Benchmarking carrots and sticks: developing a model for the evaluation of work-based employment programs

Castonguay, J.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
2. Benchmarking the policy-chain

This chapter will introduce benchmarking as an instrument for evaluating labour market policy. The question to be answered in this chapter is thus: what is benchmarking and how can it be used to evaluate active labour market policies? The first section will provide a definition of benchmarking that is applicable to private and public benchmarks. Having defined what is social benchmarking, the section will go on with discussing the policy-chain and its five elements, since this will be the basis of the benchmarking model and will be used throughout the rest of the research. A typology of benchmarks based on this policy-chain will also be defined. The second section of this chapter will discuss the methodology of benchmarking, which consists of four steps. In this section, each of these steps will be discussed in detail. Also, this section will present the various choices which must be made when benchmarking. Being aware of the impact these different choices can have will allow proceeding with the building of the benchmark model in chapter 3 and 4, so that an actual benchmark can be performed in part 2 of the research.

2.1. The policy-chain and a typology for benchmark models

Benchmarking evolved as a powerful international trend in the private sector in the 1980s. Xerox Corporation was the initiator of this new trend. By means of comparing itself to other companies, it reversed a downward profitability curve to become one of the most efficient corporations in the US (Schütz, Speckesser and Schmid, 1998). This surely incited other companies to do the same, which eventually also became a common exercise in the public sector with the assent of New Public Management. Benchmarks in the public sectors have been used in a wide variety of domains, ranging from garbage collection to hospital care. Several definitions of benchmarking can be put forward, but one definition is most suitable for both the public and private sectors: “Benchmarking as an efficiency tool is based on the principle of measuring the performance of one organisation against a standard, whether absolute or relative to other organisations” (Cowper and Samuels, 1997, in Schütz, Speckesser and Schmid, 1998). Benchmarking is an efficiency tool since it is able to point out performance gaps in the production process of an organisation. Organisational learning can thus be achieved by implementing positive changes based on the identified roots of these performance gaps (Tronti, 1998).
Benchmarking in the public sector is only different from the private sector in the fact that public services are the object of the measurements, and not privately produced consumer goods or services. Social benchmarking is therefore defined here as a benchmarking exercise applied on social policy with the intent to improve social programs by learning from best-practices and to foster efficiency and effectiveness in the delivery of social provisions.

The implications of these differences are that while the private sector will base its benchmark on the production chain, the public sector will use its policy-chain as the basis for its benchmark. The policy-chain comprises of several steps in which policies are thought up and concretized in different measures. The policy chain consists of four principal elements: the input, the process, the output and the impact (Spicker, 2006). Because some factors that are external to the policy-making process can have an important influence on the policy chain, external factors are added to the model. These five steps of the policy chain will make-up the five categories of indicators that can be measured by a benchmark of public policy. While this research will concentrate itself on active labour market policies, this section will discuss the policy-chain in its most general form, and is thus applicable to all types of public policy. Chapter 3 will take this general model and adapt it specifically for the type of active labour market policy to be benchmarked, that is to say, mandatory work-based employment programs.

![Figure 1: The policy-chain](image)

The first step of the policy chain is determining the inputs. The level of inputs answers the question of what is put in place for the policy to perform its task (Spicker, 2006). This initial component of the policy-chain first includes the general objectives of the policy. Naturally, policies are created in order to change a certain situation and have a particular impact on persons and/or environments, which will be formulated in some general objectives for the policy to achieve. These objectives will be
formulated in terms of output to be achieved (for example, serving 200 clients per day) or in terms of impact (for example, increasing labour participation rates) (more explanation on output and impact below). Also, the major target groups to be served by the policies will be determined. These target groups can be defined by many different characteristics, depending on what is relevant to the policy. For example, target groups can be defined by characteristics such as area of residency, level of income, age and gender, ethnic group, education level, and many more. Next, the objectives of the program are translated in the social security laws, where the general rules of the system are defined. These will form the outer boundary in which the program will form itself, and within which the process can be defined. This legislative framework can thus be seen as the foundation upon which a specific program will be build. Also, the program will insert itself in an already existing social security system, which can vary greatly in terms of decentralisation and concentration of tasks within one or many agencies. At last, the amount of resources available for the programs will be fixed. These resources comprise the financial resources, the human resources and the infrastructure (physical resources). These input-elements thus measure the general blue print in which the program will take place.

Input indicators (what is put in place for the policy to perform its task), for example:
- objectives
- target groups
- legislative framework
- social security system: division of responsibilities
- resources: budgets, human resources and infrastructure

This set of objectives, laws and resources will then determine the second step in the policy chain: the process of the policy. It answers the question of how the program is delivering its public services. This process can be said to have two levels. The first level is the organisational-level process. Within this level, a mode of governance will be set by the top-level decision makers. This will determine whether a traditional bureaucratic model will be applied, or a different form such as New Public Management or other more hybrid type of modes of governance. Amongst others, the organisation responsible for the delivery of the services will be chosen. For example, a government can choose to use private contractors for delivering the services to the population, while an independent governmental agency will be responsible for monitoring those contractors. Furthermore, the financial construct of the program will be designed and geared towards certain incentives for the different actors involved in the program. The elements that form the
organisational-level process need to be complemented by the program-level processes to complete this category of indicators. These program-level processes refer to more micro-level processes, that is to say, decisions which are directly related to the experience of the claimant in the program. These processes will be chosen in relation to different theories and ideas on what is needed for the individual in the program for him/her to reach the objective of the program. For example, these indicators will measure the type of services delivered, for how long, in which manner, at what cost for the customer, etcetera.

More specifically in terms of benchmarking active labour market policies, this means looking at the different theories on labour market intervention and looking at what type of program intervention will result in the greatest number of participants finding a job upon completion. For example, this means measuring the type of job search assistance made available, for how long, and at which frequency. As can be seen, the specific indicators for this program-level process will thus vary greatly from program to program, depending on the influence of various theories on the specific elements of those programs. While for some programs the type of reward given to participant will be relevant, other programs will require that the type of qualification attained will be more important. As will be shown in chapter 3, the choice of these program-level process indicators will thus require a deeper investigation than with the other categories of indicators. This is due to the fact that to be able to choose program-level process indicators, it is necessary to know the nature of the determinants of success and failure of work-based employment programs.

