



EVALUATING ALMP EVALUATIONS

JAAP DE KONING

YURI PEERS

SEOR WORKING PAPER NO. 2007/2

Rotterdam, November 2007

<i>Contact person</i>	Jaap de Koning
<i>Address</i>	SEOR, Erasmus University Rotterdam P.O. Box 1738 3000 DR ROTTERDAM
<i>Telephone</i>	+31-10-4082220
<i>Fax</i>	+31-10-4089650
<i>E-mail</i>	dekoning@few.eur.nl
<i>Website</i>	www.seor.nl

SUMMARY

In this paper we carry out a meta evaluation of the international evaluation literature. Many attempts have been made to estimate the net impact of reintegration on the individual job entry chance. So, the question is what conclusions can be drawn from the literature. How big is the net effect of reintegration measures? Our paper differs from earlier meta analyses of ALMP impact evaluations in the fact that we try to make inferences about the *size* of the net effect. To that end we analyze the size distribution of the net impact estimates resulting from the international evaluation literature. In our analysis we distinguish between different types of reintegration measures. Furthermore, we carry our regression analyses in which we explain the estimated effect found in individual studies from the type of measure, the labour market context and characteristics of the evaluation method used.

A problem with the analysis is that different studies measure different things. Using the theoretical framework of hazard models, we analyze to what extent the different approaches can be related to each other and whether it makes sense to combine the different studies in one meta analysis.

The results indicate that on average the net impact of reintegration is fairly small. As a result of reintegration job entry chances are probably not increased by more than 3 per cent points on average. The case is most convincing for training and counselling with average net effects ranging from 5,7 to 9,7 per cent points. The positive results for training are surprising. However, our sample hardly contains experimental evaluations of training, which are usually seen as the most reliable evaluations. Overall we find that the net effect estimate tends to be smaller if an experiment is used for the evaluation. Hence, the results for training may be too rosy. If we account for the method used training comes out less well, while sanctions (which are treated as one type of measures) perform better. Wage subsidies and job creation come out poorly. We also find that the net impact of ALMPs depends on the socio-economic situation: it tends to be lower when unemployment is low than during a recession period.

We see this paper as a first step and want to exploit the literature further. Recently, a lot of studies using a more developed methodology for non-experimental evaluation have appeared. It will be interesting to see to what extent inclusion of these studies in the meta analysis will alter the results.

CONTENT

Summary		i
1	Introduction	1
2	Methods and what they measure	3
3	The meta-analysis	9
4	Conclusions and final remarks	18
References		20

1 INTRODUCTION¹

In this paper we analyse the outcomes of the international evaluation literature on active labour market policies (ALMPs). The latter consist of various measures for the unemployed like job counselling and training to increase their job entry chances. During the past decades many evaluation studies have been carried out in various countries that try to measure the net impact of these measures on job entry. The net impact is the difference between the job entry chance of a person after participating in a measure and the job entry chance of the same person if he had not participated. Given the social importance of these measures and the large quantity of studies, an obvious question is what we can learn from this vast amount of studies. Do active measures have a positive effect on the transition probability from unemployment to employment? And if so, how big is this effect? Does it differ between different measures, between different social groups, etc? These are the questions we try to answer in this paper.

Although earlier review studies (Heckman, Lalonde and Smith (1999) and Kluve and Schmidt (2002)) exist, these studies do not contain a systematic analysis of the outcomes. De Koning et al (2005) made a first attempt to present the outcomes in a systematic way by counting a) the number of studies with a significant positive effect, b) the number of studies with mixed outcomes, and c) the number of studies showing insignificant or significantly negative effects. This is done for the different ALMP measures separately. Kluve (2006) goes one step further by carrying out an ordered regression on the basis of such data. This allows him to take several context variables like the unemployment situation into account. Both studies come to similar conclusions with respect to the relative performance of the various types of measures. Job counselling, sanctions and intensified job search control and subsidies for job entry appear to be most effective. Job creation schemes perform poorly, while training holds an intermediate position. According to Kluve context variables do not play a big role. He finds some indications however that the net effect is higher in situations of high unemployment than in situations of low unemployment.

Although these studies give some indications about the effectiveness of ALMPs, they do not answer the most important question, namely: how big is the net impact of ALMPs on job entry chances? In this paper we concentrate on exactly this question. To that end we analyse the size distribution of the net impact estimates emerging from the evaluation literature. Most studies are taken from the previously mentioned review studies. Other studies included in the analyses are: Blundell et al (2004), Heckman, Smith and Clements (1997), Kluve (2002) and Van Rheezen (2003).

We do this both for all ALMPs taken together and for the different ALMPs individually. Furthermore, we carry out regression analyses in which we explain the net impact estimates found in individual studies from factors like the sample size, the evaluation method used, the economic and labour market context, etc. In view of the limited number of observations this is only done for the whole sample taken together. In these analyses possible differences in effectiveness between the different ALMPs are accounted for by including dummy variables representing the latter. And even then only a limited number of variables can enter the equation at the same time owing to the limited sample size.

In judging ALMPs the net effect on individual job entry chances is a key aspect, but it is certainly not the only relevant aspect. First, costs are also relevant.

¹ The authors thank Eelco Kappe for his contribution at an early stage of the study.

According to the results found by de Koning et al and Kluge the most effective measures also seem to be the cheapest ones. However, we may not come to the same conclusion when we analyse the size of the impact.

Second, comparing measures on the basis of short-term effects as we do in this paper may not be completely fair with regard to training and job creation schemes. We only look at impacts during the unemployment spell in which the intervention took place. Vocational training may take that long that it does not reduce the length of the unemployment period in which the training takes place or even prolongs it. However, when the training results in a significant increase in skills levels, it may reduce the chance that the people concerned will become unemployed again in the future. Therefore, long-term benefits could be considerable. It should be noted, however, that although this may seem likely, there is not much empirical evidence to support it. The few studies dealing with long-term effects of training show mixed outcomes (De Koning et al, 2005).

Also with respect to job creation schemes one should be cautious in drawing conclusions from existing evaluation studies. Most of the literature concerning job creation deals with the transition from unemployment to regular employment, where people are counted as unemployed while being in the scheme. However, an important objective of job creation schemes may be to offer employment for people unlikely to find regular employment in the labour market. As the transition probability to regular employment is small for this group, such schemes may be useful even if they slightly reduce this transition probability (a result often reported in the literature).

