European
Commission

Testing Social Policy Innovation

Primer for the training

Last updated on 25rd June 2014

Employment,
Social Affairs
and Inclusion

Table of Contents

Preface

Social policy innovation means developing new ideas, services and models to help addressing the challenges to our welfare systems for delivering better social and employment outcomes. It can help nurture the current fragile economic recovery with improved social and employment outcomes in the medium and long term. It involves new ways of organising systems and therefore invites input from public and private actors, including civil society. Closer partnerships between this broad spectrum of actors are critical to help reach the Europe 2020 targets.

As a tool to provide better and innovative solutions to social challenges, social policy innovation is an essential element for Member States' structural reforms in line with the social investment approach, as described in the Social Investment Package (SIP) . The SIP stresses the need to embed social policy innovation in policy-making and to connect it to social priorities. It also stresses the need for modernisation of welfare states given the implications of the demographic change and of the financial and economic crisis. Modernisation of social policies requires systematic introduction of ex-ante result orientation in financing decisions and a systematic approach of the role social policies play in the different stages in life.

Finally, the SIP places a special focus on improving the measurement of social outcomes in particular in terms of social returns. This relates to the need to ensure that policy reforms are not only evidence-based, but also results-oriented.

In this context the role of policy makers is crucial in guiding the reform process, selecting the appropriate policy priorities and for an effective follow-up and increased sustainability of the results. In order to play this function, policy makers need tools that allow them to assess the investment returns of the chosen policies in terms of social outcomes (increase in inclusion and employment, reduction in cost of service at same quality level, contribution to the economy...).

There are different evaluation methods available to policy makers depending on the specific features of the policy to be assessed. These methods can provide evidence of positive outcomes of policies and support policy decisions. The European Commission will organise a series of training sessions on the different methodologies. This guide is part of the training material and will be often revised and updated in view of the experience with using it. It complements a similar initiative on counterfactual impact evaluations in the context of ESF- funded projects.

This guide has been prepared under coordination of and edited by LSE Enterprise.

The information contained in this publication does not necessarily reflect the position or opinion of the European Commission. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein. Reproduction is authorised provided the source is acknowledged.

Introduction

Impact evaluations produce empirical evidence of a policy's effect. This evidence foster social policy implementation by highlighting the link between policy actions and the intended outcomes. It also allows effective policies to be transferred, contributes to their continued improvement and appropriate follow-up. Policy makers, service providers and researchers are jointly committed to evaluating the impact for future policymaking and make results available to the broader policy community.

Different methods can be used to measure the impact of a policy and gain evidence on effective social policy reforms. This guide will focus on commonly used methods, including: (1) Randomised Controlled Trials; (2) Differences-in-Differences; (3) Statistical matching; and (4) Regression Discontinuity Design.

What is the purpose of this Guide?

This Guide is meant as a companion to policy-makers and social service providers wishing to implement social policy innovations and evaluate the impact of their interventions. It addresses three important and related questions:

- How to evaluate the impact of a social policy intervention. Which methods are applicable and under which assumptions they work?
- How to design an impact evaluation? The most important decisions related to an impact evaluation will be made at the planning phase. A rushed and superficial plan is likely to result in interesting questions remaining unanswered or inadequately answered due to poor or missing data. This Guide sheds light on the critical decisions that need to be made at an early stage and the trade-offs that they involve. Concrete examples are provided.
- How to assess and disseminate its results, on the basis of their reliability, transferability, and sustainability. How to use the result to create knowledge that can feed into fine-tuning ongoing reforms, inspiring new change and building-upon to create additional knowledge. This is about participating in a community that builds and shares experience; this is about policy-makers sharing reliable experience across borders.

The examples in the second part of this guide, will illustrate the role of methodologies in supporting and facilitating the implementation of planned reforms.

Who should read this Guide?

This Guide is intended to support at national, regional and local level:

– Policy-makers; i.e. those formulating policies, be it through programmes, legislation or social dialogue. Among them, this Guide aims to support those seeking to build evidence and/or use evidence about 'what works' for social policy innovation.

