institutions and the distribution of wages, 1973-1992: a semiparametric approach. In: *Econometrica*, 1996, Vol. 64, p. 1001-1045.

Dolton, P.; O'Neill, D. Unemployment duration and the restart effect: some experimental evidence. In: *Economic Journal*, 1996, Vol. 106, p. 387-400.

Fay, R. *Enhancing the effectiveness of active labour market policies: evidence from programme evaluations in OECD countries*. Paris: OECD, 1996 (Labour Market and Social Policy Occasional Papers).

Firth, D.; Payne, C.; Payne, J. Efficacy of programmes for the unemployed: discrete time modelling of duration data from a matched comparison study. In: *Journal of the Royal Statistical Society*, 1999, Vol. 162, p. 111-120.

Fitzenberger, B.; Prey, H. Assessing the impact of training on employment. In: *ifo-Studien*, 1997, Vol. 43, p. 71-116.

Fitzenberger, B.; Prey, H. Evaluating public sector sponsored training in East Germany. In: *Oxford Economic Papers*, 2000, p. 497-520.

Fitzenberger, B.; Prey, H. Beschäftigungs- und Verdienstwirkungen von Weiterbildungsmaßnahmen im ostdeutschen Transformationsprozeß: Eine Methodenkritik. In: Pfeiffer, F. Pohlmeier, W. (eds) *Qualifikation, Weiterbildung und Arbeitsmarkterfolg*. Baden-Baden: Nomos-Verlag, 1998 (ZEW-Wirtschaftsanalysen, Vol 31).

Gerfin, M.; Lechner, M. *Microeconometric evaluation of the active labour market policy in Switzerland*. Mannheim: ZEW, 2000 (Discussion Paper, No 00-24).

Gritz, M. The impact of training on the frequency and duration of employment. In: *Journal of Econometrics*, 1993, Vol. 57, p. 21-51.

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

Heckman, J.; Hotz, J. Choosing among alternative nonexperimental methods for estimating the impact of social programmes: the case of manpower training. In: *Journal of the American Statistical Association*, 1989, Vol. 84, p. 862-880.

Heckman, J. et al. Sources of selection bias in evaluating social programmes: an interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method. In: *Proceedings of the National Academy of Sciences*, 1996, Vol. 93, p. 13416-13420.

Heckman, J. et al. Characterizing selection bias using experimental data. In: *Econometrica*, 1998a, Vol. 66, p. 1017-1098.

Heckman, J.; Ichimura, H.; Todd, P. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. In: *Review of Economic Studies*, 1997, Vol. 64, p. 605-654.

Heckman, J.; Ichimura, H.; Todd, P. Matching as an econometric evaluation estimator. In: *Review of Economic Studies*, 1998b, Vol. 65, p. 261-294.

Heckman, J.; LaLonde, R.; Smith, J. The economics and econometrics of active labor market programmes. In: Ashenfelter, O.; Card, D. (eds) In: *Handbook of Labor Economics Vol. III*. Amsterdam: Elsvier, 1999, p. 1865-2097.

Heckman, J.; Robb, R. Alternative models for evaluating the impact of interventions. In: Heckman, J.; Singer, B. (eds) *Longitudinal analysis of labor market data*, Cambridge: Cambridge University Press, 1985, p. 156-245.

Heckman, J.; Singer, B. A method for minimizing the impact of distributional assumptions in econometric models for duration data. In: *Econometrica*, 1984, Vol. 52, p. 271-320.

Heckman, J.; Smith, J. Assessing the case for social experiments. In: *Journal of Economic Perspectives*, 1995, Vol. 9, p. 85-110.

Holland, P. Statistics and causal inference. In: *Journal of American Statistical Association*, 1986, Vol. 81, p. 945-970.

Holmlund, B.; Linden, J. Job matching, temporary public employment and equilibrium unemployment. In: *Journal of Public Economics*, 1993, Vol. 51, p. 329-343.

Hübler, O. Berufliche Weiterbildung und Umschulung in Ostdeutschland – Erfahrungen und Perspektiven. In: Pfeiffer, F.; Pohlmeier, W. (eds) *Qualifikation, Weiterbildung und Arbeitsmark*terfolg. Baden-Baden: Nomos-Verlag, 1998 (ZEW-Wirtschaftsanalysen, Vol. 31).

Hübler, O. Weiterbildung, Arbeitsplatzsuche und individueller Beschäftigungsumfang – Eine ökonometrische Untersuchung für Ostdeutschland. In: *Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 1994, Vol. 114, p. 419-447.