Third, indirect effects and aggregate effects might be important. The improved chances of the participants may, for example, go at the cost of longer job search spells for other job-seekers. This would imply that the aggregate effects of active policies are less favourable than the effects for the participants. However, a review of aggregate impact studies shows that aggregate outcomes are not too different from results produced by micro evaluation studies (De Koning, 2001).

Finally, it might be important to take non-market effects into consideration. By non-market effects we understand the effects on health and social participation. A recent review study by Gelderblom and de Koning (2007) provides evidence from various studies that long-term unemployment increases the chance of health problems and social isolation. Hence, ALMPs that succeed in reducing unemployment duration may entail savings in health costs and social costs. In an integral cost-benefit analysis of ALMPs these savings should of course be taken into account.

While acknowledging the limitation of just looking at short-term effects of ALMPs on the job entry chances of ALMP participants, we still think that summarizing the evidence from the international literature on this point is an important step in increasing our understanding of the effectiveness of these measures.

The paper is structured as follows. In section 2 we discuss the methods that have been used in the literature to assess the net impact of active policies on individual unemployment duration and job entry chances and how the different methods can be related to each other. This is important as we want to use the outcomes of the studies as observations for regression analyses. Hence, we want to use as many studies as possible. In section 3 the methodology of the meta evaluation is explained and the results presented. Finally, section 4 contains the conclusions and some final comments.

2 METHODS AND WHAT THEY MEASURE

Methods to measure net impacts

The most important question evaluations of active labour market measures should answer is whether a person participating in a measure gets a higher chance to find a job as a result of it. The fundamental problem posed by this question is that by definition we cannot observe what would have happened to a participant, if he or she had not participated. Therefore it is the task of the evaluator to find a convincing estimate for the latter (the so-called counterfactual).

Experiments are seen as the most reliable method to construct the counterfactual. They are based on random assignment in two steps. In the first step a random sample is taken from the group that is entitled to a measure. Then, in the second step, this sample is divided randomly over two groups: the experimental group, which gets the treatment, and the control group, which is not treated. Consequently, the two groups are monitored over a sufficiently long period to judge whether the participants have a higher job placement rate than the control group. The random assignment guarantees that the participants are representative for the whole target group and that the experimental group and the control group have the same composition (even with respect to unobserved characteristics). Hence, the difference in placement rates between the two groups can be attributed to the measure. Furthermore, the estimate for the net impact applies not only to the participants but to the target group as a whole.

This, however, is only true for an ideal experiment. In reality, experiments often do not fully meet the requirements for such an ideal experiment. One can argue that the implementation of a measure during an experiment will usually differ from its full-scale operation. Another point is that as far as the monitoring is done by surveys, some of the participants and the control group may not be willing to answer the questionnaires. This non-response is then likely to be selective. Obviously, one cannot simply assume that the bias resulting from this selectivity is the same for the two groups. Heckman and Smith (1996) mention a number of factors implying that in practice experiments can suffer from biased outcomes. The methods that can be used to correct for this bias are in fact the same as the ones that have been developed for non-experimental data.

However, before we turn to non-experimental methods we want to stress that although experiments have their limitations, they are still more reliable than the former methods. In that respect it is a pity that there is so little room for experiments in most European countries. Two factors seem to be in play. The first factor is the fact that it is often seen as unethical to exclude people (the control persons) from treatment. However, this argument is questionable. Is it ethical to apply a measure on a wide scale to unemployed people without knowing that it has favourable effects and with the possibility that it has adverse effects (prolonging unemployment duration rather than reducing it)? If an experiment can give reasonable certainty about the net effect of a measure than this seems to be a valid reason to temporarily exclude some people from the measure. The second factor is of a judicial nature: legislation in some countries seems to prohibit such an exclusion.

In the case of a non-experimental evaluation one can try to select the control group in such a way that for every participant a non-participant is selected who resembles the participant closely. So, we can choose someone of the same gender, the same age, the same education, etc. Then we can again compare the two groups as to their job placement rates. This method is known as the **matching method**. The problem with this method, however, is that we can never be sure that we control for all the relevant factors. Furthermore, it is often likely that some of the unobserved factors influences both the selection process and the chance of finding a job. Suppose, that we do not observe motivation. More motivated unemployed will

tend to have a higher job placement probability than those with less motivation. But more motivated unemployed are also more likely to enrol in a program. They will show a higher interest in participation. Furthermore, program managers will prefer higher to lower motivated participants. If we then find that the job placement rate among the participants is higher compared to the controls, it says nothing about the net impact of the program. The difference in placement rates may be wholly due to the difference in motivation².

Because matching methods deal not completely satisfactorily with the selectivity bias problem, attempts have been made to deal with the problem by using econometric methods. Within the framework of the **timing of events method** Abbring and Van den Berg (2003) have developed such an approach. They consider two random processes:

- the time that expires between entry into unemployment and entry into a program (the length of the unemployment spell up to enrolment in a program);
- the time that expires between entry into unemployment and the outflow from it (the full length of the unemployment spell).

For both of these random processes a hazard model is used. In most cases it is assumed that the program does not take time and can be seen as a point in time³. The impact of the program is then reflected by a change in the unemployment hazard from this point in time onward. Both spells will depend on observed and unobserved personal characteristics. As far as the same unobserved characteristics affect both spells, the latter will be correlated. By taking this correlation into account the selectivity problem is tackled. The hazard approach will be treated in more detail in the next section.

What do the various methods measure?

Practically all experimental studies and matching studies in our sample compare the participant group with a control group on the basis of the cumulative job entry chance⁴. The latter is defined as the percentages of the people in both groups that have found a job between the moment of entry of the participants into the program and some point in time after they left the program. So, these studies show to what degree participants have a higher chance of finding a job compared to non-participants with the same characteristics. In the remaining part of the paper we will also refer to these studies as E&M (experimental and matching) studies. Most other studies use the timings of events method (also to be referred to as TOE studies) and measure the effect of a measure on the hazard rate.

² Within the context of the matching method the problem has been dealt with by looking at the difference in placement probabilities before and after the treatment. Then these differences are compared between the two groups. This is the so-called **'difference-in-differences' method**. This method takes care of the selection problem as far as this problem is caused by unobserved factors that do not change over time. In principle, it is also possible within the framework of the matching approach to control for unobserved factors that change over time, but not in an entirely satisfactory way.

³ Some studies take the moment of exit from the program as the point from which on the hazard changes or try both options.

