International conference on Public Policy (ICPP)

26-28 June 2013 Grenoble, France

Panel 16. Experimenting with Public Policy: Issues of Design and Effectiveness

Evaluating What Works for Whom in Employment Services

Thomas Bredgaard, Associate Professor, Ph.D.

Stella Mia Sieling-Monas, Ph.d.-student

Julia Salado-Rasmussen, Ph.d.-student

Centre for Labour Market Research & Research Centre for Evaluation

Department of Political Science

Aalborg University

Fibigerstraede 1

DK-9220 Aalborg

Denmark

Email: thomas@dps.aau.dk

Ph. +45 9940 2605 / +45 2724 2128

Abstract

This paper describes a new framework that is suitable to evaluate complex policy experiments. The complexity and innovativeness of policy experimentation makes standard evaluation methods unsuitable. We use policy experimentation in employment services as a case to illustrate the shortcomings of experimental evaluation and propose to combine its strengths with program theory evaluation. Such an approach would make it possible to study what works for whom, under what circumstances. The methodology is illustrated with literature reviews of different program evaluations of employment services and with reference to two current evaluations of employment policies towards unemployed people on social assistance and sickness benefits.

Introduction

If policy experimentation is defined as an intervention to test the effects of a policy in a real world setting, then policy experimentation is on the rise in numerous policy areas. It may carry different labels in different contexts like pilot programs, controlled experiments, earmarked budgets etc. The demand for experimentation among policy makers can be understood in the context of shrinking public budgets and the rise of evidence-based policy-making and practice (cf. Pawson 2006). Experimentation and evaluation are combined to understand what works for whom and prioritise scarce resources. Experimentation is also necessary due to the complexity and interconnectedness of many policy problems.

As policymakers and public managers turn towards researchers and the evaluation community to learn what works for whom and which interventions are most cost-effective they are nevertheless often disappointed. The messages from the scientific literature and evaluations are neither clear-cut nor easily transferred to political decision-making.

To become applicable, research findings and evaluation results need to be simplified and summarised. Increasingly, empirical findings are ranked in a hierarchy, where randomized controlled experiments and econometric impact evaluations are considered the “golden standard” for evaluating what works. In systematic reviews, the best available evidence is summarised and quantified, sometimes into a single outcome measure, to conclude whether or not a program or intervention works.

This simplification process is problematic. First, the causal relations between interventions and outcomes tend to become a "black box". That makes it difficult, sometimes impossible, to answer policy-relevant questions like: Why did an intervention work or why did it not work? If it worked, for whom did it work? When did it work, and under what circumstances? Second, if complicated and complex policy experiments, programs and interventions are simplified to make experimental evaluation feasible, then there is a clear risk of misleading generalizations and policy learning.

The article proceeds as follows. First, we describe the current state of affairs in program evaluation of employment policy and the shift towards a simplistic notion of problems, interventions, and outcomes. Second, we make suggestions for overcoming the shortcomings of this mainstream version of "evidence-based policymaking". The idea is to design evaluations that combine impact evaluations with program theory evaluations to arrive at more detailed answers to the questions of what works for whom under what circumstances.

Program evaluation of employment policies

We define employment policies as a group of selective policies aimed at situations on the labour market defined as problematic (e.g. mismatch between supply and demand for labour, unemployment, skills shortages, discrimination etc.). Others use the term labour market policy. There is of course great variation in what is defined as a problematic situation in the labour market, thereby requiring public intervention. At a general level, the objective of labour market interventions is to improve the functioning of labour markets in achieving politically desirable outcomes (Schmid et al. 1997; Bredgaard et al. 2011a).

There are numerous classifications of the different functions, dimensions, and instruments of employment policies. One of the most simple and widely used is the distinction between passive (compensatory) and active (interventionist) employment policy. Passive employment policies comprise public income benefits to the unemployed or inactive (unemployment insurance benefits, social assistance/cash benefits, early retirement benefits etc.).

In this article we focus on active employment policies. Their main objective is to reduce unemployment and improve the employment opportunities of the unemployed. Different measures are used to achieve this objective. The OECD and Eurostat define active labour market policies as comprising the following types of programs:

(1) Labour market training (classroom training, one-the-job training, and work experience). Training includes both general education and specific vocational skills. The main objective is to enhance the productivity and employability of participants and to enhance human capital.

(2) Private sector incentive programs (wage subsidy and self-employment grants) aim at creating incentives to alter the behaviour of employers and/or workers (e.g. encourage employers to hire long-term unemployed).

(3) Direct employment programs in the public sector comprise the production and provision of public works or other activities that produce public goods and services. These measures are typically targeted at the most disadvantaged and aim at keeping them in contact with the labour market to preclude the loss of human capital. However, the jobs are often additionally generated jobs not close to the actual labour market.

(4) Job search assistance (job search courses, job clubs, vocational guidance, counselling and monitoring, and sanctions) are geared toward increasing the efficiency of the job matching process.

When defining active programs this way, it becomes clear that they do not in themselves create new (unsubsidised) jobs, at least not in the short term. It is in combination with other public policies (like macro-economic policy, industrial policy, and education policy) that employment policies may assist in boosting job creation in the medium and long term. This is an important point when we measure the impact and success of active employment policies.

The evaluation of active employment policies have been dominated by program evaluation of the outcomes of interventions (e.g. impact evaluation, randomized controlled experiments, and econometric evaluation). Impacts are assessed by comparing measurable outcomes for the participant group with a reference group of non-participants, so as to estimate the net impact of the program or intervention. A specified period after the termination of the program, outcomes are measured as changes in employment, unemployment, income or wages.