– Social service providers; i.e. the organisations delivering social services, either in public agencies, charities or private sector companies. This publication is particularly relevant to those seeking to evaluate the social outcomes of their programme or policy in a credible manner.

PART I - TESTING IN 7 STEPS

Step 1 Defining policies and interventions

Social policies

Social policies consist of several interconnected interventions addressing social problems. Individual provisions in a policy cannot be considered separately, and their impact will depend on the interaction among the different provisions and on the interaction with other policies, for instance fiscal, environmental, financial policies. While individual interventions can be assessed separately more reliably, entire programmes can also be evaluated. Individual and global analysis supplement each other and allow for a better understanding, looking at "the forest and the trees" at the same time.

Social policy innovation

The concept of policy innovation has been promoted by the European Commission in the context of the implementation of the Social Investment Package (SIP). It refers to social investment approaches that provide social and economic returns and it is linked to the process of reforming social protection systems and social service delivery through innovative systemic reforms.

Social policy interventions

An intervention is an action taken to solve a problem. In the area of medical research an intervention is a treatment administered with the aim of improving a health disorder. The relative simplicity of the medical treatment makes it easily replicable; this partly explains why the whole idea of experimentation hatched and the method is so compelling in the medical context. Social policy interventions, on the other hand, have different and arguably more far-fetched aims. As mentioned in the SIP, welfare systems fulfil three functions: social investment, social protection and stabilisation of the economy. To assess social policy interventions, we might need to combine several methods.

Selecting a relevant intervention, programme or policy

While the EU2020 strategy and the SIP offer clear guidance on policy priorities, it is very important to carefully identify only the few most relevant policy interventions to be evaluated. For instance, there is limited added value in evaluating the impact of interventions on a very limited number of people, or in testing a policy question that is already supported by an extensive solid evidence-base. Important social protection system changes, innovative interventions pilots or demonstration projects, and all interventions whose conclusions can be of high relevance to the broader policy community are all excellent candidates.

While assessing the potential impact, it is important to keep in mind which features of the interventions can be reliably tested; this guide is meant to help with this. Ultimately, the decision to test a programme will rest on two legs: policy relevance and feasibility.

Evaluating entire programmes and policies

A policy can be evaluated at different levels, from the 'macro' to the 'micro' level. The appropriate level depends on the policymaker's needs. There is a practical trade-off between obtaining robust evidence on the impact of a single intervention and the concrete policy relevance that a less strict and therefore less robust methodology might provide for a broader reform. One could first evaluate the impact of a programme *as a whole*. The aim of the evaluation is then to find out whether the programme, including all its components, made a global difference for its beneficiaries.

Evaluating an entire programme create a 'black box' problem for policymakers and researchers, who are unable to distinguish the whole from its parts. Even if the programme is found to have no significant effect overall, this result is open to speculation as whether one intervention had an effect, but not another; no single intervention had an effect or the effect of the interventions was cancelled out by their interaction.

Evaluating an entire programme also has implications on the available methodological tools. For instance the use of counterfactual analysis, by its own nature, is not possible. However, the methodological limitations might be counterbalanced by more timely policy indications.

Evaluating interventions

At a lower level, it can be interesting to test each intervention separately. By comparing the impact of each intervention, policymakers can identify the most effective alternative to address a given policy goal. It is important to note that the cost of the evaluation must be carefully taken into consideration. However, an evaluation testing different hypothesis may yield more comprehensive results. Besides, it may be much more efficient, in terms of time and resources spent, to test different relevant hypotheses in one single evaluation.

Example: Intensive job-counselling

In 2007, a team of researchers tested the impact of an intensive job-counselling programme when provided by the French public employment services, and when provided by private providers. They first measured the impact of supplementing standard public employment counselling with more intensive counselling, and then compared the relative effectiveness of intensive counselling services when provided by the public employment agency and by the private providers[1].

Interventions must be precisely and carefully defined; extensions or future applications can yield much different results than during the assessment.

What's next?