Process indicators (how are the policies delivering their programs)
   - Organisational-level processes:
     - mode of governance: privatisation and performance pay
   - Program-level processes:
     - Determinants of success and failure of micro-level interventions

Using the input, the process “makes” the output, the “product” of the social program. The output answers the question of what is being produced by the policy. This thus means looking at the quantity and quality of the different services delivered by the social program. Analysing the level of output is important because the same amount of financial and human resources does not imply the same quantity and quality of the services. By comparing the input to the output, the influence of the process on the quality and quantity of services delivered can be measured. This opens the black box of policy implementation that was long left closed by evaluations which looked only at output or impact. Furthermore, some objectives can also be formulated in terms of
output, such as delivering a public service to a target number of citizens per month or year. Besides looking at quantity and quality of services, elements of the legislative framework can also be measured with regards to their actual use. This will thus show how policy and practice can differ from one another. For example, the number of sanctions actually being used will be measured as part of the output of the program. Third, the coverage rate of the program, in terms of the percentage of unemployed which took part in the program, is also important to measure in order to show the relative size of the program within the unemployment benefit scheme.

Outputs (what are the “products” of the policy):
- type of services delivered and number of participants for each
- quality of the provision
- sanctions, type and quantities used
- coverage rate

The last step in the policy chain will then be the impact the policy has on the participants of the employment program. This impact is the ultimate result of the policy. This impact is very relevant since for many social policies the objective will not be defined in terms of their amount of output, but rather in terms of their impact. Indeed, the difference between the output and the impact is that the impact is not an actual product of the public program, but rather its effect on the participants. For example, in the case of employment policy, the impact refers to participants finding a job on the regular labour market, something that is not directly "given" to the participant, but rather the impact of several interventions such as training and job search assistance. This distinction also implies that the impact is not directly under the control of those delivering the services, but that several external factors can come between the direct actions of the program and the impact it has on the individuals. As already mentioned, the objectives of a public program are often formulated in terms of not only output, but rather on their intended impact, making this indicator crucial to include in the performance measurement. Furthermore, the sustainability of this impact is also important to measure, since it can vary from one type of policy to another. The impact can also relate to other aspects not directly mentioned in the objective of a policy, such as preventing claimants to enter the benefit, or to stay too long on the benefit. In addition, some impacts are indirect or unwanted, and should also be taken into account in choosing the indicators.

Impact (results of the policies):
- rates of outflow to work, caseload reductions
- rates of inflow prevention
- sustainability of influence on individual (for example: status after one year)
- increases in revenue / salary
- other types of impact relevant to the program (fraud reduction, reduction in child poverty, reduction in crime rate, etcetera)

Besides these four steps of the policy chain, there is an important last factor that needs to be added to the evaluation framework. The whole policy chain is very much influenced by various external factors, which cannot directly be included in the policy chain but have some direct influence on all of its steps. The reason they cannot be included in the policy chain is that they are outside the control of the policy maker. Such elements are, for example, the unemployment rate in a country or region, the institutional environment of the employment program, different labour and social security laws acting at the periphery of the employment program, and more. These external factors are thus directly linked to the policy chain and have such a great impact on it that it is essential to include them in the evaluation framework. They can be regrouped in different categories, such as macro-economic factors, political context, and the juridical context.

External factors (influence the input, process, output, and impact, but outside the policy-chain, thus outside the reach of the policy-maker):
- macro-economic factors: unemployment rates, job vacancies, poverty and inequality rates, total welfare caseload, government debt and budget balances
- political context: parties in places and political scope for reform
- juridical context external to social security laws (for example, employment protection laws, human rights legislation, tender laws)

Obviously, depending on the specific type of active labour market program, the precise indicators which define best each section of this policy-chain will vary. The different indicators which were given here are not meant to be an exhaustive list of all components of the policy-chain. They are rather meant to provide some examples of the types of elements that are making up the links of the chain. In fact, they are kept purposefully very broad in order to be applicable to most activation programs found around the world, disregarding their specific content. However, it is clear that when attempting to benchmark a specific active labour market program, these indicators will require more specificity and depth in order to be useful as a benchmark model.
This section will now go on with a definition of different types of benchmarking, keeping these different parts of the policy-chain as the basis of the analysis. However, in order to keep the discussion on a general level that is applicable to most active labour market policies, the specific indicators found within each of these parts of the policy-chain will not yet be given. As can be imagined, the choice of these indicators will require a more extensive analysis, and will thus be tackled in the next two chapters.

**Benchmark Typology**

Measuring different parts of the policy-chain will create different types of benchmarks. Throughout the literature on benchmarking there are accounts of many different typologies, each with their own types and definitions. The great variety of typologies found in the literature shows that not one model is being seen as the standard way to benchmark (see amongst others, Schütz et al. (1998) for a review of different benchmark typologies). The typology developed here will draw on the policy-chain and show the differences between four types of benchmark practices.

- **Input Benchmark**
  This type of benchmarks is mostly focused on the finances of employment programs, looking at how much budget there is available in each of the benchmarked programs. For example, this means comparing different countries on the budgets for ALMP as a percentage of GDP, or as a ratio to the number of unemployed in the country. Often, it is assumed that the higher the financial input for a program, the more effective it will be at reaching its aim. It is thus common, for example in OECD publications (see Employment Outlook, yearly), to see public spending on ALMP being compared for different countries.

- **Impact Benchmark**
  These types of benchmarks are focused on the other end of the policy-chain, and compare the impact of each program in the benchmark. By measuring the impact with regards to the objective of a program, the effectiveness is measured. These benchmarks can thus be used to compare the effectiveness of programs, but it should be noted that this requires that all programs have the same objectives. Depending on the question this benchmark attempts to answer, either the gross impact or the net impact will be calculated. Net impacts are often used since they only show the impact of the program itself on its participants, and thus do not take into account effects that would have been present even without participation in the program. More specifically for labour market policies, the net impact will subtract from its calculation those participants who
would have returned to the labour market without the help of the program. Although most impact-benchmarks will only calculate the impact of the program with regards to the main objective of the program (for example, calculating exit-to-work ratios if the aim is the return to the labour market of participants), it can also be that secondary impacts, such as an increase in earnings for participants, or also prevention-of-entry ratios are calculated.

- **Output Benchmark**
  First, output-benchmarks can be centred on comparing the number of provisions which are offered to benefit claimants. For example, comparing the number of completions of a training program or comparing the number of participants in a Work First program would be considered to be an output benchmark. Comparing client satisfaction would also be considered to be an output benchmark. These benchmarks are relevant when the objective of a program is not per se related to the impact of the program on its participants (for example, the return to the labour market). Indeed, output benchmarks can be used to measure effectiveness when programs have the general objective to provide skills or assistance to a participant, for his/her own sake. In these cases, it would be relevant to assume that participating in a program is in itself a result that is achieved. The effectiveness of such programs can thus be compared on their output, assuming they have the same output-related objective.

- **Process Benchmark**
  This is a very traditional way of benchmarking, associated with the “re-engineering production strategy” common in many companies in more technical branches. In these types of benchmarks the production process of different companies is compared in order to make improvements in the way the products (or services) are made. It is less likely to be used on its own for the comparison of labour market programs. Process Benchmarks would compare the design of programs, comparing their length, the type of activities they offer to their participants, the number of hours needed to complete the program, etcetera. Indeed, this information might be interesting to compare on its own, but will be much more informative when combined with the other parts of the policy-chain, as is done in the last benchmark type: the best-practice benchmark.