Impact evaluations do not necessarily document that changes in outcomes are the causal effect of the program. That depends on the possibilities for establishing a “true counterfactual” situation (a situation where all but the intervention is equal in the participant and control group). Moreover, impact evaluations

have difficulties in explaining which elements of the program that produce outcomes. By implication, the content of programs and their implementation may become a black box.

In the impact evaluation literature, three important outcomes of active employment programs have been distinguished:

1. Motivation effects: Mandatory job search requirements or obligations to participate in active programs may increase the transition rate from unemployment to employment before participation.
2. Locking-in effects: Participants in active programs may reduce their job search intensity during participation, thereby reducing their transition rate from unemployment to employment.
3. Participation effects: Participants may improve their qualifications and competencies during participation, thereby improving their employment opportunities.

These are the direct effects of active employment programs. Outcomes are measured in terms of unemployment rates, employment rates, income or wages. These indicators are observable and measurable if data are available and the "right" methods and techniques are applied, like randomised controlled trials, quasi-experiments, matching techniques, duration analysis etc. There are, in addition, some important indirect effects, which are more difficult to observe and measure:

4. Selection effects: The chances of finding regular employment may not be significantly higher for program participants than non-participants because some participants are selected at the expense of other participants (e.g. creaming effects).
5. Deadweight effects: Employers may hire subsidized program participants that they would have hired even in the absence of employment subsidies.
6. Displacement effects: The policy may have improved the employment opportunities of participants at the cost of a decline in job opportunities for non-participants.
7. Substitution effects: A program may have unintended side effects that lie beyond the intended target area, for instance, firms employing subsidized program participants gain a competitive advantage over firms not benefitting from program subsidies.

Selection effects are addressed in microeconomic evaluations that strive to establish a counterfactual basis against which to compare net program outcomes. Here randomized controlled experiments are the optimal solution (cf. Heckman & Smith and Björklund & Regnér in Schmid et al. 1997). Deadweight, displacement,

and substitution effects can be addressed by aggregate impact analysis with the aim of measuring the general effects of labour market policy on macroeconomic performance (like aggregate employment, unemployment, and wages).

Reviewing the evidence: What works?

When asking "what works" in employment policies we are directed towards impact evaluations, quantitative methods, econometrics, and randomized controlled experiments. Other types of evidence based on more qualitative methods as well as process or formative evaluation tend to be excluded, even before their contribution to the evidence base is considered.[1]

Before we discuss the shortcomings of this scientific reductionism and possible alternatives, we briefly review some of the current evidence on what works in active employment policies. It is neither possible nor necessary to review all the evidence. Instead we select the most widely used and cited references in the literature on impact evaluation. In recent years, a number of meta-evaluations of the impact of active labour market policies have been undertaken.[2]

One prominent example is Card, Kluve, and Weber (2010), who made a statistical meta-analysis of 97 studies conducted between 1995 and 2005 (see also Kluve 2010 for a similar approach). Each study was categorised as either significantly positive, insignificant, or significantly negative. This method allowed the authors to make comparisons across studies that used very different dependent variables (ranging from duration of time in registered unemployment to average quarterly earnings) and very different econometric modelling strategies. The authors found that job search assistance programs were most likely to yield positive impacts especially in the short run, whereas subsidised public sector employment programs were less likely to yield positive impacts. Classroom and on-the-job training programs yielded relatively positive impacts in the medium term (after two years); although in the short term these programs often had insignificant or negative impacts. Comparing across participant groups, they found that youth programs were less likely to yield positive impacts than untargeted programs. There were no large or systematic differences by gender.

Martin and Grubb (2001) did not use statistical analysis of program impacts, but made a narrative meta-analysis of the experiences of OECD countries with active labour market programs. They drew lessons from the impact evaluation literature and combined the insights with country reviews and analytical studies of

the interactions between active and passive labour market policies and the role of public employment services. They concluded, among other things, that evaluation findings were not terribly encouraging, especially for disadvantaged youth. But there were some success stories. Job search assistance, wage subsidies in the private sector, and labour market training did seem to work for some groups, although their impacts were not always large.

Another study that is often quoted in the economic impact literature is Calmfors, Forslund, and Hemström (2002). They reviewed the lessons of the Swedish active labour market policy of the 1990s. The authors concluded that active labour market policies probably reduced open unemployment, but also reduced regular employment mainly because of displacement effects. Their policy recommendation was not to use active labour market policies on such a large scale as in Sweden in the 1990s, and especially not to use active labour market policies as a means to renew unemployment benefit eligibility.

In the UK, Hasluck and Green (2007) made a meta-analysis of the evidence on what works for whom in the various programs implemented by the jobcentres towards specific target groups. The review mainly summarised the research and in-house publications from the Department for Work and Pensions and Jobcentre Plus.

A recent review of economic and quantitative impacts evaluations was carried out by Rosholm and Svarer (2011). The study was commissioned by the Danish national labour market administration with a special focus on the impact of activation programs implemented in (private or public) enterprises. Around 30 international and Danish articles were selected and reviewed. They authors found strong empirical evidence in favour of using private job training while public job training was mainly found to have positive effects when used on disadvantaged unemployed.