Get more information	National Audit Office (2011). *Auditing Behaviour Change*. P.11-12 Available at: http://www.nao.org.uk/report/auditing-behaviour-change/ OECD. *Labour market programmes: coverage and classification*. Available at: http://www.oecd.org/els/emp/42116566.pdf Morris S, et al. (2004). *Designing a Demonstration Project.* Cabinet Office. Chapter 1. Available at: http://www.civilservice.gov.uk/wp-content/uploads/2011/09/designing_demonstration_project_tcm6-5780.pdf Glennerster R, Takavarasha K (2013). Running Randomized Evaluations: A Practical Guide, Princeton University Press, p. 8-12 and p.73-77

[1] *See Part 3, Example 2*

Step 2 Specifying a 'theory of change'

A '**theory of change**[2]' (ToC) is to social policy reform what plans and foundations are to a building structure. The section below provides a fairly succinct description of this approach. Further information on the most important steps in a ToC is given in subsequent sections.

A map to the desired outcome

The emphasis on design comes from an observation that has been made countless times by researchers, trainers and advisers: a lot of important questions in an evaluation remain unanswered or poorly answered due to superficial design. Whilst there is no such thing as a perfect design, some steps can be taken so that the energy used in the development and conduct of an impact evaluation get rewarded as it should. These steps have been integrated into a single framework known as a 'theory of change'.

A ToC has been defined as "the description of a sequence of events that is expected to lead to a particular desired **outcome**"[3]. It is the causal chain that connects resources to activities, activities to outputs, outputs to outcomes and outcomes to impacts.

A good ToC uses six different building blocks:

1. *Needs:* is the assessment of the problems faced by the target population.
2. *Inputs*: are the resources that will be consumed in the implementation of the intervention. Those include the time spent by the agents implementing and evaluating the project and the costs involved *(i.e.* the services and goods service providers will need to purchase). The critical question is: to what extent will these resources enable the delivery of the intervention?
3. *Outputs:* is what will be delivered. It can be information, a subsidy or a service. The key question here is: how likely is the intervention to produce the intended short-term outcome?
4. *Outcomes:* are the results of interest likely to be achieved once the service has been delivered. Outcomes in the social policy area usually appear in the medium-term.
5. *Impact:* is the change in outcomes that is caused by the intervention being tested.
6. Finally, a ToC should document the **assumptions** used to justify the causal chain. These assumptions need to be supported by research and stakeholder consultations. This will strengthen the case to be made about the plausibility of the theory and the likelihood that stated outcomes will be accomplished.

Table 1 – An example of ToC: Pathways to Work

The table below shows the implicit theory underlying reforms of incapacity benefits such as the British Pathways to Work programme. It shows that for programme beneficiaries to *get a job* (impact), they must *apply for a job* in the first place (outcome). Likewise, for beneficiaries to apply for a job, they must been *encouraged* or *compelled* to do so (output). This causal chain holds to the extent that the assumptions made by programme managers are credible. Here, the connection

[2] *Note that a ToC is sometimes referred to as 'programme theory', 'outcome model', 'intervention logic' or 'logical framework'*

[3] *Rick Davies, April 2012: Blog post on the criteria for assessing the evaluability of a theory of change http://mandenews.blogspot.co.uk/2012/04/criteria-for-assessing-evaluablity-of.html*

between the expected outcome 'beneficiaries apply for jobs' and the expected impact 'beneficiaries get a job' is conditional on the competitiveness of beneficiaries on the labour market.

ToC	Programme description	Assumptions
Goal	The sustainability of the Disability Insurance scheme is ensured	
Impact	Beneficiaries get jobs	Beneficiaries are competitive on the labour market
Outcomes	Beneficiaries apply for jobs	– Beneficiaries are convinced by case managers; – Financial incentives are high enough.
Outputs	– Mandatory work focused interviews; – Financial incentives to return to work; – Voluntary schemes to improve work readiness.	– Beneficiaries comply with their obligations; – Beneficiaries participate in voluntary schemes.

Inputs	– Guidelines for work-focused interviews;
	– Training for case managers;
	– Financial resources;
	– Software.
		Budget, staffing and equipment are available.