- **Best-practice Benchmark**
  This last type of benchmarking looks at all the components of the policy-chain at the same time, and thus includes all previous four benchmark into one single benchmark model. As already discussed, the impact-benchmark model was useful for comparing the absolute effectiveness of
different programs, that is to say, the extent to which programs reach their goals, assuming that all programs have exactly the same goals. If one wants to take into account variation in objectives and thus measure relative effectiveness, an input benchmark must be combined with the impact benchmark. Moreover, comparing the financial inputs to the output and the impact of the policy will allow for measurement of the efficiency of the policy. In this way, efficiency can show that the performance of one policy is higher than that of another policy because it achieved the same results with fewer resources, or more results with the same resources.

Furthermore, a measure of the process is necessary in order to explain why a certain level of efficiency and effectiveness is being achieved. By comparing the process of each country with input, the output and the impact, one can draw conclusions on the efficiency of the implementation of a policy. This means that by including elements of the process into the benchmark, the black-box of how the input becomes the output can be opened. Furthermore, measuring the importance of external factors will also help explain why certain levels of efficiency or effectiveness are being achieved. For example, they will help explaining why a certain level of impact is noticed while it does not seem to be coherent with the level of input and the process being used. As an example, a high level of unemployment in a region may explain why a low rate of return to the labour market is noticed in a region with high budgets and an efficient process. Therefore, in order to be able to explain the reasons behind the differences in efficiency and effectiveness between the subjects of the benchmark, process-elements and external factors are included in this benchmark. Their purpose is focused on explaining the background of the policy, in order to make a realistic evaluation. Clearly, adding the process elements and the external factors to the benchmark corrects for an important flaw that many benchmarks models have, namely that the subjects to be compared are so different that comparison becomes unsound.

2.2. Social benchmarking for evaluation and public management

Now that social benchmarking has been defined and that the various types of benchmarks have been presented, the usefulness of benchmarking active labour market policies will be discussed. As already mentioned in the introduction chapter, current methodologies for the evaluation of active labour market policies do not provide a clear blue-
print of how to best evaluate those policies. This next section will thus present the shortcomings of the available methods and discuss how benchmarking could intervene in solving some of these.

It is clear that all social benchmarks will have the first intention to evaluate the performance of a public organisation or program. However, what policy-makers will actually do with the results which come out of a benchmark can have various effects on the way in which these policies are being delivered. In fact, besides providing evidence of what works and what doesn’t, benchmarking results can also be used within the organisation in order to improve efficiency of program delivery. Although the benchmark in this research is not meant to be used as a management tool, section 2.2.2 will discuss how benchmarking can in fact also be used for this purpose. It will be shown that through the dissemination of benchmarking results, first, a quasi-market can be created which will increase competition and therefore efficiency, second, objectives will be more easily communicated and monitored, and third, decisions and actions can be legitimised based on performance measures.

2.2.1 Benchmarking for evaluation: opening the black box of policy implementation

In general, evaluation can be defined as the determination of the merit or worth of something. From the previous section it should be clear that benchmarking is in fact a specific type of evaluation methodology, where merit or worth is based upon the comparison of the performance of different organisations. In order to answer the question of why to benchmark labour market policies, arguments for benchmarking as an alternative to the current evaluation methods for labour market policies should be given. As will be shown, benchmarking does solve some of the criticisms of the current methodologies and should thus be considered as part of the evaluation toolbox for ALMP.

The evaluations of active labour market policies are mostly done from either of two perspectives. The first one is the macro perspective, which attempts to find a causal relation between macro-level variables, such as spending on ALMP and the effect on the unemployment rate. The other one is the micro perspective, which uses econometric models at a micro-level in order to estimate the influence of ALMP programs on the chances to find work for the participants.

The macro perspective evaluates active labour market policies by comparing the scope of these policies with the effect these should have on
the labour market. For example, a macro-evaluation might estimate how much unemployment has decreased with the increase in use of vocational training. An econometric model would be used to analyse this relationship between these two variables. However, accounts of evaluations of the macro-perspective are seldom, which Grubb (2004) explains by the fact that these evaluations measure effectiveness at a level that does not coincide with concerns of the public and the political arena. In addition, Martin (2000) mentions how there is growing interest in this type of studies, but that these are faced with important methodological issues. One of these is the fact that the number of statistical units – countries – remains often too small to make significant conclusions about the effect of a policy. One solution to this has been to also use variation in time for each country as a different unit of analysis. Also, most macro-level studies are faced with the problem that spending on unemployment rises as the unemployment level increases, also called the simultaneity bias (Martin, 2000). All of these drawbacks have meant that no clear link can be said to exist between the scope of ALMP (either in terms of spending, or number of participants, or other types of macro-level indicator of "size"), and the extent to which unemployment is reduced and economic growth fostered. In Martin’s review of the existing evidence across all OECD countries, his conclusions was that “the jury is still out on the matter” (p.89). This coincides with the findings of De Koning (2001), who reviewed 26 international macro-level studies of the impact of ALMP on the labour market, to find that 6 studies presented positive effects, 8 studies showed no significant effects and the rest showed negative effects.

On the other side, the micro perspective looks at how specific instruments and programs of ALMP (for example, schooling or subsidised labour) increase the chance of finding a job for different individuals. Micro-level evaluations are by far the largest type of evaluations of labour market policies nowadays. These evaluations could solely report the number of participants that did find a job after participating in the program. For example, a micro-perspective evaluation could conclude that through a specific program, half of the participants found a job on the labour market. However, many academics advocate that such figures are misleading. In reality, the true effect of participating in the program is not clearly illustrated by only looking at rates of exit-to-work, because other factors may have come to influence this return to work. These other factors can be grouped into three categories: dead-weight effects, substitution effects and displacement effects. The dead-weight effect refers to participants who would have found a job even though they would have not participated in the program. Their return to the labour market can therefore not be attributed to the effectiveness of the program,
but to a combination of favourable personal and labour market conditions. The substitution effect refers to participants who do find a job on the labour market, but this job would have nevertheless been filled by another unemployed. This also means that the participation in the program is not the cause of this return to the labour market, since someone who did not participate could have just as well filled the vacancy. The last effect is the displacement effect, where the participants in the labour market programs actually take-up jobs for which someone else was already employed. In other words, one job is gained but a new unemployed person is created. This situation is most realistic when considering programs for which a wage-subsidy is available for the employer who hires a benefit claimant. Hence, in order to calculate the true effect of the program on the unemployment rate, these three effects must be removed from the gross rate of return to work (De Koning, 2005).