Few evaluations investigate whether subsidised employment in the private sector has a displacement effect. One exception is Pons and Arendt (2010) who found that in Denmark private job training did not seem to displace ordinary employees and that the net employment rates in the private companies increased. Similarly, the Danish Economic Council in a recent impact evaluation controlled for possible displacement effects, and found that subsidised employment in the private sector, unsubsidised training in the private sector, as well as counselling and training did in fact increase transition rates from unemployment to employment. The results were robust during the economic boom (2006-2008) as well as the current economic slump (2009-2011) (Danish Economic Council 2012).

The majority of impact evaluations show that active programs are least effective towards the most disadvantaged unemployed (cf. Rosholm & Svarer 2009a; Skipper 2010). Graversen (2012) made a review of 35 selected impact evaluations that focus on disadvantaged unemployed (e.g. unemployed on social assistance, unemployed with social, physical, or mental problems, long-term unemployed etc.). He found "strong evidence" that private job training had a positive employment effect and "moderate evidence" that public job training had a positive effect. The findings were corroborated in the report of the Danish Economic Council (2012) mentioned above.

Jensen and Rosholm (2011) made a narrative summary of the micro-econometric literature on the impact of Danish active labour market policy. They emphasised that the major share of the positive impact of activation measures in Denmark originates from the motivation effect, i.e. increased transition rates from unemployment to employment before participation in activation programs. This motivation effect is only found for men. Moreover, randomised quasi-experiments show that especially intensive contact interviews (conducted more often than at the standard three-month interval) and sanctions have a positive motivation effect, especially for less disadvantaged unemployed. Subsidised employment in the private sector (private job training) is found in a number of evaluations to be the most efficient active labour market program. The highest impact was found for unemployed without vocational qualifications and women as well as unemployed above 50 years of age.

In recent years, the Danish labour market administration has commissioned a number of controlled field experiments to study what works for whom in active employment programs. The experiments use randomisation to select participant and control groups. Sometimes impact evaluations are combined with implementation studies to assess whether the experiment was carried out as intended (Graversen et. al 2007; Rambøll 2008, 2009; Rosholm & Svarer 2009a, 2009b). In the most recent experiments, different packages of programs have been introduced to evaluate what type of interventions works best. This idea of combining impact evaluation and process evaluation is similar to the approach I suggest below, except from the important difference that the former does not include explicit program theories in evaluating what works for whom.

What works for whom, under what circumstances?

As we saw above, some field experiments do in fact combine impact evaluations and implementation studies to understand what works for whom. They do not, however, include explicit program theories. That

is problematic because variables and program activities tend to be conflated with mechanisms. Mechanisms explain why programs and experiments work. Mechanisms establish generative causality between programs and outcomes (Pawson & Tilley 1997). They are usually hidden underneath the observable surface of variables and activities (Astbury & Leeuw 2010). Program theories are assumptions about what works for whom in specific contexts (i.e. Context-Mechanism-Outcome configurations). If program theories and mechanisms are not included in the evaluation it becomes difficult to interpret evaluation results correctly. If the program or intervention failed to achieve its intended results, it may be because the program did not work, but it might also be because it was not implemented properly. An evaluation using program theory allows us to systematically distinguish between implementation failure and theory failure (Dahler-Larsen 2001; Funnell & Rogers 2011).

Another challenge with evaluations without program theory occurs if the evaluation finds that the program is successful. We do not know exactly why the program achieved its intended outcomes and sometimes not even whether the outcomes were the result of the program or some other intervening variable or program. That makes learning difficult. In principle we, therefore, need to copy the program exactly as it was for fear of missing something essential. The evaluation does not provide any guidance for adapting the policy to other settings (Funnell & Rogers 2011: 5). This is why, as described above, some field experiments introduce variation in programs and/or target groups (see also Peck 2012 for adaptations of the experimental logic in randomized control groups).