⁴ Some studies compare the participant group with a control group on the basis of the percentage that has a job on a specific point in time after the participants left the program. This situation is dealt with in the annex to the paper. Because most studies that report this percentage also report the cumulative job probability, we do not distinguish the former as a separate category in our analysis.

Since the hazard reflects the outflow rate from unemployment one might be inclined to think that the latter studies measure the same thing as the former studies. However, as we will show below this is not the case.

Why is it important to compare these two categories of studies (E&M studies versus TOE studies)? First, combining the two categories in one sample is important because of the relatively small sample size that we dispose of. After splitting up the sample we are left with two sub samples that are hardly large enough to allow us making reliable inferences for the different ALMPs. But even if the latter was possible, it would be important to relate the outcomes of the two sub samples to each other. Do they point to the same conclusion concerning the net effect or to a different conclusion? Ending up with two conclusions drawn from two sub samples of studies is not entirely satisfactory.

A comparison between the two categories of studies is not straight-forward. Experiments and matching methods do not make use of models while the TOE method does use a model for the assessment. So, in order to show the differences we have to use a theoretical mathematical model that encompasses both approaches. At first instance we use a simple model hazard model based on the assumption that the unemployment duration of an individual spell is a random variable from an exponential distribution. Later we discuss the implications of relaxing some of the assumptions of this model.

Starting from an exponential duration model the probability density function of unemployment duration x is equal to:

$$f(x) = \beta \exp(-\beta x) \quad (1)$$

In equation (1) f is the density function and β is the hazard. The latter is defined as the chance of escaping unemployment in an infinitely small time interval (where P refers to the probability distribution associated with f):

$$\lim_{h \downarrow 0} \frac{P(t \leq x \leq t + h; x \geq t)}{h} = \beta \quad (2)$$

Often evaluators assume that the program they are evaluating leads to a constant percentage change of the hazard:

$$\beta^* = \beta \exp(\gamma) \quad x \geq t_0 \quad (3)$$

Where γ denotes the effect of the program on the hazard rate and t_0 the time of the intervention. Studies using the TOE method estimate γ using econometric techniques and often only report the estimate for γ (as well as the estimators for the other explanatory variables entering the hazard model)⁵.

Now the question is what E&M studies are measuring and whether this can be related to γ . Only then both categories can be directly compared. Let us assume again that the intervention

⁵ Note that γ is not equal to the effect on the mean of the unemployment duration distribution. This effect also depends on the timing of the intervention during the unemployment spell.

took place at time t_0 . Both participants and controls were unemployed at the time. The probability that a control person who was unemployed at time t_0 found a job within a period of length d is equal to:

$$P^c = P(t_0 \leq x^c \leq t_0 + d; x \geq t_0) = 1 - \exp(-\beta d) \quad \text{for a control person} \quad (4)$$

and

$$P^d = P(t_0 \leq x^d \leq t_0 + d; x \geq t_0) = 1 - \exp(-\beta^* d) \quad \text{for a participant} \quad (4)^*$$

From (4) and (4)* we can deduce that:

$$\gamma = \log \left\{ \frac{\log(1 - P^d)}{\log(1 - P^c)} \right\} \quad (5)$$

An important limitation of equation (5) in view of our intention to combine the two categories of models in one empirical analysis is that it does not define γ as a function of the *difference* between P^d en P^c . The problem is that E&M studies usually present their results in terms of this difference. Often P^d and P^c are not given individually. This would mean that E&M studies that report the difference between the two probabilities and TOE studies that only provide an estimate for γ cannot be compared and cannot be combined in one meta analysis. However, there might still be a way out of this problem if it were possible to approximate (5) by a relatively simple function in the difference between P^d en P^c . In order to explore the relationship between γ and $P^d - P^c$ we did the following:

- we took all combinations of the two probabilities in the range between 0,02 and 0,98 with 0,02 intervals (so, we took 0,02, 0,04, 0,06, etc.);
- for each combination we computed both γ and $P^d - P^c$.

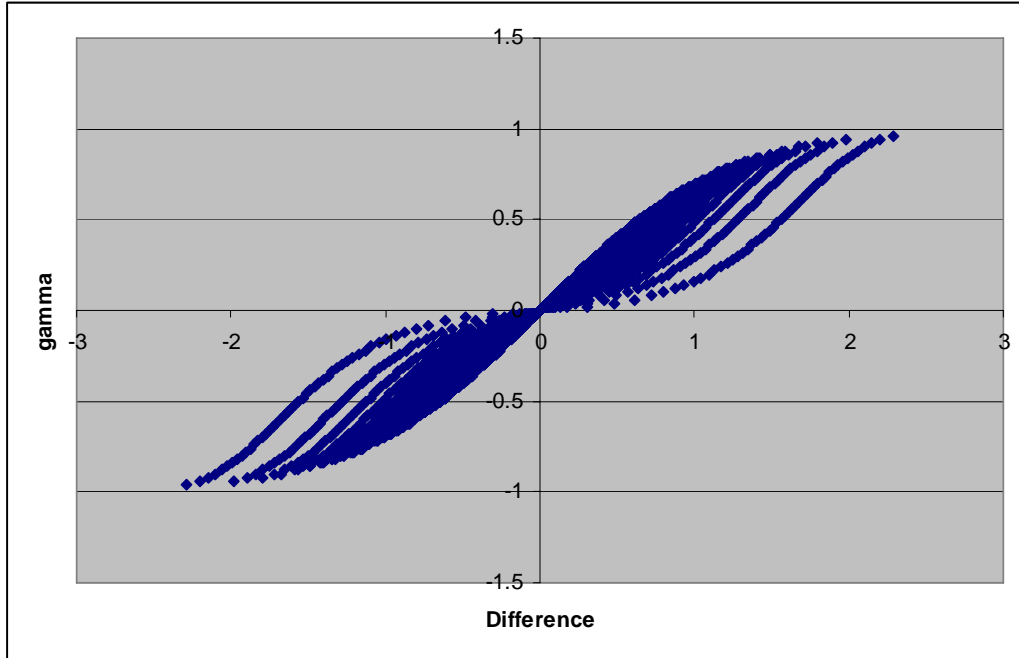
Figure 1 contains the data. Clearly, there is a positive relationship between γ and $P^d - P^c$. This relationship can be approximated by a linear equation fairly well. A linear regression gives the following results:

$$\gamma = 4.5E - 17 + 1.71(P^d - P^c)$$

(0.004) (0.098)

$$\bar{R}^2 = 0.93; N = 2401$$

Figure 1 Scatter diagram of the simulated values for γ (gamma) and $P^d - P^c$ (difference).