In fact, most micro-level evaluations also concentrate on developing a proper way to calculate the net-impact of active labour market programs. The difficulty lies in the fact that the benefit claimants can only be observed in one stage: either as a participant in the program, or as a non-participant. Sophisticated econometrical models are therefore necessary in order to estimate what would have happened if the participants would have not taken part in the program. This requires the use of quasi-experiments, since direct experiments in both stages with one person are impossible. Quasi-experiments are set up such that one group of individuals is taken in charge by the employment service provider while a control-group is statistically designed (Smith, 2000). This control group can be formed through a matching-function, in which non-participants with similar characteristics as the participants are compared to the participants. These non-participants can be found in a group of claimants which was in the benefit scheme at the same time but was not targeted by the program, or found by looking back in time to a period in which participation in the program was not applicable to all claimants (Smith, 2000). Nevertheless, it has recently been shown by Lalove, Van Ours and Zweimüller (2008) that these differences in methodology can have important consequences for the conclusions being made. In their study, they compared two different methodologies for measuring net-impacts using the same program. They found that each method resulted in different net-impact calculations, even though they were measuring the same program.

Furthermore, as discussed by De Koning (2005) these methods imply in fact that the models are corrected for individual characteristics, such as level of education and family situation, or even the level of motivation to find a job. Other academics contest this method, arguing that these
individual characteristics are essential to the success of employment services and not taking these into account leads to unrealistic evaluation (Glebbeek, 2005). These differences in methodology have consequences for the policy recommendations that will be made. While those advocating for the calculation of net-effectiveness suggest concentrating efforts to help those with the smallest chance to re-enter the labour market (De Koning, 2005 and Koning, 2005), those rejecting the correction for individual characteristics actually propose helping those who have the highest chance to returning to the labour market (Glebbeek, 2005). Accordingly, cream skimming will then be seen as a failure of the system, or as a necessary policy for the efficiency of the employment policies (Glebbeek, 2005). These two different methodologies are therefore clearly influenced not only by different views of how efficiency is to be measured, but also by the interpretation of efficiency itself. Efficiency being a measure of how much resources are used to reach a certain goal, this debate then comes down to fundamental differences in the goals to be attained by the labour market policy. At the end, this political choice is a matter of whether one seeks to reduce the amount of benefits paid out by either preventing entry into the system or by facilitating exit out of the system.

The social benchmark constructed in this research is a compromise between these two approaches. The first reason why this is a compromise is that different success criteria can be measured side-by-side when measuring the impact of the program. This means that performance can be concluded to be high according to gross-effectiveness but low according to net-effectiveness. Second, by measuring the whole policy chain, this social benchmark model will allow to also understand how such divergence in gross and net effectiveness is created. The drawback of net-efficiency is that instead of shedding light on the reason why one program has a higher gross effectiveness than another program, these differences are eliminated and camouflaged. In terms of learning best practices, it is more interesting to know which choices lead to lower gross-effectiveness (such as choices in the target group), then to know that “if all things where equal” a certain program would turn out to be more effective than another. This benchmark has as its objective to assist policy-makers in designing highly effective work-based employment program, so demonstrating how gross-impacts are created will be central to its analysis.

The need for linking impacts to the process of policy-making was pointed out by many researchers such as Schmid, O’Reilly and Schömann (1997), Meager and Evans (1997) and Mosley and Sol (2000). These researchers made a significant contribution to evaluation theory from the perspective
of labour market policies by showing how the shortcomings of micro-level evaluations could be alleviated by “opening the black box of policy implementation”. These authors contested the tradition of simple program evaluation of labour market policies. As already mentioned, most evaluations focus on a micro-perspective and look at the impact different reintegration instruments have on single individuals. These types of studies simply evaluated separate instruments of a policy programme, by comparing its measurable outcomes to the policy goals. The major critique was that these studies treated policy formation and implementation as a black box, neglecting to take into account the interaction of different instruments and neglecting to consider the cumulative impacts of all instruments of a program (Meager and Evans, 1997). The proposed “Target-oriented evaluation approach” therefore differs from program-oriented approaches by analysing policies in their whole context, looking at which broadly defined targets can be achieved by which policy combination (Meager and Evans, 1997). In order to open the black box of policy design, the evaluation of the processes that can create the most optimal impact is crucial. A different term for such type of evaluations is “Process-evaluation”, which stresses the need to focus on why the targets are or are not reached, which can only be achieved by looking at the policy formation and implementation stages (Pierre, 1999).

This vision was also shared by Campbell (2000), who stated that programs were treated by micro and macro evaluations as black-boxes, where content, nature resourcing, staffing and quality of provisions are not part of the performance indicators. Macro-evaluations were also pointed at for not looking at the core of the programs being evaluated. Esping-Andersen (1990) criticised only looking at the amount being spent on social policy as a measure of performance, arguing that welfare states of similar size and scope might have entirely different structures which have entirely different effects on how social risk is being alleviated for different groups. This general comment which applies to all social policy is surely applicable to active labour market policies.

The importance of including the evaluation of the nature of the design or the implementation of a program was said to be even more apparent when the design or implementation varied widely at the local level (Campbell, 2000). Indeed, it is often the case that social policies are being implemented by local governments within a broad national regulatory framework. Ignoring the variation in implementation of those policies would then have an important impact on the findings of such an evaluation. Additionally, Campbell discussed how the evaluations that do take into account the varying implementation of policies at the local level are often of qualitative nature. Only a few local evaluations were
found by Campbell to use large scale quantitative evaluation methods. Nevertheless, the combination of a national and a local level of evaluation with a quantitative and a qualitative focus would maximize the chance to find out what really works in social programs (Campbell, 2000). As pointed out by Finn (2000), there is thus a great need for the development of new evaluation methodologies that are able to adequately test social policy arrangements in which the design and the implementation varies from one unit of delivery to the next.

Clearly, a meso-perspective is very much appropriate for the evaluation of active labour market policies. The meso perspective can be defined as one looking at performance at the level of programs or organisations. The focus is on measuring the effect of the design and the implementation of a program on the level of performance, as measured by the impacts and the outputs. Meso-level evaluations therefore use a wide range of variables. These comprise macro-level variables, variables on the level of the organisations and programs implementation, and micro-level variables. Benchmarking thus clearly belongs to this level of evaluation. In addition, cost-benefit analyses and more qualitative analyses are also common to the meso-level (Pierre, 1999). Unfortunately, meso-level benchmark models of public employment services are rare. Even rarer is meso-level evaluation research that reached publication in international, or even national, scientific media. Most of these evaluations are in fact very needs-specific and do not reach further than policy administrators who directly influence the policy implementation. Indeed, as Helgason (1997) mentioned, these types of benchmarks often have a bottom-up approach and are mostly meant for internal use.