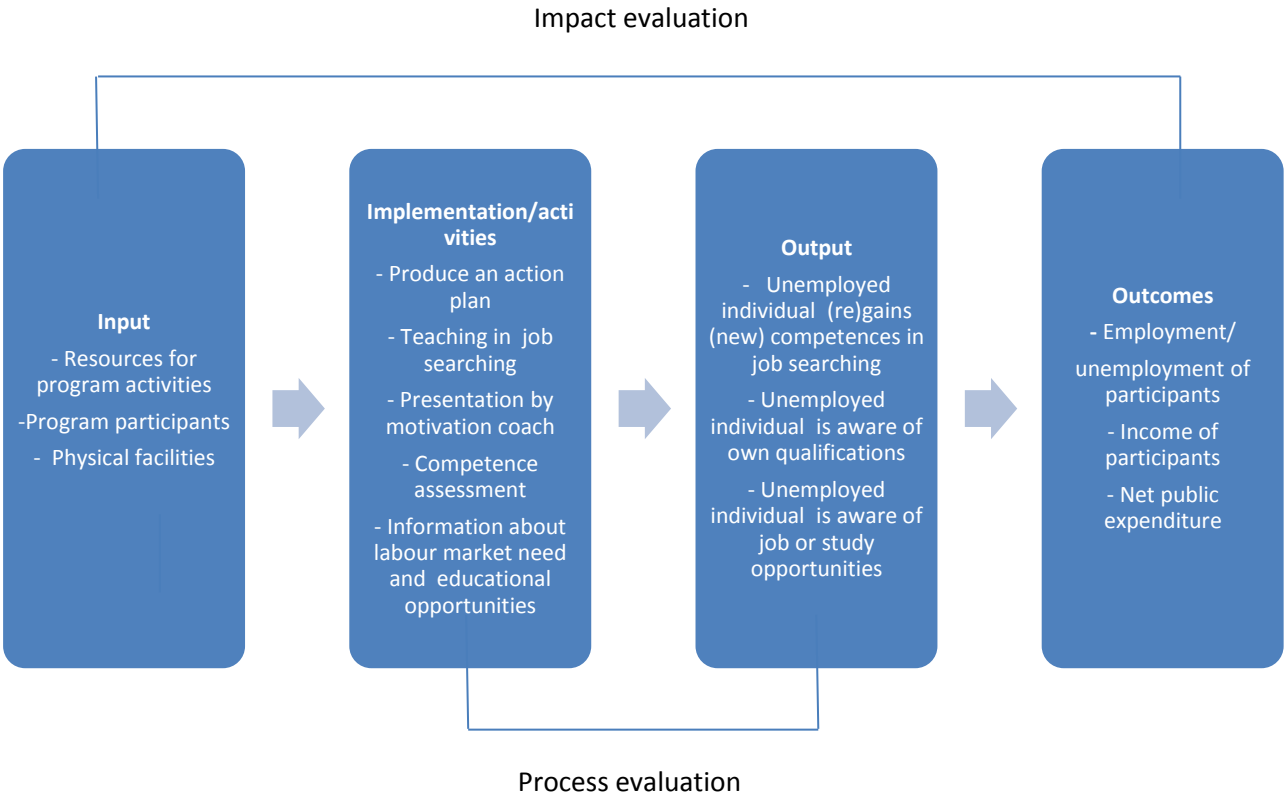
Program theory is useful to identify and evaluate for whom and why a program worked or did not work and under what circumstances (Pawson & Tilley 1997). Program theory can be defined as an explicit theory or model of how an intervention, such as a project, a program, a strategy, an initiative, or a policy, contributes to a chain of intermediate results and finally to the intended or observed outcomes (Funnell & Rogers 2011: xix).[3] A program theory ideally consists of two components: A theory of change (the central processes or drivers by which change comes about for individuals, groups, or communities) and a theory of action (how programs or other interventions are constructed to activate these theories of change).

In the evaluation literature, there is a tendency to treat evaluations with and evaluation without program theory as alternatives (Funnell & Rogers 2011; Pawson & Tilley 1997). The ambition here is to combine the two approaches. Similarly, Cook (2000) argues that treating theory-based evaluation and experimentation as alternatives is a false choice. Bloom (2005) suggests that randomised experiments should be refined by combining them with non-experimental methods.

In the case of active employment programs, we should combine impact evaluation with program theory evaluation. Such an alternative evaluation design can best be described by using an example. In a current evaluation of employment policies towards unemployed people on social assistance we are focusing on job search assistance programs. As we saw above, impact evaluations find that job search assistance programs increases the individual transition rate from unemployment to employment. The obvious conclusion is that job search assistance programs do indeed work, and are an example of an effective employment policy. But we intend to open the "black box" and investigate whether job search assistance programs work for every target group of unemployed, why they work for specific target groups, how they work, and under what circumstances.

In order to design such an evaluation, we need to understand the nature of the intervention and the situation in which the intervention is implemented. In the figure below, we present a simple pipeline diagram for a job search assistance program offered in a local municipality in Denmark.

Figure 1. Evaluating job search assistance programs



Impact evaluation is necessary to evaluate whether the program works and whether it is cost-efficient. In the case above, we need to evaluate the outcomes in terms of subsequent employment, unemployment, and/or income of participants compared to a similar group of non-participants. This can be combined with cost-benefit analysis to evaluate whether the program is cost-efficient.

We need to supplement impact evaluation with process evaluations to understand how the program worked (implementation), what types of outputs the program actually delivered and whether intermediate outcomes were achieved. In the case of job search assistance programs, we can open the "black box" by investigating the most efficient means to enhance the unemployed competences and motivation to job search, to investigate for what type of unemployed the program works best, to attempt to understand why the program works for some groups and not for others, to explore whether there are certain characteristics in the environment of the program that promote or inhibit success, etc. In this case, the construction and test of program theories is precisely a method for fusing impact evaluation with process evaluation.

The nature of the intervention

In order to select the most appropriate evaluation design, it is essential to analyse and understand the nature of the intervention or policy experiment. Experimental evaluations tend to treat all programs and interventions as if they were simple. There are, nevertheless, important aspects of active employment policies that are complicated or complex. This situation can lead to evaluations that are insensitive to the inherent unpredictability of the contexts within which programs function and the way in which they are implemented. Treating an intervention as simple when it in fact has complicated or complex aspects also means disregarding the fact that programs should and can adapt to changing circumstances in the environment and the likelihood, and in many cases desirability, of emergent outcomes (Funnell & Rogers 2011: 48-49). Some researchers, on the other hand, tend to characterise active labour market policies as complex, flexible, and adaptive and unemployment problems as "wicked" and ungovernable (Rothstein 1998: 78). This is not entirely precise either. We need to make a situation analysis of the specific problem and intervention at hand to select the most appropriate evaluation design.

Rather than treating the entire problem as either tame or wicked and the intervention as either simple or complex, we need to acknowledge that policies have simple, complicated, and complex aspects and apply evaluation methods that are appropriate to the situation and type of intervention (cf. Funnell & Rogers

2011). Program theories are a way to combine the simple, complicated, and complex aspects of public interventions.