Where the figures between brackets denote the standard errors. So, we find a very high adjusted R-squared. The constant term is practically zero. This suggests that for a wide range of values γ and $P^d - P^c$ are indeed almost proportional. We only run into trouble with values of the probabilities very near to zero and very near to 1. However, such values are not realistic.

One might also argue that even the fact that we derived a relationship between γ on the one hand and $P^d - P^c$ on the other hand does not mean a lot as we derived it from a very simple duration model which assumes that apart from the ALMP intervention the hazard rate is a constant. However, it is important to note that equation (5) holds for a much wider class of hazard models, namely the mixed-proportional hazard models that are most often used in evaluation studies. The latter class of models allows the hazard to depend on time and on unemployment duration (taking duration-dependence of the hazard into account, for example). For other duration models equation (5) holds in specific cases only. In the case of the Accelerated Lifetime Model (ALM) for example, where the explanatory variables scale the time rather than the hazard, equation (5) only holds for the exponential baseline hazard. So, equation (5) seems to apply to a fairly wide range of hazard models.

Other assumptions made concerning the hazard model may be more restrictive. One of these assumptions is that the effect remains constant after the intervention. Bijwaard (2002) estimates both a model with a time invariant effect of the intervention and a model where the impact may differ between the first ten time periods after the intervention and the subsequent periods. The outcomes of the latter model suggests that the initial effect is higher than the long-term effect.

A second assumption that is often made in TOE studies but seems to be questionable is that program participation does not take time. In reality, participants in active labour market measures often stay in these measures for some time. Training courses, for example, may take several months or even longer. During participation the hazard rate will be different than before⁶. A priori it is difficult to say whether it is higher or lower. In case of a training program it is probably lower. People will usually participate in a training program to increase their human capital under the assumption that this improves their chances in the labour market. Hence, they will reduce their job search intensity and concentrate on the training. However, other types of programs may have the opposite effect. Examples are the so-called Work First programs that have become popular in active labour market policy. Often an important element of these programs is that the unemployed have to carry out simple work in order to obtain their benefit. For those with higher qualifications being obliged to do such work will encourage them to look for a regular job.

Relaxation of these assumptions will make equation (5) even more complicated, which further undermines the case for a common analysis. The fact that TOE studies often make use of assumptions that do not seem plausible also implies that one might question the validity of the outcomes of many of these studies. A study that treats the selectivity bias problem in a satisfactory way but makes unrealistic assumptions on other points may still produce unreliable outcomes. Non-experimental matching studies may suffer from the selectivity bias problem but are not subject to the other problems because they do not work with econometric models. So, the TOE studies in our sample may not be superior to the latter studies. May be this is different for recent TOE studies that start from more general assumptions.

Conclusion

Experimental evaluations and matching (E&M) studies often only report the difference between the job placement probabilities of participants and non-participants, while most studies using the timing of events method (TOE studies) only report the effect on the hazard rate. It appears that for a fairly general class of models (the mixed-proportional hazard model and even specific variants of even more general models) a direct relationship can be derived between the program effect on the hazard rate on the one hand and the job placement probabilities of the participants and the non-participants on the other hand. However, this does not define a relationship between the effect on the hazard rate and the **difference** between the latter probabilities. Unfortunately many E&M studies only report this difference and not the underlying probabilities. Fortunately, for a wide range of values this relationship is almost proportional, making it relatively easy to use outcomes from both types of studies in one analysis.

Different methodologies are used in the literature. Experimental studies and matching studies compare participants with a control group of look-a-likes. These studies are not based on econometric models and thus do not rely on specific model assumptions. A disadvantage of the matching method is that it does not deal with the selection bias problem in a satisfactory way. TOE studies are a second class of studies using non-experimental methods. A major advantage of TOE models is that they allow for a satisfactory treatment of the selectivity bias

⁶ It is even possible that the hazard rate changes before entry into a program, once people know that they are going to participate. Also here the effect can go in two directions. The prospect of participating in a program may imply that job search intensity is already reduced, if the program is seen as positive. This effect is also referred to as the Ashenfelter dip. However, if a client has a negative attitude towards the program he may try to find a job before entry into the program.

problem. However, being based on econometric models the TOE studies in our sample often make simplifying assumptions. The most common assumption is that the impact of the intervention is constant after the intervention, which does not seem to be realistic. So, the TOE studies in our sample may not be superior to the matching studies contained in it. May be this is different for the younger generation of TOE studies. If the outcomes of TOE studies are unreliable owing to unrealistic assumptions, this would distort the comparability with M&M studies. So, it is an empirical matter whether the outcomes of TOE studies and M&M studies can be combined in one meta analysis.

3 THE META-ANALYSIS

Methods

Plotting the size distribution of the net impact estimates from the available studies is an obvious way to get an impression of what these studies tell us. Furthermore, statistics of this distribution like the mean and the standard deviation are informative from this perspective. However, as we will see the variation in the estimates is considerable, even we look at each ALMP individually. Hence, it is interesting to look more closely at the possible sources of this variation. As we have shown in the previous section the method used for the evaluation is an obvious source for this variation, but there are many more like contextual differences (the studies in our sample refer to different countries and to different periods in time). Given our limited sample size it is not possible to deal with all these aspects in a descriptive way. An obvious way to deal with this problem then is to control for the various factors by conducting a regression analysis in which we use the outcomes of net impact studies as observations and the factors mentioned as explanatory variables. So we have:

$$NI_i = X_i\beta + \varepsilon_i \tag{6}$$

where NI denotes a net impact estimate, X stands for a set of factors influencing the size of the estimate and i is an index for the study. Furthermore, β is vector with unknown parameters and ε an error term.

The vector x may, in principle, contain the following types of factors:

- a. the type of measure: some measures may be more effective than others;
- b. design features of a given measure: training measures, for example, can take different forms as to duration, level and field;
- c. characteristics of the participants: a given measure may be more effective for some groups than for others;
- d. features of the implementation system: a system in which government agencies implement the interventions may produce different results than a system in which implementation is outsourced to private companies that have to compete for the government contracts;
- e. features of the implementation strategy: a given method may be implemented in different ways, with a different quality of the staff involved, etc.;
- f. characteristics of the evaluation method used: some methods (particularly non-experimental methods that do not correct for selection bias) may be less reliable and may systematically produce too optimistic results about the net impacts;
- g. the number of observation used: estimates are more precise when they are based on more observations;

- h. the economic context: net effects may be different in a situation of low unemployment and high economic growth compared to a situation of high unemployment and low economic growth.
- i. The institutional context: the net effects of interventions may, for example depend on the financial incentives for unemployed jobseekers to actively search for jobs and to accept jobs even if the latter are less attractive.