Needs	Large increase in the number of recipients of the various Disability Insurance, possibly leading to a 'fiscal crisis'. This could be due to: – A deterioration of labour market opportunities; – A policy framework combining generous disability benefits with lenient screening and monitoring.	

An essential tool in social policy management

There are several advantages in using a ToC to underpin social policy. Firstly, a ToC will help policy makers make better decisions throughout the entire lifecycle of the policy. At an early stage, it will support the formulation of a clear and testable hypothesis about how change will occur. This will not only improve accountability, but also make results more credible because they were predicted to occur in a certain way. During the implementation, it can be used as a framework to check milestones and stay on course, as well as a blueprint for evaluation with measurable indicators of success. Once the policy is terminated, it can be updated and used to document lessons learned about what really happened.

Secondly, a ToC is a powerful communication tool to capture the complexity of an initiative and defend a case to funders, policymakers and boards. The tough economic context, as well as the intense pressure on governments and organisations to demonstrate effectiveness, means that leaders are increasingly selective when it comes to supporting research projects. A visual representation of the change expected in the community and how it can come about ought to reassure them as to the credibility of the initiative. It can also keep the process of implementation and evaluation transparent, so everyone knows what is happening and why.

Doing it right

A ToC is the outcome of two parallel and simultaneous processes involving research and participation. The *research* process aims to generate the evidence base underpinning the programme and to inform its assumptions. Expectations that a new intervention will lead to the desired outcome are often justified by our 'experience' or 'common sense'. Inasmuch as possible, impact evaluations should refrain from relying on such subjective measures in that they are highly debatable and do not offer any warranty that the intervention will succeed. To be truly 'evidence based', the causal link between the intervention and the outcome should rely on social science research. An effective intervention will require insights from economics, sociology, psychology, political science, etc. Thus, it is crucial to involve experts very early on in the project. The *participatory* process usually includes a series of stakeholder workshops. The objective is (a) to get feedback on the conclusions and implications of the preliminary research; and (b) to secure stakeholder buy-in, which is an essential success factor (see FS6 on implementation).

What's next?

Get more information	Anderson, A. (2005). *The community builder's approach to theory of change: A practical guide to theory and development*. New York: The Aspen Institute Roundtable on Community Change. Website of the Center for Theory of Change: http://www.theoryofchange.org/

Step 3 Defining outcomes, outcome indicators and data collection plans

Impact evaluations test hypotheses regarding the expected outcome of an intervention. But what are well-defined outcomes? What type of metric should be used? And when should the outcome be measured? The following section gives some guidance to make the best decisions.

Prioritising intended effects …

An intervention can have two types of effects: intended and unintended. The design of an evaluation is essentially concerned with identifying and evaluating the former. Given the complexity of social mechanisms as well as the limited traction of most social policy interventions, policy-makers are advised to identify the one outcome they are most keen to change and focus all energies and resources towards it.

Now they might have a good reason to think that the intervention will have other positive effects. In this case, it is good practice to clearly prioritise those outcomes and indicate which is the primary outcome and which is the secondary outcome.

Example: the Job Retention and Rehabilitation Pilot
Between 2004 and 2006, the UK Department for Work and Pensions conducted the Job Retention and Rehabilitation Pilot, a set of workplace and health interventions meant to help people with long-term health issues stay in work. Given the nature of the intervention, it was decided that the primary outcome of the pilot was the employment situation of participants. Their health situation, which might have been affected by the intervention was considered a secondary outcome.

… while looking out for unintended effects

Always keeping the initial purpose of the intervention in mind does not mean that any unexpected pattern and signal emanating from the programme should be ignored. Some of them might be policy-relevant. Most evaluations – if not all – generate serendipitous findings which can challenge and extend our understanding of economic and social mechanisms. Such phenomena are 'natural' in social research. They should be recorded, reported and discussed. They might justify additional research.

Choosing the right metric

Once priorities are established, it is important to identify the metric which will give the most accurate estimate of the intervention's impact. The challenge here is to ensure that what gets measured accurately reflects what was meant to be measured. In other words, that the evaluation has **construct validity**.