Hujer, R.; Caliendo, M. Evaluation of active labour market policy – methodological concepts and empirical estimates. In: Becker, I. et al. (eds) *Soziale Sicherung in einer dynamischen Gesellschaft*. Campus Verlag, 2001, p. 583-617.

Hujer, R.; Maurer, K. O.; Wellner, M. Kurz- und langfristige Effekte von Weiterbildungsmaßnahmen auf die Arbeitslosigkeitsdauer in Westdeutschland. In: Pfeiffer, F.; Pohlmeier, W. (eds) *Qualifikation, Weiterbildung und Arbeitsmark*terfolg. Baden Baden: Nomos-Verlag, 1998, p. 197-221 (ZEW-Wirtschaftsanalysen Band 31).

Hujer, R.; Maurer, K. O.; Wellner, M. Estimating the effect of vocational training on unemployment duration in West Germany: a discrete hazard rate model with instrumental variables. In: *Jahrbücher für Nationalökonomie und Statistik*, 1999a, Vol. 218/5+6, p. 619-646.

Hujer, R.; Maurer, K. O.; Wellner, M. The effects of public sector sponsored training on unemployment duration in West Germany: a discrete hazard rate model based on a matched sample. In: *ifo-Studien*, 1999b, Vol. 3, p. 371-410.

Hujer, R.; Maurer, K. O.; Wellner, M. Analyzing the effects of on-the-job vs. off-the-job training on unemployment duration in West Germany. In: Bellmann, L.; Steiner, V. (eds) *Panelanalysen zu Lohnstruktur, Qualifikation und Beschäftigungsdynamik*. IAB – Beiträge zur Arbeitsmarkt- und Berufsforschung 229, 1999c, p. 203-237.

Hujer, R.; Wellner, M. *The effects of public sector sponsored training on individual employment performance in East Germany*. Bonn: IZA, 2000a (Discussion Paper, No 141).

Hujer, R.; Wellner, M. *Berufliche Weiterbildung und individuelle Arbeitslosigkeitsdauer in West- und Ostdeutschland: Eine mikroökonometrische Analyse*, 2000b (Discussion Paper).

Jackman, R.; Pissarides, C.; Savouri, S. Labour market policies and unemployment in the OECD. In: *Economic Policy*, 1990, Vol. 5, p. 450-490.

Jensen, P.; Nielsen, M. S.; Rosholm, M. *The effects of benefit, incentives, and sanctions on youth unemployment*. Arhus: Center of Labour Market and Social Research – CLS, 1999 (Working Paper, No 99-05).

Johansson, P.; Martinson, S. *The effect of increased employer contacts within a labour market training program*. Uppsala: IFAU – Office of Labour Market Policy Evaluation, 2000 (Working Paper, No 10).

Kluve, J.; Lehmann, H.; Schmidt, C. Active labour market policies in Poland: human capital enhancement, stigmatization, or benefit churning? In: *Journal of Comparative Economics*, 1999, Vol. 27, p. 61-89.

Koenker, R.; Bilias, Y. Quantile regression for duration data: a reappraisal of the Pennsylvania reemployment bonus experiments. In: Fitzenberger, B.; Koenker, R. Machado,; J. A. F. (eds) *Studies in empirical economics: economic applications of quantile Regression*, 2002, p. 199-220.

Kraus, F.; Puhani, P. A.; Steiner, V. *Do public works programmes work? Some unpleasant results from the East German experience*. Mannheim: ZEW, 1998 (Discussion Paper, No 98-07).

Kromrey, H. Evaluation – ein vielschichtiges Konzept. In: *Sozialwissenschaften und Berufspraxis*, 2001, Vol. 2, p. 105-132.

Lalive, R.; Van Ours, J. C.; Zweimüller, J. T*he impact of active labor market programmes and benefit entitlement rules on the duration of unemployment*. Bonn: IZA, 2000. (Discussion Paper No 149).

LaLonde, R. Evaluating the econometric evaluations of training programmes with experimental data. In: *The American Economic Review*, 1986, Vol. 76, p. 604-620.

LaLonde, R. The promise of public-sector sponsored training programmes. In: *Journal of Economic Perspectives*, 1995, Vol. 9, p. 149-168.

Larsson, L. *Evaluation of Swedish youth labour market programmes*. Uppsala: IFAU, 2000 (Working Paper, 2000:1).