It is clear from this list that many factors may be of influence. Potentially, assessing the effect of some of these factors on the net impact is of high policy relevance. This is particularly true for the effect of design features and implementation strategies on net impact, because that type of information would provide recommendations about what could be done to improve ALMPs. However, at the moment we only have information about a more limited set of variables only. Collecting data on design and implementation of individual programs might be possible but would require a considerable effort. En even then the data will probably be far from complete.

An aspect not included, but highly important is the destination after leaving unemployment. It may be outflow to any destination or only outflow to a job. The studies in our sample are often unclear about this, although it will probably matter to the outcomes.

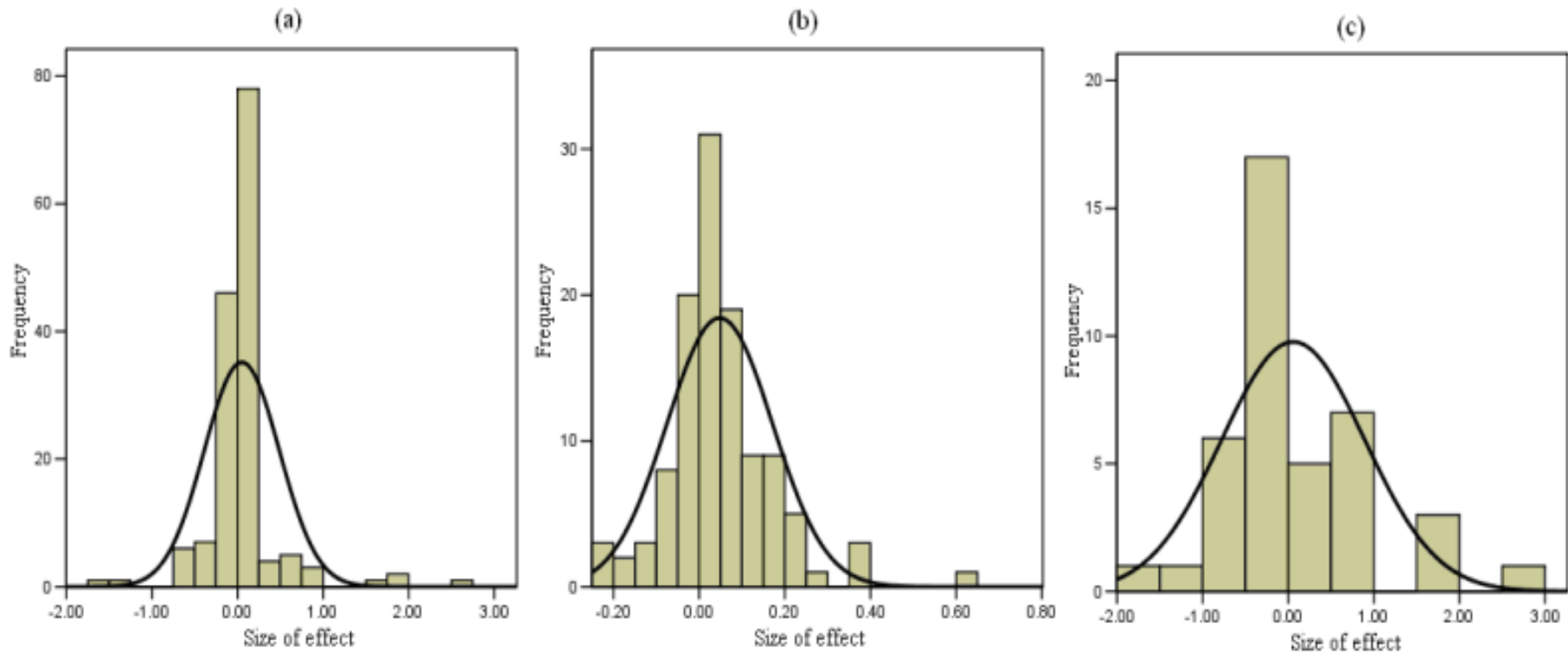
Data and descriptive analyses

Our sample contains 28 articles and papers. Since some of these articles and papers report results for more than one participant group or for more than one type of measure, the total number of cases is much bigger, namely 155. Of these 155 cases, 73 cases report a positive effect, 43 a negative effect and 39 a non significant effect. Table 1 gives a summary per type of measure.

Table 1 *Breakdown of the available cases according to the sign of the effect*

Measure	Sign (number of cases)		
	Positive	Negative	Not significant
Training	41	12	13
Counselling	10	1	3
Sanctions	10	3	-
Placement subsidies	2	7	8
Job creation	2	7	13
Other	8	13	2
Total	73	43	39

Figure 2 Size of the effect using: (a) all observations (155), (b) observations except TOE (114), (c) observations TOE (41)



Three figures are depicted that show the size distribution of the net effects. Figure 2a contains all the cases. From this figure we can conclude that predominantly on the positive side there are a number of considerable outliers. From the figures 2b and 2c we can conclude that the outliers primarily come from the applications of the TOE method. The figures suggest that the mean value of the net effect is close to zero.

Table 2 contains a number of statistics with regard to the size distribution of the effect (minimum, maximum, average and standard deviation). Based on all observations the mean net effect is 0,051, which means that on average the job placement rate for participants is 5,1 per cent points higher than for non-participants. However, some of the TOE studies report extreme values, which cannot be interpreted within the same framework as the E&M studies. This leads to an extreme high standard deviation. Based on E&M studies only (the sub sample emerging after leaving out the TOE observations) we find a mean net effect of 0,035 or 3,5 percent points. But even here the standard deviation is considerable, implying a 95 per cent confidence interval ranging from minus 12 per cent to plus 19 percent.

If we look at the various ALMPs individually we can observe that the mean net effect is positive for training, counselling and sanctions. The means for training and counselling are not affected that much if we leave out the TOE observations. For sanctions we do not have TOE studies. The difference in mean outcomes for the whole sample compared to the sub sample based on E&M studies is somewhat larger for subsidies and job creation, although for both measures the mean net effect is negative in both cases. For these measures TOE measures give a more pessimistic picture than E&M studies do. Finally, we have a category of other measures, which is highly heterogeneous. Hence, it is more difficult to interpret the results for this category.