Nevertheless, the potential for benchmarking as an evaluation tool is promising. Indeed, benchmarking is on one side more general than micro-level evaluations, since these evaluations focus on single provisions and their net-impact on the individual’s chances on the labour market. On the other side, benchmarking is more specific than macro-level evaluations, as macro-evaluations focus on the economy or labour market as a whole. As a complement to macro and micro level evaluations, benchmarking should therefore be regarded as a welcomed addition to the evaluation tool-kit.
2.2.2 Benchmarking as a public management tool: facilitating competition, monitoring and legitimisation

With the emergence of New Public Management, evaluations have taken the centre-stage in public policy-making. These days, not only academics are interested in researching the effect of different government policies, but the government itself is taking an active role in finding out whether its policies have the intended effect. In times where public spending is under increasing pressure, policy-makers are in need of knowing whether social investments are having the desired returns. From various policy domains such as health care, education, childcare, police and fire departments, evidenced-based policy making has become a must. Governments ought to implement policies that have been proven to work.

Besides improving public policy through mutual learning, social benchmarking can actually assist in improving the efficiency and effectiveness of the delivery of social programs by acting as a management tool. In fact, the role that benchmarking can play in management functions is often what first comes to mind when looking at benchmarking. Since the benchmark to be undertaken in this research is not directly linked to any government body, these objectives are not explicitly part of this research. Nevertheless, it is worth mentioning shortly here how benchmarking can be used in order to improve the efficiency and effectiveness of program delivery besides doing so through mutual learning. Three functions of benchmarking for management purposes can be distinguished: the creation of a quasi-market through competition, the communication of objectives and monitoring of those objectives, and the legitimisation of decisions and actions based on evidence of good or poor performance. Each of these will briefly be discussed in the following pages, before moving on to the second step in the construction of the social benchmark to be used in this research.

Two types of inefficiencies were found to affect the performance of governments at delivering services: productive inefficiencies from low-powered incentives, and allocative inefficiencies from political interference in business management (Lunsgaard, 2002). Besides full-privatisation and almost complete withdrawal from the government, a solution to this is to create a market-like environment in the public sector by introducing competition. As explained by Lunsgaard (2002), creating a quasi-market can be done through performance related funding, benchmarking, contracting out, and vouchers. When properly set-up, each of these elements will give an incentive to provide the best-possible service at the lowest costs.
When using benchmarking in order to make government service provision more efficient, a quasi-market situation can thus be created by increasing the competition between its different production units, often departments. Benchmarking mimics competition by replacing a fight for the greatest market-share and the biggest profit by a fight for the highest rank on the benchmark list. This is often called the “naming-and-shaming” mechanism of benchmarks. The question that managers can answer by using benchmarking is “how are we doing compared to the others?”, which fulfils a monitoring purpose. But this monitoring function is assumed to increase the stakes since organisations do not want to look inefficient to others or to higher organisational levels.

A well-know example of an international benchmark is the Open Method of Coordination of the European Union, and in particular the European Employment Strategy (EES). In this macro-benchmark, the countries of the European Unions agree on common sets of goals regarding employment (De la Porte, Pochet and Room, 2001). These objectives are then translated into impact and output indicators on which each member country must report yearly (De la Porte, Pochet and Room, 2001). In fact, what happens is that the European Commissions asks each member country to report on the steps taken to reach the EES goals, and recommendations on best-practices are based on the reports from those countries that scored well on the benchmark (Trubek and Moscher, 2003). These best-practices can thus not be drawn from the benchmark indicators, as they only refer to the results of policies. Even though the EES is considered by many as a great tool for fostering mutual learning, the OMC is broadly acknowledged to aim more at creating a competitive field in which each country is trying to improve its ranking with respect to other countries. In order to improve their ranking on the benchmark, countries will reform their employment policies to increase effectiveness and efficiency, and in that way improve their performance level regarding the ESS indicators. Although the Open Method of Coordination cannot make use of sanctions in order to force countries to adopt the reforms needed to improve performance, it is assumed that the “naming-and-shaming” mechanisms in the benchmark of indicators as well as in the peer-review procedure will put enough pressure on the member countries to reform their policies (De la Porte, Pochet and Room, 2001). Zeitling (2005) indeed mentions how there is evidence that rankings do have an influence on the behaviour of European countries, partially since countries see these rankings as a potential source of political embarrassment if these are picked up by opposition parties, interest groups and the media.
It has nevertheless been recognised that competitive incentives arising from benchmarking could be made more powerful by attaching other instruments to the “naming-and-shaming” mechanism. In fact, in many countries, competition has been introduced not only by introducing meso-benchmarks between different departments or local governments, but rather by various combinations of performance-related funding, benchmarking, contracting-out, and vouchers (see Sol, forthcoming; and Lunsgaard, 2002). It is believed that these four elements and their combined actions can result in even more efficient systems. This is surely true when combining benchmarking with some type of financial incentives, such as contracting out and performance-related pay. If a high rank on the benchmark will assure a more interesting contracting position or increase the amount of resources to be allocated, competition for the best rank in the benchmark will surely increase.

Besides creating a competitive playing field, benchmarking can also improve performance through allowing for a closer monitoring of those responsible for the direct delivery of the programs to the population. This need for increased monitoring arises from the increased decentralisation which is found in the delivery of active labour market policy in many countries. As explained by Finn (2000), the reason for the inefficiency of bureaucracies was thought to be the way decision-making was done in a centralised way. Typically, national programs were designed by the ministry of social affairs of the country and implemented in a uniform fashion for all eligible participants. These employment programs could thus be characterised as “one size fits all” program, although some were targeted at some specific groups such as young people or immigrants. However, such a centralised decision making apparatus did not allow for much variation between regions, cities, or even individuals, and was in consequence believed to be inefficient at meeting needs. The rise of New Public Management meant that the delivery of policies was greatly decentralised and that the implementation of policies was allowed to vary according to specific needs. As a result, local governments in some countries were given full power over the delivery of ALMP, while in other countries agencies were allowed to influence the form the policy would take on a local level. Another reason why decentralisation was thought to increase efficiency is that experiments were made possible, where different types of approaches could be tested on their results. By allowing variation in the design of programs, decentralisation meant that best-practices could be highlighted.

Furthermore, decentralisation also meant that the decision-making within departments responsible for labour market policies was also brought to a lower level. As opposed to the decentralisation of the power to design
policy, called political decentralisation, this type of decentralisation is of administrative value, and is consequently called administrative decentralisation (Moreno and McEwen, 2005, p.12). Moyihan and Pandey (2004) indeed showed that effectiveness was increased by delegating decision-making at the lowest possible level, thus confirming the assumptions made by New Public Management. As a result of NPM being implemented as a new governance model, case-managers saw the leeway they were granted in the implementation of the programs increase greatly.