Table 1. Simple, complicated and complex aspects of interventions

	Simple aspects	**Complicated aspects**	**Complex aspects**
Objectives	Single	Multiple	Emergent
Nature of intervention	Standardised	Objectives and outcomes are valued differently	Non-standardised, changing, adaptive, and emergent
Nature of organisation	Implemented by a single organisation	Implemented by multiple, identifiable organisations in predictable ways	Implemented by multiple organisations with emergent and unpredictable roles
How interventions work	Almost the same everywhere	Different in different situations and for different people or in different environments	Generalisations rapidly decay Results are sensitive to initial conditions and context
Evaluation	Impact evaluation	Impact and process evaluation	Process evaluation

Source: Adapted from Funnell and Rogers (2011): Purposeful Program Theory, San Francisco: Jossey-Bass (p. 73).

Some aspects of programs are simple if they involve a discrete, standardized intervention with an agreed set of objectives that is implemented by one agency and carried out the same way everywhere. For example, the administration of benefit payments can be characterized as a simple intervention in most countries. Eligibility criteria for unemployment insurance benefits are standardised, the payment level is defined in advance for different target groups, there is only one separate organization responsible for transferring the money, the same amount is paid out everywhere for each target group.

Another example is standardised interviews at specified time intervals. A checklist (standard operating procedure) can be followed and job search requirements can be standardised. However, the interpretation

of rules and procedures in each individual case has some more complicated aspects. Impact evaluations have documented positive general outcomes of standardised interviews on transitions to regular employment (Rosholm & Svarer 2010). But we do not know much about why it works, for whom and in which contexts. The causal mechanisms are unclear. Do regular interviews lead to intensified job search activity? Does it lead to more realistic job search activity? What is the impact of the counselor's level of qualifications and knowledge of local labour markets? Impact evaluations cannot respond to these questions.

A simple intervention does not imply that the intervention is trivial or necessarily easy to implement, but it makes evaluation in terms of what works easier. RCT, quasi-experiments, and impact evaluations are suitable for evaluating whether the program works or not. Program theory can supplement these methods in order to open up the "black box".

In practice, few, if any, interventions meet exactly the criterion of simplicity. Most interventions, especially in interventionist human services like employment policy, also have complicated and complex aspects. An intervention has complicated aspects when different individuals or organizations value impacts differently, when the intended impacts are multiple and competing, or when impacts are needed at different levels of a system. A program may just be one piece of the puzzle needed to achieve intended results, and it will work only if the other components are in place, for example, other interventions, a favourable implementation environment, or particular participation characteristics (Funnell & Rogers 2011: 75). Active employment policies have many complicated aspects: Political disagreement over objectives and intended impacts, various implementing agencies with different, and sometimes competing, interests (public and private employment services, unemployment insurance funds, trade unions, employers etc.) and fluctuating environments (changing business cycles, different local contexts). Evaluations using program theory can recognize different legitimate views on the objectives of an intervention and address the different intended outcomes of an intervention. If agreement cannot be reached or negotiated, multiple program theories can be constructed (Funnell & Rogers 2011: chapter 6; Hansen and Vedung 2010). Several chains of assumptions of different theories can be included in the program theory to see which of them is best supported by data (Weiss 2000).

Complex interventions are dynamic and responsive to changing needs, opportunities, and challenges. Both intermediate as well as long-term outcomes can be adaptive and emergent. Funnell and Rogers (2011: 79ff.) argue that conventional evaluation methods are challenged when interventions change substantially

over time and specific impacts cannot be identified in advance. The solution is not to make program theories increasingly complicated, but rather to revise and adapt them accordingly. Evaluation design and measures must also become dynamic and emergent. An example is employment programs for young unemployed individuals with complex and interrelated problems (cf. Bredgaard et al. 2011b). Impact evaluations find very limited evidence of success in integrating this target group on the labour market. There is no standardised recipe or simple intervention that will be successful in all such cases. The public intervention should in theory be adapted to the needs and capacity of each individual participant and local community. Each situation and individual should be treated as unique, which makes it difficult to specify objectives, outcomes, and causes and effects in advance. Multiple organisations need to work in cooperation to be successful. Such emergent and dynamic interventions and situations are difficult to evaluate, but program theories are helpful in clarifying what actually happened and in identifying lessons that may be generalised from the specific case.

Causal inference

One of the main challenges in program theory evaluations is that of causal inference (Davidson 2000). Program theory is intimately linked to concerns about causality. A good program theory describes the causal links between the program and the intermediate and ultimate outcomes, and tests whether or not they worked as intended. Program theory should include testable causal hypotheses about what works for whom, under which circumstances.

Pawson and Tilley (1997) label these Context-Mechanism-Outcome (CMO) configurations. The approach applies mixed methods and different data sources, and does not accept the “evidence-hierarchy” in meta-analysis, where RCT and quasi-experimental studies are regarded as the only methods that can provide “strong evidence” (Pawson 2006; Reiper & Hansen 2007).

Randomisation is important, but not the only way to make causal inferences (Davidson 2000). In situations where random assignment is not practically possible or ethically defensible, then program theory evaluation may assist by investigating whether intermediate outcomes have been achieved and whether there are alternative explanations for outcomes and pattern matching (Rogers et al. 2000).

Funnell and Rogers (2011: chapter 15) advise being scientific, but also pragmatic. They describe a number of evaluation methods and techniques that are applicable to infer causation when interventions have complicated and complex aspects.