If we compare these results with the ones earlier obtained by de Koning et al and Kluge there is one major difference. In these studies training comes out relatively poorly, while it performs relatively well in our results. Subsidies, on the other hand, are performing worse in our case compared to the previously mentioned studies. One should add to that, however, that the latter measure shows the biggest standard deviation in our case (although this statistic is also very high for training). It is also important to note that our sample is not the same as the ones used by de Koning et al and by Kluge.

It is clear that the variation in the outcomes is considerable, even if we leave out the TOE observations and look at the various ALMPs individually. Hence, it would be important to see whether we can explain at least some of this variation from characteristics of the various studies.

Table 2 Descriptive statistics with respect to the various measures

Measure	Based on all 155 observations					Without TOE observations				
	Number	Minimum.	Maximum.	Average	Standard deviation	Number	Minimum	Maximum	Average	Standard deviation
Training	66	-1.75	2.57	.109	.528	52	-.20	.39	.092	.117
Counselling	14	-.10	.62	.130	.197	9	.03	.62	.120	.187
Sanctions	13	-.01	.13	.034	.040	13	-.01	.13	.034	.040
Subsidies	17	-1.43	1.65	-.056	.575	8	-.24	.17	-.004	.153
Job creation	22	-.74	.85	-.084	.392	12	-.13	.08	-.027	.056
Other	23	-.25	.92	0.051	.278	20	-.25	.12	-.029	.089
Total	155	-1.75	2.57	.051	.406	114	-.25	.62	.035	.077

Estimation results

In order to compare our data with the data used by Kluge (2006) we have started by carrying out the same type of analysis as he did, namely an ordered logit analysis in which the dependent variable indicates three possible situations: 1) the net effect is negative and significant, 2) it is insignificant or 3) it is positive and significant. The results are given in table 3. In the analysis ‘other’ measures are taken as the reference category. According to the results training, counselling and sanctions come out relatively well, while placement subsidies and job creation seem to be less effective. The unemployment rate has a significant influence: the lower the unemployment rate is, the lower the net effect. A possible explanation for this result is that in situations of low unemployment many unemployed will find a job anyhow. GDP growth is not significant in the equation.

So, also in this regression we can observe the two differences with Kluge’s outcomes that were pointed out in the previous sub section: training comes out more positively in our analysis while subsidies perform less well according to our results.

Table 3 *Ordered logit model*

Dependent variable :sign of the effect	
Number of observations: 152	
Explanatory variables	Coefficient (standard error)
Dummy Training	1.443 (0.531)*
Dummy Counselling	2.015 (0.735)*
Dummy Sanctions	1.956 (0.805)*
Dummy Placement subsidies	-0.551 (0.659)
Dummy Job creation	-0.086 (0.576)
Economic growth	-0.00069 (0.144)
Unemployment	0.132 (0.053)*
Threshold value between negative and not significant	0.436 (0.642)
Threshold values between not significant and positive	1.723 (0.659)*
Pseudo R-squared	0.12

The reference measure is “other measures”

Inclusion of economic growth and unemployment leads to the loss of two observations owing to missings.

* = significant at least at the 10 per cent level.

We now turn to the regressions with regard to the **size** of the net effect starting from equation (6). We have estimated both models with and without the TOE observations. In the model with the TOE observations only placement subsidies differ significantly from the other measures (table 4). This measure appears to be less effective than the other measures. Job creation is also strongly negative compared to the other measures, but it is not significant. In this model the unemployment rate has a strong positive (and significant) influence. If we omit the TOE observations the adjusted R-squared increases considerably, which illustrates the point made earlier that the results of the TOE cannot

easily be compared with the other studies. In this model training is significantly more effective than ‘other measures’, the reference category. Job creation is now the worst performing measure, although its effectiveness does not differ significantly from that of ‘other measures’. Both the unemployment rate and economic growth are significant in this equation. Both variables indicate that the net effects are lower the better the economic situation is. A possible explanation for the influence of economic growth is that this variable is related to the creation of new jobs and thus positively affects job the opportunities for the unemployed. In both regressions the relative performance of training, counselling, sanctions, placement subsidies and job creation is similar, but their relative performance compared to ‘other measures’ differs.

Table 4 *Regressions with regard to the size of the effect with en without TOE observations (the latter without correction)*

Dependent variable: size of the net effect		
TOE observations included?	Yes	No
Number of observations:	152	111
Explanatory variables	Coefficient (standard error)	Coefficient (standard error)
Dummy Training	-0.008 (0.110)	0.083 (0.028)*
Dummy Counselling	-0.016 (0.1445)	0.052 (0.039)
Dummy Sanctions	-0.108 (0.148)	0.033 (0.035)
Dummy Placement subsidies	-0.276 (0.145)*	0.004 (0.042)
Dummy Job creation	-0.203 (0.129)	-0.023 (0.035)
Economic growth	0.020 (0.029)	-0.028 (0.007)*
Unemployment rate	0.045 (0.011)*	0.008 (0.004)*
Constant	-0.190 (0.133)*	0.044 (0.033)
Adjusted R-squared	0.11	0.29

The reference measure is “other measures”

**significant at least at the 10 % level*

In the previous section we have seen that TOE observations can be made more comparable with the other observations by multiplying the latter with a constant factor. This factor has been determined by taking the coefficients of the regression without TOE observations as our starting point. We assume that these coefficients represent unbiased estimates. Then a regression is carried out with the TOE observations to estimate the correction factor given the coefficients from the regression without the TOE observations. We find a value of 1,77, which is quite comparable with the value found in the previous section (1,71) based on simulated data.

When we apply the factor 1,77 to the TOE observations and carry out a new regression including the adjusted TOE observations we get the results shown in table 5. The correction leads to a improvement of the adjusted R-squared. It is in between the values obtained with the two regressions shown in table 5 (0,11 and 0,29 respectively). In this

sense it helps to correct the TOE observations. The results are similar to the ones from the regression with uncorrected TOE observations. The only difference is that job creation is now also significant.