However, the increased distance between those who manage and those who deliver the services meant that a principal-agent relationship was created. In a principal-agent relationship, the agent carries out work on behalf of the principal. However, the interest of the principal - efficient service delivery - does not always coincide with the interest of the agent (Mosley and Sol, 2005, p.4). Managers found themselves in need of instruments to diminish the perverse effect of information asymmetries typical to principal-agent types of relationship, as these effects can hurt the overall performance level of the system.

Firstly, benchmarking can alleviate some of the perverse effect of the principal-agent relationship between the central government and the agency or local government. By monitoring the results of the programs of each delivery unit, the central government can make sure that the agents are acting in line with their objectives. Secondly, the privatization of service delivery raises some issues concerning the accountability of those private sector agencies that are providing services in a publicly financed sector. By benchmarking the approach and the results of private providers, the government is able to alleviate some of the information asymmetries and act consequently to the findings of the benchmark.

However, the benchmarking targets must be very carefully set as they may create room for goal-displacement. Goal-displacement occurs when too much emphasis is put on reaching targets at the expense of achieving overall effectiveness (Van Bruggen, 1998). Those goals that are left unmeasured or improperly measured will consequently lose importance for the organisation being benchmarked. De Bruijn (2001) showed how performance measurement of public servants must be carefully built if it wants to avoid creating more disincentives than incentives to increase productivity. Three principles should be followed for sound performance measurement of the public sector: 1) Interaction between the principal and the agent on the design, use and implications of the performance measurement, 2) Diversity in the choice of indicators and in their definition, and 3) Hybrid use of process-measurement and product-
measurement which are dynamically interpreted. One of the major points De Bruijn (2001) made in his book is that the more performance measurement is used for control, the more easily the perverse effects of “gaming the numbers” will arise. While the discussion above points towards the need to increase monitoring of the agents in a principal-agent relationship, this does not imply that a “command-and-control” benchmark must be set. As noted also by Hupe and Hill (2007), this top-down view is highly outdated and normatively biased. In fact, a benchmark at the organisational level should be seen more as a contractual agreement between the professional case-manager and his/her supervisor, similar to the contractual agreements which arise from other levels of decentralisation. De Bruijn (2001) actually stressed how both the professionals and their supervisors must be equal partners in designing and using a system of performance measurement by which all parties can meet their needs. For the supervisors, this means the need to communicate goals and monitor their objectives. The next property then relates to the needs of the professionals for instruments which can facilitate the legitimization of actions and decisions.

The third and last way benchmarking can assist management is by being instrumental in legitimizing actions and decisions. With the rise of New Public Management, bureaucracies were also criticised for the scattered nature of service delivery, which was the product of the specialised nature of public work (Finn, 2000). In the case of labour market policy, it was the situation in most country that one office was responsible for the disbursement of benefits, while another office provided for employment services. New Public Management brought changes to this segregated structure since it was believed that the individual needs would be better met by combining functions that are closely related to each other. New Public Management thus meant the integration and coordination of services that were to be delivered by integrated agencies, often called “one-stop-shops”.

The integration and coordination of services for the unemployed meant that much more responsibilities were concentrated in one single organisation. Since large amounts of public resources are transferred to organisations which control the largest part of labour market policies - both passive and active - the dangers resulting from the principal-agent relationship are amplified. Benchmarking at the meso-level can thus be used by organisations, either public or private, in order to demonstrate their performance level to their “principal”. Indeed, since much freedom is given to the agencies in the way they choose to approach the delivery of the services, there is an increased need for them to legitimize their decision-making. One aspect of this is political, since the central
government level is no longer designing the policies in the details, but is still required to be accountable to its electorate. Benchmarking can thus assure politicians that agencies are employing strategies which are efficient and effective according to the benchmark, and are able to take actions towards those agencies that are performing poorly. Agencies that are taking a different course than they used to, for example by adopting a Work First approach to labour market policies, might also feel the need to demonstrate that their approach is successful even though it faces much political opposition.

At the micro-level, combining the effect of the trend of decentralisation with this trend of integration gave rise to an increase in the degree of professionalism of case-managers. Under NPM, the case-manager became a central actor in the implementation of programs, enjoying an important degree of freedom his/her decision-making and possessing a wide spectrum of available options (Hupe and Hill, 2007). In other words, the case manager not only has a large degree of power in deciding whether a claimant has the right to a benefit and whether he or she must participate in an employment program, but the case manager also has many different instruments available from which he/she can choose to give the best possible assistance to the claimant. Consequently, professionals might need to legitimize certain choices that conflict with the usual way of implementing the program. This need for legitimization can also be fulfilled by an evaluation model which allows seeing whether the interdependent work of the case-managers is performing up to standard. As Lipsky (1980) argued in his classic work on street-level bureaucracy, public policy is made up of the individual decisions of the professional case-managers which may or may not coincide with what the “top-floor policy-makers” have in mind with the policy. Benchmarking can therefore be used by professionals to legitimize their actions and decisions. In addition, the case-manager is increasingly provided with the possibility of evidence-based actions, as the performance measurement of his/her work and the work of his/her colleagues allows for the evaluation of what has worked, and what has not.

Discussing these three management functions of benchmarking could easily make-up a research topic on its own. Since this benchmark does not intend to assist managers in a direct way in the management of their work-based employment programs, the objective of learning from best-practices will be the only objective on which this research will be based. Without going into more detail on the topic of New Public Management, it should be emphasized how the use of benchmarking for management purposes is not straight-forward and a great deal of care should be given to the perverse effects performance measurement could have on
productivity, especially when individual case-workers are being benchmarked. The reader is referred to the large body of literature on Management by Objectives and benchmarking for an extensive discussion on how to best set-up a benchmark such that these various management objectives are properly attained (see amongst others, Grubb (2004); Mosley, Schutz and Breyer (2001); Heinrich (2002 and 2007); Van Bruggen (1998), De Bruijn (2001)).

2.3. Benchmarking methodology

The first section of this chapter explained the concept of benchmarking for social policy, or so-called social benchmarking. The previous section discussed how benchmarking can be very useful for the evaluation of labour market policy. In other words, these previous sections have served in answering two crucial questions: first, “what is social benchmarking?” and second, “why use social benchmarking for evaluating active labour market policies?”. The rest of this thesis will therefore concentrate on the more arduous question of “how to benchmark active labour market policies”.

When using a benchmark in order to measure the performance of a program, four steps must be undertaken. The first step is to choose the type of benchmark to be used based on the needs of policy-makers in terms of what the benchmark should achieve. Second, the performance standard has to be chosen against which the performance of the organisation will be measured. Third, the indicators which will make-up the benchmark must be chosen, and last, these must be measured in each program and an analysis of these results must be performed.