First, we should ask whether the final outcomes of the program are congruent with the program theory. We need to make plausible causal links between outcomes and the program or intervention. This can sometimes be done by logical reasoning (cf. Dahler-Larsen 2001). We should also investigate whether the program was implemented as intended, and if so, whether the theory of the program worked as intended. At a more specific level, we should investigate what lies behind the average outcomes, if the program worked for some groups in some sites rather than others. If the sample becomes too small to make statistical generalisations, we could ask participants whether they understand how the program worked and whether it contributed to change their behaviour (Bredgaard et al. 2011b).

Second, we can make counterfactual comparisons by asking what would have happened in the absence of the program. When control groups or comparison groups are used, program theories can improve these methods by identifying mediators and moderators that determine the size and direction of causal mechanisms.

Third, we should make critical reviews of evaluation findings by asking whether there are other possible explanations for the final results than the program. A useful method is to ask participants and key informants or to make comparisons across cases.

Applications of the evaluation framework

As an illustration of the points made above we draw on examples from two current research projects designed to evaluate employment policies towards unemployed on social assistance and sickness benefits. The two research projects apply realistic evaluation and the aim is to determine what works for whom and under which circumstances in specific active employment programmes.

As mentioned in the introduction we are often directed towards impact evaluations when we evaluate active employment policies. Working within the framework of realistic evaluation, however, the research projects incorporates a mixed-method approach that allows for an opening of the “black-box” of

interventions and thus deals not only with “what works” but also for whom it work, in which context, and why (Pawson & Tilley 1997).

As discussed previously, the nature of the intervention undergoing evaluation is vital to the design. Some aspects of an intervention towards unemployed can be simple while other aspects can be complicated or even complex. The two week job search assistance program mentioned above (cf. figure 1) is part of the research project about employment policies towards unemployed on social assistance. The assumptions is that the program will make the unemployed individuals able to find more vacant positions (which increases the likelihood that they find a vacant position they like), improve their applications (thus making it more likely that their application is chosen) and give them a realistic view on job or study opportunities (i.e. lowering their expectation and showing them alternatives).

Each aspect of the job search assistance program can be characterized as mainly simple, e.g. the teaching in job searching where the unemployed is informed about different job databases and is given lectures on how to write a CV. The teaching in job searching can be standardized with a clear set of objectives or learning outcomes and it can be carried out almost the same way everywhere and typically by the same organization or types of organizations. Furthermore experiences with offering the program gives solid knowledge of how it should be implemented and what to expect of it (Patton 2011, p. 86). In other words the intervention can be controlled and planed in advance, as can the design of the evaluation. To be successful in increasing transition from unemployment to employment, however, the job search training need to be adapted to the specific local labour market context.

The level of complexity further increases when the aim is to evaluate whether the entire job search assistance program leads to the desired outcome - employment. First, it is likely that the objectives and outcomes of the program will be valued differently by different people in different environments and thus creating variation in the implementation of the program. Second, it is questionable whether a two week program alone will lead to employment. Finally, the unemployed will typically be offered several interventions simultaneously or continuously thus making it difficult to separate the effect of one intervention from another. Hence an impact evaluation after the two week program has finished cannot stand alone. Thus a program theory should be applied to construct and test assumptions about why it works for whom and under which circumstances, e.g. by interviewing staff and unemployed and analyzing documents describing the program and the possible progression of the unemployed.

Often active employment policies are more complicated than this example. Another example is a program of intensive support targeted at people on sickness benefits suffering from stress, depression or anxiety. This 10 week program consists of meetings and consultations every day with either a therapist and/or a social worker. The objective is to enable participants to return to work through daily support and a range of activities such as physical exercise, therapy, stress- and anxiety management, etc. Thus, the target group has complex barriers to reemployment and the intervention attempts to address this complexity.

The program assumption is that an early, coordinated and integrated intervention will ensure a quicker transition from sickness benefits to regular employment. The early intervention is expected to help prevent long-term sick leaves and thus prevent a worsening of the condition due to isolation, exclusion, loss of skills etc. The coordination of relevant professions (i.e. psychologists, physician, social workers etc.) is expected to help establish a more holistic treatment program and prevent misunderstandings, inapt overlaps, etc. The nature of different types of therapy, conversations with psychologist and/or social workers, and the combination with physical training and labour market integration increase program complexity. Objectives and outcomes vary to a great degree and interventions are likely to work differently in different situations and for different individuals in different environments (Funnell and Rogers 2011: 75).

It is therefore appropriate to apply impact evaluation and program theory in order to fully grasp all outcomes and variations of the intervention and thereby become able to asses if indeed the program works, and for whom it works and under what circumstances.

Conclusions

In modern governance, the complexity of policy problems has increased in tandem with the complexity of policy interventions. Systematic policy experimentation and evaluation are suitable methods to address this complexity.

The higher the complexity of the policy experiment, the less likely that experimental techniques are an appropriate evaluation method. We, therefore, suggest combining the strengths of experimental evaluation with program theories to better understand the mechanisms that explain why experiments work for some groups in some sites, some of the time. We have demonstrated this combined evaluation design and methodology with reference to one case of interventionist human services, namely active employment programs.