Table 5 Regressions with TOE observations and with TOE correction

Dependent variable: size of the net effect	
Number of observations:	152
Explanatory variables	Coefficient (standard error)
Dummy Training	0.006 (0.090)
Dummy Counselling	-0.0170 (0.110)
Dummy Sanctions	-0.077 (0.134)
Dummy Placement subsidies	-0.199 (0.105)*
Dummy Job creation	-0.190 (0.129)*
Economic growth	0.038 (0.025)
Unemployment rate	0.034 (0.008)*
Constant	-0.216 (0.111)*
Adjusted R-squared	0.19

The reference measure is "other measures"

** = significant at least at the 10 % level*

We also experimented with different ways to adjust the TOE observations. An obvious alternative is to take account of the possibility that there is not only a scale factor, but also a constant in play. However, the results hardly differ from the outcomes in table 6 when this idea is put to practice.

By using the sample means for economic growth and the unemployment rate we can compute average net effects for the various measures. Table 6 gives the results.

Table 6 Average net effects for each measure based on the sample means of the context variables

Measure	Model without TOE observations (see table 5)	Model with corrected TOE observations (see table 6)
Training	0.088	0.097
Counselling	0.057	0.088
Sanctions	0.038	0.013
Placement subsidies	0.009	-0.109
Job creation	-0.018	-0.100
Other measures	0.005	0,090
Average net effect	0,030	0,013

Using the model with TOE observations, the average net effect varies between $-0,018$ (for placement subsidies) and $0,088$ (for training). On average the net effect is fairly small (approximately $0,030$). The model without TOE observations leads to an even smaller net effect (on average approximately $0,013$). The main difference is that placement subsidies and job creation come out much more negative in the TOE studies. For training, counselling and sanctions the differences are much smaller. Training appears to have a net effect of $8,8$ to $9,7$ per cent points; for counselling this is $5,7$ to $8,8$ per cent points. For sanctions small positive effects are found in both cases. For ‘other measures’ inclusion of the TOE observations improves the outcomes leading to a similar net effect as is found for training and counselling.

The results can be interpreted as follows. A value of x for the net effect implies that a person increases his probability of finding a job by x per cent points by participating in an ALMP. Assuming that a person’s chance to find a job is 30 per cent without participation in a measure, this becomes $31,3$ to $33,0$ per cent in case of participation. For some measures (like training and counselling) it might be much higher and for others (like placement subsidies and job creation) much lower (even lower than the reference situation of 30 per cent, implying a negative net effect). It is important to repeat that most studies do not provide information about the level of job placement rates. For disadvantaged groups the chance of finding a job without participation in a measure can be quite low⁷, implying that an increase of $5,7$ to $9,7$ per cent points which was found for training and counselling may still be relatively high. But we do not know whether net impact for these groups is similar to the average level found in this paper⁸.

⁷ In the Netherlands the job placement rate for unemployed social assistance clients is approximately 20 per cent on an annual basis, for example. For those with an unemployment insurance benefit it is considerably higher (35 per cent). These are averages. For disadvantaged groups among the unemployed these figures are much lower.

⁸ De Koning et al (2005) conclude on the basis of a limited number of studies that net impact is probably higher for disadvantaged groups than for other groups, but this conclusion is based on the sign rather than the size of the effect.

We also carried out regressions with the number of observations added as an explanatory variable. The problem is that information on this variable is available for a limited number of studies only. If we take the sample mean for the missing values and include this variable in the regression it appears to be significant with a negative sign. So, the net impact tends to be lower the larger the number of observations.

Furthermore, we have run regressions that included a dummy indicating whether the study made use of an experimental evaluation or not. In regressions without the TOE observations the sign of this dummy is negative. So, evaluations with a better methodology seem to lead to smaller effects. The coefficient involved is, however, not significant. An interesting feature of the results is that in these regressions the coefficient for training, although still significant and positive, becomes smaller in relation to the coefficient for counselling.

The latter also happens in an extended model with TOE observations and dummies in relation to cases where the TOE method has been applied. In these regressions the dummy for experiments is negative **and** significant. Furthermore, in this model studies that use an experimental evaluation only produce positive effects in case of counselling or sanctions, not anymore for training. Hence the net effect of training might be smaller than table 7 suggest, while the case for sanctions is probably stronger than the table suggests. However, the sample does hardly contain experimental evaluations of training programs. So, final conclusions cannot be drawn.

4 CONCLUSIONS AND FINAL REMARKS

In this paper we have used the outcomes of evaluation studies from various countries to make inferences about the average net impact of active labour market policies on job entry chances. Our conclusion is that this average net impact is fairly small. It is not higher than 3 percentage points if all ALMPs are taken together. This average effect is based on regression analyses in which the results of the studies are used as observations. The models allow for variation of the net impact between the various types of measures. Furthermore, the models control for the influence of the economic situation. For training and counselling the net impact may be around 7 per cent points. This may still be relatively high compared to the job entry chances that unemployed people have when they do not participate in such a program, which are probably not higher than 35 per cent and often much lower.

Our initial analyses suggest that training, counselling and to a lesser degree sanctions and 'other measures' have positive net effects. Placement subsidies and job creation measures have a negative effect. The effect of training may have been somewhat over-estimated and that of sanctions somewhat under-estimated. The reason is that there are hardly any experimental evaluations in our sample with regard to training. Although the available data is not sufficient to obtain a reliable estimate for the influence of the evaluation method (experimental evaluation or not), some of the regressions suggest that studies that do not use an experiment, or non-experimental evaluations that do not deal with the selection bias problem in a satisfactory way, may produce overly optimistic results.

According to our results the net impact of active measures is higher in situations of low GDP growth and high unemployment than under more favourable economic and labour

market conditions. In the latter situation unemployment jobseekers are more likely to find a job anyhow.

Clearly, our results should be treated with caution. First, the number of cases in our study is relatively small given the fact that we want to make estimates for the various ALMPs individually. Second, the number of studies on which the cases are based is even more limited. Third, it is not easy to relate the results of studies using the timing of events method to the results of the other studies. But if we exclude the TOE studies we are left with only 111 observations or so. Furthermore, the TOE studies are important as these are the ones among the non-experimental studies that deal in the most satisfactory way with the problem of selection bias. In principle, from the estimation results of the TOE method the net impact of a measure on the job entry probability and on expected unemployment duration can be computed. Unfortunately, many papers do not contain this type of information. It would be recommendable for authors to present the results in such a way that a comparison with other studies is easier. In the present situation the best way to proceed is to collect more cases, both studies using the TOE method and studies using other methods, and apply separate analyses to both types of studies. Then in a second step we may investigate to what degree the results point to the same direction. When we have more cases it will also become possible to include more factors in the regression like target group characteristics. This would enable us to test whether measures work better for some groups than for others.