Step 1: identifying the aim and choosing the benchmark model

The first step to be undertaken is to identify the aim of the benchmark. This sounds rather logical, but can quickly be ignored by those who take part in a benchmark just in order to follow the “trend”. Indeed, as benchmarking becomes increasingly popular in the public arena, taking part in a benchmark could be used more as window-dressing than for real performance improvement. The public organisation that performs a benchmark could therefore use this exercise only to pretend efficiency and effectiveness are strived for, while the lessons from this benchmark would not have any purpose in reality. By clearly stating the aim of the benchmark, such window-dressing can be avoided, since it will be made explicit how this benchmark will contribute to policy-making. Obviously,
the benchmark will only be useful to the policy-makers if it can contribute to their decision-making process in one way or another.

From the discussion in sub-section 2.2.1, it is clear that evaluations of ALMP programs aim to answer the questions of “what works best and why”. The works of, amongst others, Schmid, O’Reilly and Schömann (1997), Mosley and Mayer (1999), Meager and Evans (1997), Pierre (1999), Grubb (2004) and Schütz, Speckesser and Schmid (1998) all point towards evaluation practices that address the “bigger picture” of the policy program. Their conclusions insist on looking further than impacts or outputs only. This means that the social benchmark model will need to follow the holistic approach which characterises best-performance benchmarks. The holistic approach means that evaluations of active labour market policies should not only look at the impact of the policies, but should assess the whole policy chain. By following the policy chain, a holistic evaluation would start by looking at the input of the system, it should then go on evaluating its method of implementation, then measure its output, and finally look at the impact the policy has on its beneficiaries and their environment. The five elements of the policy chain will thus form the core of the social benchmark for active labour market policies.

That is not to say that all benchmarks of active labour market policies will necessarily have the intention to evaluate those policies in order to learn from best-practices. As discussed in the previous sub-section, benchmarks can also be set-up for the purpose of making the delivery of the programs more efficient. The choice of the type of benchmark model will, however, not be influenced by this and a best-practice benchmark is still deemed most appropriate. This is because for all benchmarks the key to their usefulness lies in the fact that they provide an explanation of the level of performance reached. For example, while one could think an output or an impact benchmark would be enough in order to create a competitive playing field, the reason behind differences in competitiveness will remain unknown. Different ways to improve competitiveness can then not be easily found, and further research on the source of performance gaps will be needed. This could have already been achieved if the benchmark would have been a best-practice benchmark.

Step 2: Choosing the performance standard and level of analysis

Choosing the performance standard means choosing the value against which the performance of the programs will be compared. As discussed at the beginning of this section, Xerox Corporation was the initiator of the benchmark trend in the 80’s. Interestingly enough, Xerox Corporation did not benchmark itself against competitors in its branch, but sought after
companies that were doing extremely well in one part of Xerox's production chain, for example warehouse logistics or human resources (Schütz et al. 1998). These types of operations are fairly similar between significantly different companies, making "learning-from-the-best" possible, even outside a company's sector of production. However, benchmarking has become increasingly focused on direct competitors, and this type of methodology is now the most common in both the private and the public sector.

According to Grubb (2004) and Mosley and Mayer (1999), three general methods can be used to define a benchmark standard:

- Theoretical standards: based purely on what theories say the performance value should be.
- Institutional benchmark: the performance targets are set by leadership or management, on an ad hoc basis.
- Best-performance benchmarks: the best performing program is the standard.

The problem of using theoretical standards is that they difficulty include qualitative indicators, as these are often too complex to include in predictive models. For impact-benchmarking, this should be less of a problem than for the other types of benchmarks, since most impact indicators are quantitative. In addition, theoretical goals are often unclear or ambiguous (Mosley and Mayer, 1999). On the other hand, by using ad hoc values coming from institutional choices as the performance standard, the benchmark might be legitimate for those who participated in the negotiation, but is denied of any further scientific ground (Grubb, 2004). This makes the lessons to be learned from the benchmark thornier to accept by outsiders who may object the use of that particular standard. A performance standard set by management could be applicable in some cases, since management might wish to attain a certain level of output or impact without this being based directly on theory or best-practice. Clearly, benchmarking based on the performance level of the best-performing program is the most objective method of the three. This will thus be the method used in the social benchmark of labour market programs which is being constructed. All rankings will consequently be made with respect to the best performing program for each indicator separately.

This benchmark compares programs in different countries, and can thus also be described as an international benchmark. As mentioned by Helgason (1997), international benchmarking in particular can be very useful when trying to assess the general level of performance of a public sector program, since benchmarking across countries may reveal more in
In terms of best-practices. Also, it is an excellent way to find innovative and alternative ways of providing a service or designing a policy that are not otherwise found within the country itself. This means that the best-practice standard could very often refer to programs found abroad. However, since choosing which international programs should take part in this benchmark requires more knowledge about work-based employment programs, this discussion is postponed to the next chapter. At that point, after having introduced and defined work-based employment programs, the reasons behind the choice for each of the programs will be given.

It is, however, crucial to note at this point that international benchmarks will almost always require choosing the best-practice type of benchmark model. A major criticism of many benchmarks is that comparisons of performance are not possible because of different backgrounds and underlying structures of the programs to be compared. This is even truer when talking about international benchmarks, which try to compare policies and services that are based on varying national structures (Schütz et al., 1998). Evaluators have always been warned about the dangers of comparing programs that are actually very different from each other. The expression “comparing apples to oranges” has often been used in order to warn for the dangers of concluding that one program is better than another program, while the very different nature of these two programs has been ignored. In fact, it is not the comparison of “apples to oranges” which is dangerous, but what is inappropriate is to conclude that “this apple is very bad”, when in fact it is an “orange” you have in hand. What is therefore crucial for international benchmarks that might face important differences in the evaluated programs is first to clearly define the type of program to be evaluated. That way, a set of basic common denominators will be established. Then, it is imperative that the indicators found in the benchmark model are able to take into account the variations between the different countries. Input, process and external factors indicators will thus be crucial to the understanding of the differences in the results of the programs. Obviously, the more different the programs are from each other, the more indicators are needed in order to take into account these differences.

Step 3: Choosing the indicators

As can be imagined, the benchmark model will be very sensitive to the indicators that are chosen. For this reason, the choice of the specific indicators is considered to be a part of the research on its own, and thus not a part of the general social benchmark model. In other words, extensive knowledge of the type program to be benchmarked is needed in
order to choose the proper performance indicators. This refers mostly to the process variables, since this element is made up of various components that all can have a different influence on the output and the impact of the program. Empirical studies or theoretical arguments should be surveyed in order to find the most relevant components of the process, which should be included in the benchmark.

While choosing the indicators, it should be kept in mind that these can be of either quantitative or of more qualitative nature. Mosley and Mayer (1999) pointed out the importance of giving greater consideration to qualitative dimensions of labour market performance. These qualitative indicators will nevertheless need to be quantified in some way in order to be incorporated in the benchmark. By using a ranking from low to high, poor to good, a benchmark model can make sure not only the purely quantitative indicators are included in the model. As will be seen in chapter 4, many of the input, process and output indicators are actually of qualitative nature. This surely will add to the understanding of all the facets of the programs in the benchmark, and not just of those facets which are already quantified.