The advantages of combining experimental evaluation with program theory evaluation are that they: (1) Allow us to open "the black box" and investigate how and why programs succeed or fail, (2) address for whom programs are particularly successful or unsuccessful, thereby allowing for better targeting of interventions, (3) provide information about what aspects of programs lead to success or failure, (4) systematically distinguish between how programs should work (construction of program theory) and how programs actually work (test of program theory), (5) address whether success or failure can be attributed to the program or some other explanation, (6) allow the inherent (complicated and sometimes complex) nature of active labour market interventions and situations to be reflected in the evaluation and, (7) allow for a constructive dialogue and exchange between program staff, participants, and evaluators.

There are also disadvantages in combining impact evaluation with program theory evaluation: (1) Evaluations become more ambitious and time-consuming, and by implication, more expensive. Therefore, tough choices and delimitations of the scope and ambitions of the evaluation are necessary. (2) Mixed methods and cross-disciplinary evaluation competencies are needed, but often in short supply. Public organisation need to build evaluation capacity and engage with interdisciplinary evaluation teams. (3) Reviewing, synthesising, and simplifying the findings of such nuanced and complex evaluations is much more difficult, but still necessary if they are to inform policymakers and other decision makers. Realist reviews seem to be a promising method (cf. Pawson et. al 2005; Pawson 2006).

References

Astbury, B. & F.L. Leeuw (2010): Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation, *American Journal of Evaluation*, 31(3), pp. 363-381.

Berkel, R. van & V. Borghi (2007): New Modes of Governance in Activation Policies, *International Journal of Sociology and Social Policy*, Vol. 27, no. 7/8.

Bloom, H.S. (2005), ed. *Learning more from social experiments – Evolving analytic approaches,* New York: Russell Sage Foundation.

Bredgaard & Larsen (2007): Implementing Public Employment Policy: What Happens when Non-Public Agencies take over?, *International Journal of Sociology and Social Policy*, Vol. 27, no. 7.

Bredgaard, T., H. Jørgensen, P. K. Madsen & S. Rasmussen (2011a): *Dansk arbejdsmarkedspolitik*, Jurist- og Økonomforbundets forlag.

Bredgaard, T., H. H. Jørgensen, R. Madsen, M. R. Dahl & C. Hansen (2011b.): *Hvad virker i aktiveringsindsatsen*, Report commissioned from Employment Region Northern Jutland (Rapport udarbejdet for Beskæftigelsesregion Nordjylland).

Boone, J. & J. C. van Ours (2004): Effective Active Labour Market Policies, IZA Discussion Paper no. 1335.

Bruttel, O. (2005): Contracting-out and Governance Mechanisms in the Public Employment Service; Berlin: Wissenschaftszentrum, *Discussion paper* 2005-109.

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

Card, D., J. Kluve & A. Weber (2010): Active Labour Market Policy Evaluations: A Meta-Analysis, NBER Working Paper Series, 16173 (www.nber.org/papers/w16173).

Chen, H.-T. (1990): *Theory-Driven Evaluations*, California: SAGE Publications.

Considine, M. (2001): *Enterprising States – The Public Management of Welfare-to-Work*, Cambridge: Cambridge University Press.

Considine, M, O' Sullivan, S., E. Sol (forthcoming): *The End of entitlement – Activation and Contracted Welfare in Australia, Britian and the Netherlands* (working title).

Cook, T. D. (2000): The False Choice Between Theory-Based Evaluation and Experimentation, *New directions for Evaluation*, no. 87, pp. 27-34.

Dahler-Larsen, P. (2001): From Programme Theory to Constructivism: On Tragic, Magic and Competing Programmes, *Evaluation*, 7(3), pp. 331-349.

Danish Economic Council (2012): *Dansk Økonomi Efterår 2012,* Copenhagen: Det Økonomiske Råd.

Davidson, J. E. (2000): Ascertaining Causality in theory-Based Evaluation, *New directions for Evaluation*, no. 87, pp. 17-26.

European Commission (1993): *White Paper on Growth, Competitiveness and Employment*, Brussels: European Commission.

European Commission (2006): *Employment in Europe 2006*, Directorate-General for Employment, social Affairs and Equal Opportunities.

Funnell, S. C. & P. J. Rogers (2011): *Purposeful Program Theory – Effective Use of Theories of Change and Logic Models*, San Francisco: Jossey-Bass.

Graversen, B. K., B. Damgaard & A. Rosdahl (2007): *Hurtigt I Gang. Evaluering af et forsøg med en tidlig og intensive beskæftigelsesindsats for forsikrede ledige.* SFI, Socialforskningsinstituttet.

Graversen, B. K. (2012): Effekter af virksomhedsrettet aktivering for udsatte ledige, en litteraturoversigt, SFI – Det Nationale forskningscenter for Velfærd, Rapport 12:20.

Hansen, M. B. & E. Vedung (2010): Theory-Based Stakeholder Evaluation, *American Journal of Evaluation*, 31(3), pp. 295-313.

Hasluck, C. & A. E. Green (2007): *What works for whom? A review of evidence and meta-analysis for the Department for Work and Pensions*, Department for Work and Pensions, research Report no. 407.

Jensen, P. & M. Rosholm (2011): Arbejdsmarkedet – hvad virker og hvad virker ikke? *Samfundsøkonomen*, no. 1, pp. 31-35.

Kluve, J. (2010): The Effectiveness of European Active Labour Market Programs, *Labour Economics*, 17, pp. 904-918.

Koning, J. & H. Mosley (2001): *Labour Market Policy and Unemployment – Impacts and Process Evaluations in Selected European Countries*, Cheltenham: Edward Elgar.