From this exercise we conclude that the existing research into the impact of active labour market policies is to a large extent aimed at improving evaluation methodology. Researchers are less interested in presenting their results in such a way that these can easily be compared with the results of other studies. This leads to a situation in which on the one hand many studies are carried out, but on the other hand not much can be learned from the outcomes for policy purposes. We can observe a huge variation in the outcomes, but we can explain only a small part of it.

REFERENCES

Abbring, J.H. en G. van den Berg (2003), The non-parametric identification of treatment effects in duration models, *Econometrica*, vol. 71.

Bijwaard, G.E.(2002), “Instrumental Variable Estimation for Duration Data: A reappraisal of the Illinois Reemployment Bonus Experiment”, *Econometric Institute Report EI 2002-39*

Blundell, R., Meghir, C., Dias, M.C. en Van Reenen, J. (2004), “Evaluating The Employment Impact of a Mandatory Job Search Program”, *Journal of the European Economic Association*, 2, 569-606

Gelderblom, A. and J. de Koning with support by K. Lachhab (2007), Effecten van “zachte” kenmerken op de reïntegratie van de WWB, WW en AO populatie. Een literatuurstudie (‘Effects of “soft” factors on the reintegration of social, unemployment and disability benefit claimants’), SEOR, Rotterdam

Heckman, J., R. Lalonde and J. Smith (1999), The economics and econometrics of active labor market programs, in: *Handbook of Labor Economics, Volume 3*, Ashenfelter A. en D. Card (eds.), Elsevier Science, Amsterdam.

Heckman, J., J. Smith en N. Clements(1997), Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts, *Review of Economic Studies*, vol. 64.

de Koning, J. (2007), Is the changing pattern in the use of almps consistent with what evaluations tell us about their relative performance?, in: J. de Koning (ed.), *Evaluating active labour market policies: measures, public-private partnerships and benchmarking*, Edward Elgar, Cheltenham.

de Koning, J., Gelderblom, A., Zandvliet, K. and Van den Boom, L.(2005), Effectiviteit van Reïntegratie. De stand van zaken – Literatuuronderzoek (“Effectiveness of reintegration. The state of the art – literature review), Ministry of Social Affairs and Employment, Den Haag.

de Koning, J. (2001), Aggregate impact analysis of active labour market policy: a literature review, *International Journal of Manpower*, Vol. 22, No. 8.

Kluve J. (2006), “ The Effectiveness of European Active Labor Market Policy”, *IZA Discussion Paper No. 2018* , Bonn

Kluve, J. and C.M. Schmidt (2002), Can training and employment subsidies combat European unemployment, *Economic Journal*, Vol. 17, Issue 35.

Van Rheeën, J. (2003), “Active Labour Market Policies and the British New Deal for the Young Unemployed in Context”, *NBER Working Papers 9576*

ANNEX: A MODEL FOR THE CASE WHERE THE LABOUR MARKET STATUS OF PARTICIPANTS AND CONTROLS IS REVIEWED ON A CERTAIN DATE

In section 2 we stated that studies using experiments or matching methods use two ways of representing the effect on the job entry rate. The first one was treated in this section and refers to the chance that a person has found a job during some time interval after completing the program. The second one refers to the effect on the chance that someone has a job on a specific point in time (after completing the program). We now turn to the latter case.

If we look at the chance of having a job on a specific point in time than we have to account for the fact that people who find a job after completing the program may lose that job and become unemployed again (and then find another job, etc.). In the most basic case where the spells of unemployment and employment are each distributed according to an exponential distribution this process can be described by a continuous Markov process, more precisely as a birth-mortality model with two statuses. The statuses are unemployed (status 0) and employed (status 1). Unemployment has a exponential distribution with parameter β and job duration with parameter μ . Using Kolmogorov's Backward equation the probability that a person who was unemployed at time 0 (status 0) is employed at time t (status 1) is equal to:

$$P_{01}(t) = \frac{\beta}{\beta + \mu} - \frac{\beta}{\beta + \mu} \exp(-(\beta + \mu)t) \quad (\text{A1})$$

The parameters β en μ can both be interpreted as hazard rates. From equation (A1) we can conclude that if the probability to escape from unemployment becomes big then the probability of being in a job tends to 1 with the elapsing of time. If the chance that a person with a job becomes unemployed (μ) is very small the equation (A1) tends to the result obtained in section 2. So, only in the latter case the two ways of measuring job placement probabilities lead to the same result.

Now suppose that at time 0 an intervention takes place which changes the hazard rate from unemployment to employment according to the following equation:

$$\beta^* = \beta \exp(\gamma) \quad (\text{A2})$$

Then the chances of having a job at time d for those treated (P^d) and for the control group (P^c) are equal to:⁹

$$P_{01}^d(d) = \frac{\beta^*}{\beta^* + \mu} - \frac{\beta^*}{\beta^* + \mu} \exp(-(\beta^* + \mu)d) \quad (\text{A3})$$

and

$$P_{01}^c(d) = \frac{\beta}{\beta + \mu} - \frac{\beta}{\beta + \mu} \exp(-(\beta + \mu)d) \quad (\text{A4})$$

It is not possible to derive an explicit relationship between the job chances in the two cases and the impact of the intervention. Only asymptotically, for high values of d , this is possible. Then we have:

$$P_{01}^d(d) = \frac{\beta^*}{\beta^* + \mu} = \frac{1}{1 + (\beta^*/\mu)} = \frac{1}{1 + \{\exp(\gamma)\}(\beta/\mu)} \quad (\text{A5})$$

and

$$P_{01}^c(d) = \frac{\beta}{\beta + \mu} = \frac{1}{1 + (\beta/\mu)} \quad (\text{A6})$$

implying:

$$\gamma = -\log\left\{(1/P^d - 1)/(1/P^c - 1)\right\} \quad (\text{A7})$$

⁹ Notice that we assume in equation (10) that in case of repeated spells of unemployment the effect of the intervention stays intact.

This is approximately the same as equation (5) as long as the difference between P^e and P^d is small.

So far, we have assumed that the hazard rate from employment to unemployment is not influenced by the intervention. If one takes a measure like training into consideration this assumption may not be realistic. Training often takes so much time that the effect on the length of the unemployment period during which the training takes place may be small. However, training may considerably reduce the chance of future unemployment, which would be reflected in a lower value of μ . Assuming that the effects of the intervention on both hazard rates is constant over time, we could compute the effects if we would have information on the two types of chances (the chance of ever finding a job after the intervention and the job chance on a specific time point after the intervention) for the participants and the control group.