Layard, R.; Nickell, S.; Jackman, R. *Unemployment: macroeconomic performance and the labour market*. Oxford: Oxford University Press, 1991.

Lechner, M. Mikroökonometrische Evaluationsstudien: Anmerkungen zu Theorie und Praxis. In: Pfeiffer, F.; Pohlmeier, W. (eds) *Qualifikation, Weiterbildung und Arbeitsmarkterfolg*. Baden-Baden: Nomos-Verlag, 1998. (ZEW–Wirtschaftsanalysen, Vol. 31).

Lechner, M. Earnings and employment effects of continuous off-the-job training in East Germany after unification. In: *Journal of Business Economic Statistics*, 1999c, Vol. 17, p. 74-90.

Lechner, M. An evaluation of public sector sponsored continuous vocational training programmes in East Germany. In: *Journal of Human Resources*, 2000a, p. 347-375.

Lechner, M. The effects of enterprise-related continuous vocational training in East Germany on individual employment and earnings. In: *Annals d'Economie et de Statistique*, 1999d, Vol. 55-56, p. 97-128.

Long, D. A.; Mallar, C. D.; Thornton, C. V. B. Evaluating the benefits and costs of the job corps. In: *Journal of Poilicy Analysis*, 1981, Vol. 1(1), p. 55-76.

Lubyova, M.; Van Ours, J. Effects of active labour market programmes on the transition rate from unemployment into regular jobs in the Slovak Republic. In: *Journal of Comparative Economics*, 1999, Vol. 27, p. 90-112.

Martin, J. *What works among active labour market policies: evidence from OECD countries*. Paris: OECD, 1998 (Occasional Papers).

O'Connell, P.; McGinnity, F. What works, who works? The employment and earnings effects of active labour market programmes among young people in Ireland. In: *Work, Employment and Society*, 1997, Vol. 11, No 4, p. 639-661.

OECD. *Economic Survey Italy 1986/87*. Paris: OECD, 1987.

OECD. *Economic Survey Belgium 1987/88*. Paris: OECD, 1988a.

OECD. *Economic Survey Denmark 1987/88*. Paris: OECD, 1988b.

OECD. *Economic Survey Finland 1987/88*. Paris: OECD, 1988c.

OECD. *Economic Survey France 1987/88*. Paris: OECD, 1988d.

OECD. *Economic Survey Ireland 1987/88*. Paris: OECD, 1988e.

OECD. *Economic Survey Portugal 1987/88*. Paris: OECD, 1988f.

OECD. *Economic Survey Spain 1987/88*. Paris: OECD, 1988g.

OECD. *Economic Survey Austria 1988/89*. Paris: OECD, 1989a.

OECD. *Economic Survey United Kingdom 1988/89*. Paris: OECD, 1989b.

OECD. *Economic Survey Greece 1989/90*. Paris: OECD, 1990a.

OECD. *Economic Survey Germany 1989/90*. Paris: OECD, 1990b.

OECD. *Economic Survey Netherlands 1989/90*. Paris: OECD, 1990c.

OECD. *Economic Survey Sweden 1990/91*. Paris: OECD, 1990d.

OECD. *Employment Outlook.* Paris: OECD, 1991.

OECD. *Employment Outlook.* Paris: OECD, 1993.

OECD. *Employment Outlook.* Paris: OECD, 2001.

OECD. *Employment Outlook.* Paris: OECD, 2002.

Pannenberg, M. *Weiterbildungsaktivitäten und Erwerbsbiographie – Eine empirische Analyse für Deutschland*. Frankfurt: Campus-Verlag, 1995.

Pannenberg, M. *Zur Evaluation staatlicher Qualifizierungsmaßnahmen in Ostdeutschland: das Instrument Fortbildung und Umschulung (FuU)*. Halle: Institute for Economic Research, 1996 (Discussion Paper, No 38).

Pannenberg, M.; Schwarze, J. Labour market slack and the wage curve. In: *Economics Letters*, 1998, Vol. 58, p. 351-354.

Payne, J.; Lissenburg, S.; Withe, M. *Employment training and employment action: an evaluation by the matched comparison method*. London: Policy Studies Institute, 1996.

Pohnke, *C. Wirkungs- und Kosten-Nutzen Analysen: Eine Untersuchung von Maßnahmen der aktiven Arbeitsmarktpolitik am Beispiel kommunaler Beschäftigungsprogramme*. Frankfurt: Lang, 2001.