This is sometimes fairly straightforward. For example employment programmes all have the same goal: increase the number of people in work. Thus, an evaluation will aim to compare the number of people who got a job in both the intervention and control group. Research and consultations will still be needed to define what qualifies as work – for example a minimum number of hours a week and a minimum number of weeks in work will need to be established as part of this definition – however measuring 'objective' outcomes is mostly unproblematic.

Some other outcomes are trickier to measure because they refer to higher-order notions which seem difficult to capture with just one indicator. Those include cognitive and quality-of-life indicators, etc. Those outcomes are best measured through composite scales or proxy indicators.

Example: the Medicare Alzheimer's Disease Demonstration
In their evaluation of the Medicare Alzheimer's Disease Demonstration (a long-term care provision programme), Yordi and colleagues (1997) estimated the impact of the intervention on the quality of life of patients with the *Lawton Instrumental Activities of Daily Living* (IADL) scale. The scale assesses independent living skills using eight instruments, including food preparation, housekeeping and laundering. The IADL is still commonly used in the US to evaluate the impact of healthcare.

Primary outcome measures should, where possible, be objective measures of facts or behaviours. 'Softer' measures (e.g. attitudes, self-reported behavioural change, website hits following an advertising campaign) should be used for **triangulation** or in cases where it is not possible to resort to observed measures. When applicable, it is recommended to list the different indicators used in the literature and to discuss them in an expert panel.

Stating expectations

Inasmuch as possible, one should try and use the same outcome indicators as in previous evaluations of similar interventions. This includes evaluations conducted domestically and abroad. Using the same indicator will not only make **systematic reviews** and **meta-evaluations** easier, it will also help estimate the **effect size** of the new intervention. Interventions can have positive or negative effects and this effect can be large or small. For reasons that are outlined below, impact evaluations should always aim to evaluate interventions with a positive and reasonably large expected effect – although what makes a 'reasonably large' effect is highly policy- and context-dependent. In addition, the evaluation should seek a measure that is likely to be sensitive to the results.

For instance, if a change in disability insurance provisions is expected to primarily affect the return to work of those with an intermediate level of disability, it may pay to measure the after-intervention employment rate of this specific group. Comparing employment across all disabled people may yield a less clear (and less reliably estimated) result, even if, considering all disabled people, more people would be (potentially) affected and motivated to join the labour market.

This is important for political and analytical reasons. Firstly, setting reasonable aims to the intervention and documenting this aim will help garner political support for the programme. Policy-makers will make a stronger case for a reform if they can show that the proposed intervention can potentially reduce the unemployment rate of the target group by 2% than if they are not able to give any estimate. Likewise, they will make a stronger case if they can demonstrate that option A can reduce unemployment by between 0 and 4 % and option B by between 2 and 6%.

Secondly, larger impacts are easier to estimate. For instance, in the case of RCTs, the larger the expected the size of the intervention, the fewer participants will be needed to warrant that the observed impact is not simply due to chance.

If the intervention builds on an existing reform, then a review of previous evaluations will give an indication of the likely impact magnitude. Evaluators should undertake a systematic review of any studies on the existing reform. If none are available, they might want to conduct one as part of the design phase.

Minimum Detectable Effect

The outcome of an intervention must be substantial. This means that it must be

1. sufficiently large to justify the cost of the intervention; this is a question for policy-makers, who have to weigh the intervention cost against its benefits. This means that it must be above a "*minimum economic effect*".
2. sufficiently large to be detected in a "reasonable" sample size, or number of individual people involved in the assessment; here reasonable means that it cannot be larger than the available population, and cannot be as large as to make the assessment too expensive. This means that it must be above a "*minimum detectable effect*".

Analysing the impact of the intervention in detail

Demonstrating that a new intervention was successful (or not) in resolving a policy issue across the target group is valuable in itself. However, such a result might be perceived as a 'thick' analysis of the intervention's impact. In addition to helping frame the 'big picture', evaluation reports usually provide a wealth of details about the conditions and circumstances under which the intervention is most effective. This 'thin' analysis can be useful for policymakers.