Moreover, sensitivity analyses could be important to consider when choosing the indicators, since performance levels can easily be influenced by different factors. First, in case a choice has to be made between indicators that cannot be dictated by theoretical and practical considerations, the difference in overall performance should be calculated when different choices are made. Second, should an assumption be made about an indicator (for example a discount rate or an exchange rate) a sensitivity analysis should be performed to see whether the results are affected by changes in the assumption (Pierre, 1999).

Clearly, the specific indicators are difficult to put into a general blue-print for social benchmarking. Additionally, the availability and contextual relevance of the chosen indicators should be considered, especially considering the countries to be included in the evaluation. The next two chapters of this research will therefore concentrate on defining those performance indicators that are necessary for benchmarking work-based employment programs. Indeed, while the research question relates to evaluating active labour market programs in general, it will be necessary to define a narrower type of program for the benchmarking exercise. This is because it would be impossible to define a concise list of performance indicators that would be useful for benchmarking the complete active labour market program of different countries. Such a benchmarking exercise would go beyond the scope of this research project. Hence, in order to keep the benchmark model manageable, a single activation
program will be benchmarked as an example of how activation programs in general can be benchmarked. Chapter 1 has already explained that work-based employment programs have been chosen as this specific program to be benchmarked, mostly because of their increased popularity in many countries, but also due to the large amount of criticism these programs raise. All things considered, chapter 4 will thus present the specific performance indicators that are relevant for the benchmarking of work-based employment programs.

Step 4: Measuring the indicators and aggregating the benchmark

The crucial step of a benchmarking exercise is undoubtedly the collection of data, since the scope of the evaluation will depend on the quantity and quality of the data used (Pierre, 1999). Three types of data can be used for social benchmarking: survey data, administrative data and secondary source data from reports and academic research. Survey data originates from interviews with individuals. The main disadvantage of such database is that it may be biased and thus not truly represent the reality of the individuals (Pierre, 1999). It is, however, very useful to use international databases which are based on survey data (such as the European Community Household Panel database) whenever possible, since this will foster comparability. Various types of surveys are performed by many stakeholders, and a careful selection of the data needs to be performed in order to retain only the most reliable studies. Second, administrative data comes from administrative records of the agencies part of the policy chain. This data is generally very reliable and free of sampling error (Pierre, 1999). However, this data is created for administrative purposes and thus offers only limited scope for academic research (Pierre, 1999). The last source of data is academic research, evaluation reports, and other secondary sources of data. The benchmark model as defined earlier attempts to analyze the full context of the policy and services, and it will need to look broader than data found in administrative records and surveys. Most data will thus need to originate from secondary sources of reliable quality, such as ministerial reports, academic research, and legal documents. Reliability will be a key issue here, which should be attained if internationally well trusted sources are used. Pierre (1999) suggested that in order to improve the quality of the benchmark and compensate for the weaknesses of different types of data, data from various sources should be used. This means finding the most appropriate data source for each indicator of the model. The author also mentions that qualitative data has been found to be a good complement to quantitative data.
In order to be able to conclude on the effectiveness and the efficiency of the programs, these indicators will therefore need to be aggregated into one single indicator representing each of the links of the policy chain. This is not to say that no conclusion will be possible when the indicators are taken on a single basis. On the contrary, some important conclusions on the programs will be reached while comparing single indicators, such as the sanctioning procedure and quantity of sanctions used. Nevertheless, a measure of the size of the input as a whole and the extent to which the process as a whole can be expected to reach good output and impact will also be needed in order to make generalised conclusions on the performance of the different programs. What is thus needed is an aggregation technique, which makes it possible to analyse single indicators and calculate an overall score for each component of the best-practice benchmark.

One technique which will make these two conditions possible is the use of radar charts. Radar charts have been proposed by Schütz et al. (1998), Tronti (1998), Mosley and Mayer (1999) and Jones (2002) as a way to clearly represent performance indicators. A radar chart is made of several axes each corresponding to one performance indicator, the highest value on the axis corresponding to the performance standard. For each country, the values of each axis are connected in order to allow for clear comparison. Moreover, radar charts can be used not only to compare different countries, but also to compare the same country at different point in time. Figure 2 shows a radar chart for two countries with six performance indicators.

Figure 2: A Radar Chart for two countries and six performance indicators
Using radar charts for benchmarking has two main advantages. As already mentioned, radar charts provide a simplified presentation of multiple performance indicators in a way that is intuitive, even to non-experts and can thus be used by the many stakeholders of the evaluation (Mosley and Mayer, 1999). Indeed, radar charts are most useful for comparing performance on multiple dimensions simultaneously (Mosley and Mayer, 1999). For example, they have been used by Bonny and Bosco (2002) in order to illustrate the various dimensions of income support in 13 European cities. Moreover, Plantenga and Hensen (1999) used a benchmark that was made operational by radar charts to compare the level of gender equality in European countries.

The second main advantage of using radar charts is that the surface area that is formed by the joined lines of each performance indicator generates one single aggregated performance indicator (Schütz et al., 1998). This approach is also called the SMOP-approach, which stands for “Surface Measure of Overall Performance”. The formula for calculating this measure of performance is shown in appendix A. Consequently, the SMOP value can be used to aggregate the benchmark indicators in order to present both complete and deep information, as well as easily readable and summarised information. The previous sections showed that detailed performance indicators are needed to be able to analyse the source of performance gaps and their possible remedies. Yet, without an index or aggregation of these performance indicators, it becomes very difficult to see the relative position of different countries, since they might score high on some indicators by low on others. Radar charts are thus an excellent instrument for benchmarking, since the score on each indicator is left intact and visually analysable, while an aggregate performance measure is also made available. The benchmark model in part 2 of this research will be aggregated with the use of a radar chart and the SMOP for each of the category of indicator, that is to say, the input, process, output, impact and external factors.

2.4. Conclusions

This research has the objective to first develop a benchmark model which can be used for the evaluation of employment programs and then to use this benchmark on work-based employment programs around the world. Within this chapter, the first two steps of the benchmarking methodology have been performed. First, a best-practice benchmark model was chosen in which the whole policy chain would be included as part of the performance indicators. Second, the performance of the best program in
the benchmark will be the standard against which each program will be judged. Also, chapter one already discussed which programs will be included in the benchmark.

This chapter has already answered the first sub-question of this research by presenting different types of benchmarks and showing how in particular best-practice benchmarks based on the entire policy chain can be useful for evaluating active labour market policies. The benchmark model has been presented, but the performance indicators have been purposefully left abstract. This is because the choice of the indicators will mostly depend on the specific type of policies to be benchmarked. As a result, the discussion on social benchmarking in the specific case of work-based employment programs will now become the focal point of the research.