Prey, H. *Wirkungen staatlicher Qualifizierungsmaßnahmen. Eine empirische Untersuchung für die Bundesrepublik Deutschland*. Bern, Stuttgart, Wien: Paul Haupt-Verlag, 1999.

Prey, H. *Wirkungen von Maßnahmen staatlicher Arbeitsmarkt- und Beschäftigungspolitik*. Konstanz: University of Konstanz, 1997 (CILE Discussion Paper, No 45).

Richardson, K.; Van den Berg, G. *Swedish labour market training and the duration of unemployment*, 2001 (Working Paper, forthcoming).

Rosenbaum, P.; Rubin, D. The central role of the propensity score in observational studies for causal effects. In: *Biometrika*, 1983, Vol. 70, p. 41-50.

Rosenbaum, P.; Rubin, D. The bias due to incomplete matching. In: *Biometrics*, 1985a, Vol. 41, p. 103-116.

Rosenbaum, P.; Rubin, D. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. In: *The American Statistican*, 1985b, Vol. 39, p. 33-38.

Roy, A. Some thoughts on the distribution of earnings. In: *Oxford Economic Papers*, 1951, Vol. 3, p. 135-145.

Rubin, D. Estimating causal effects to treatments in randomised and nonrandomised studies. In:

Journal of Educational Psychology, 1974, Vol. 66, p. 688-701.

Rubin, D. Assignment to treatment group on the basis of a covariate. In: *Journal of Educational Studies*, 1977, Vol. 2, p. 1-26.

Rubin, D. Using multivariate matched sampling and regression adjustment to control bias in observational studies. In: *Journal of the American Statistical Association*, 1979, Vol. 74, p. 318-328.

Rubin, D. Comment on Basu, D. Randomization analysis of experimental data: The Fisher randomization test. In: *Journal of the American Statistical Association*, 1980, Vol. 75, p. 591-593.

Schmidt, C. *Knowing what works: the case for rigorous program evaluation*. Bonn: IZA, 1999 (Discussion Paper, No 77).

Schmid, G.; Speckesser, S.; Hilbert, C. Does active labour market policy matter? An aggregate impact analysis for Germany. In: *Labour market policy and unemployment. Evaluation of active measures in France, Germany, The Netherlands, Spain and Sweden*, Cheltenham: Edward Elgar, 2000.

Scriven, M. *Evaluation thesaurus* (Fourth edition). Newbury Park, CA: Sage Publications, 1991.

Sianesi, B. *An evaluation of the active labour market programmes in Sweden*. Uppsala: IFAU, 2001 (Working Paper, 2001:5).

Smith, J. Evaluating active labour market policies: lessons from North America. In: *Mitteilungen aus der Arbeitsmarkt und Berufsforschung, Schwerpunktheft: Erfolgskontrolle aktiver Arbeitsmarktpolitik*, 2000, Vol. 3, p. 345-356.

Smith, J.; Todd, P. *Does matching overcome LaLonde's critique of nonexperimental estimators*. University of Western Ontario, University of Pennsylvania, 2000. (Working Paper).

Staat, M. *Empirische Evaluation von Fortbildung und Umschulung*. Baden-Baden: Nomos-Verlag, 1997(Schriftenreihe des ZEW, 21).

Steiner, V.; Hagen, T. Was kann die Aktive Arbeitsmarktpolitik in Deutschland aus der Evaluationsforschung in anderen europäischen Ländern lernen? In: *Perspektiven der Wirtschaftspolitik 2000*, 2002, Vol. 3, No 2, p. 189-206.

Steiner, V. et al. *Strukturanalyse der Arbeitsmarktentwicklung in den neuen Bundesländern*. Baden-Baden: Nomos Verlag, 1998.

Van Ours, J. *Do active labour market policies help unemployed workers to find and keep regular jobs*? Tilburg: Center for Economic Research, 2000 (Discussion Paper, No 0010).

Winter-Ebmer, R. *Evaluating an innovative redundancy-retraining project: the Austrian Steel Foundation*. London: CEPR – Centre for Economic policy research 2001 (Discussion Paper, DP2776).

Worthen, B. R.; Sanders, J. R.; Fitzpatrick, J. L. *Program evaluation: alternative approaches and practical guidelines* (Second edition). New York: Longman, 1997.

Zweimüller, J.; Winter-Ebmer, R. Manpower training programmes and employment stability. In: *Economica*, 1996, Vol. 63, p. 113-130.