Le Grand, J. & W. Bartlet (1993): *Quasi-markets and social policy,* Houndsmil: MacMillan
Press.

Martin, J. P. & D. Grubb (2001): What works and for whom: A review of OECD countries' experiences with active labour market policies, *Swedish Economic Policy Review*, 8, pp. 9-56.

OECD (1994): *The OECDs Jobs Study: Facts, Analysis, Strategies*, Paris: OECD.

OECD (2012): *Employment outlook 2012*, Paris: OECD.

Osbourne, D. & T. Gaebler (1992): *Reinventing Government – How the entrepreneurial spirit is transforming the public sector,* Reading, Mass: Addison-Mesley.

Patton, M. Q. (2011): *Developmental Evaluation – Applying Complexity Concepts to Enhance Innovation and Use*, The Guilford Press.

Pawson, R. & N. Tilley (1997): *Realistic Evaluation*, London: Sage Publications.

Pawson, R., T. Greenhalgh, G. Harvey & K. Walshe (2005): Realist Review – a new method of systematic review designed for complex policy interventions, *Journal of Health Services Research and Policy*, 10(1), pp. 21-34.

Pawson, R. (2006): *Evidence-based Policy – A Realist Perspective*, London: Sage Publications.

Peck, L.R. (2012): *What Works for Addressing the What Works Question in Field Experiments*, ABT Thought leadership paper, July 2012.

Reiper, O. & H. F. Hansen (2007): *Metodedebatten om evidens*, København: AKF-rapport.

Rogers, P, A. Petrosino, T. A. Huebner & T. A. Hacsi (2000): Program Theory Evaluation: Practice, Promise, and Problems, *New Directions for Evaluation*, no. 87, fall 2000, pp. 5-13.

Rossi, P. H., M. W. Lipsey & H. E. Freeman (2004): *Ealuation – A Systematic Approach*, Califonia: SAGE Publications (seventh edition).

Rothstein, B. (1998): *Just Institutions Matter – The moral and Political Logic of the universal Welfare State,* Cambridge University Press.

Pons, G. & J.N. Arendt (2010): The effect of a wage subsidy on employment in the subsidized firm, Copenhagen: AKF-rapport.

Rambøll (2008): Hurtigt i gang 2. *Kvalitativ evaluering baseret på spørgeskemaundersøgelse blandt deltagere.* Arbejdsmarkedsstyrelsen.

Rambøll (2009): *Kvalitativ evaluering af ´alle i gang´* Arbejdsmarkedsstyrelsen. København.

Rosholm, M. & M. Svarer (2009a): *Kvantitativ evaluering af Alle i gang*. Oktober 2009.

Rosholm, M. & M. Svarer (2009b): *Kvantitativ evaluering af hurtigt i gang 2*. september 2009.

Rosholm, Michael, Svarer, Michael (2010): *Effekten af samtaler i den aktive arbejdsmarkedspolitik*.

Rosholm, M. & M. Svarer (2011): Effekter af virksomhedsrettet aktivering i den aktive arbejdsmarkedspolitik, dated 23.05.2011 (download from www.ams.dk).

Schmid, G., J. O'Reilly & K. Schömann (1997), eds. *International Handbook of Labour Market Policy and Evaluation*, Cheltenham: Edward Elgar.

Skipper, L. (2010): En mikroøkonometrisk evaluering af den aktive beskæftigelsesindsats, AKF-rapport.

Sol, E. & M. Westerweld (2005): *Contractualism in Employment Services – A New Form of Welfare State Governance*, The Hague: Kluwer Law International.

Struyven, L. & G. Steurs (2005): Design and redesign of a quasi-market for the reintegration of jobseekers: empirical evidence from Australia and the Netherlands, pp. 211-229, *Journal of European social Policy*, Vol. 15(3).

Weiss, C. H. (2000): Which Links in Which Theories Shall We Evaluate?, *New directions for Evaluation*, no. 87, fall 2000, pp. 35-45.

Notes

[1] An alternative to micro-econometric impact evaluations is macro-economic or aggregate impact evaluation (cf. Koning & Mosley 2001; European Commission 2006). Macro-economic evaluations estimates the correlation between labour market performance in selected countries (e.g. employment and unemployment rate) and different explanatory variables or institutions (e.g. duration and generosity of unemployment benefits systems, expenditures on active labour market policies) (for a review see European Commission 2006: chapter 3). Sometimes such macro-economic evaluations come to surprising and opposite conclusions of the micro-econometric evaluations. Boune and Ours (2004) for instance used data from 20 OECD countries covering the time period 1985 to 1999 and found that labour market training was the most effective program to bring down unemployment, while subsidized jobs were not effective at all. Macroeconomic evaluations are, however, conducted at a too aggregate level of generalisation to be particularly helpful in addressing what works for whom, under what circumstances.

[2] Meta-analysis has its origins in health care analysis (The Cochrane Collaboration), where it is usually used to generate evidence on the effectiveness of certain clinical interventions by aggregating data from a set of clinical trials of the same drug, all of which were ideally subject to the same laboratory conditions. Meta-analysis is also used in several fields of social policy (The Campbell Collaboration), including labour market policy.

[3] In the evaluation literature, there are different concepts that are used more or less with the same meaning as in program theory (like theory of change, logic models, intervention theory) and in the evaluation of program theories (like realistic evaluation, theory-driven evaluation) (see Chen 1990, Rossi et al. 2004).