Planning the data collection : when to measure the impact

An impact evaluation is meant to test the following hypothesis: the difference between the respective effects of the two (or more) interventions is too large to be attributed to chance. The effect of each intervention is the difference between the chosen outcome *after* the implementation of the new intervention and *before* the implementation, holding all other variables constant (also called the ***ceteri paribus*** assumption; see section 5).

The measurement done before the implementation of the intervention is called the **baseline**. The **end-line measurement** should take place once the intervention is thought to have produced its *definitive* impact. The exact timing depends on a number of factors including the type of intervention: whilst some, like information campaigns, have a fairly immediate effect, others might take years to have an effect (*e.g.* education programmes). Also, complex interventions often require 'pilot phases' to let the programme 'bed in' and allow frontline agents to familiarise themselves with their new tasks. According to the *Magenta Book*[4], although there is no set duration for an impact evaluation, they usually take "at least two to three years"[5]. Regardless of when the end-line measurement takes place, it is important to comply with the research protocol.

Baseline and end-line measurements are the two ends of an impact evaluation. In addition, it might a good idea to set up a survey halfway through the implementation of the programme for monitoring and testing purposes. This measurement can be seen as a 'dress rehearsal' for the end-line measurement. Given the importance of these surveys, it is essential that everyone gets their role right. If anything goes wrong – such as an excessively high number of people dropping out of the programme – it is better to find out at the interim measurement than at the end-line measurement.

[4] *The Magenta Book is the recommended UK government guidance on evaluation that sets out best practice for departments to follow. It is available at: https://www.gov.uk/government/publications/the-magenta-book*

[5] *http://www.civilservice.gov.uk/wp-content/uploads/2011/09/chap_6_magenta_tcm6-8609.pdf*

It is recommended that interim evaluations be made publicly available, as all other research outputs, once the evaluation has come to an end. From a political viewpoint, it can be tempting to interrupt the evaluation after the interim measurement. However such pressures should be strongly resisted; as most social policy interventions do not have the same effect after six or twelve months. Ultimately, it is the long-term effect of the intervention that matters for policy-makers, so short-term outcomes should always be considered with caution.

Choosing the most appropriate data collection method

The most inexpensive way of collecting outcome data is to use those generated by service providers and public authorities as part of their administrative procedures. For example, many vocational training providers keep a detailed record, for each of their clients, of the number of training sessions attended, the exit date (for those who interrupted their programme) as well as the reason for this interruption (e.g. the client might have found a job, or might have been removed from the programme on health grounds). However, experience shows that such data is often too basic to answer all the questions raised by the programme. Conversely, large-scale national surveys tend to collect very detailed data (e.g. on income, number of hours worked, etc.) but only for a subset of the population. Thus it might very well be that the information is only available for a fraction of the participants. In any case, a detailed audit of the available data, its completeness and reliability is an essential criterion to the design of an adequate impact evaluation.

The most common method of collecting data for impact evaluations is to survey participants. Surveys are a very 'flexible' instrument: they can be used to get information about participants, to measure their views and attitudes regarding the intervention and to collect outcome data (e.g. employment status, number of hours worked, income, etc.). However, in some cases, a concern may be that the information provided by participants in surveys is unreliable because they may have forgotten what happened in the past or they may, intentionally or unintentionally, misreport sensitive information such as their income. Given the importance of surveys, it is important to design all questionnaires with great care and with the help of the evaluators.

What's next?

Get more information	Morris S, et al. (2004). *Designing a Demonstration Project.* Cabinet Office. Chapters 4 and 6. Available at: http://www.civilservice.gov.uk/wp-content/uploads/2011/09/designing_demonstration_project_tcm6-5780.pdf Glennerster R, Takavarasha K (2013). Running Randomized Evaluations: A Practical Guide, Princeton University Press, p. 8-12 and p.73-77

Step 4 Estimating the counterfactual

Impact evaluations seek to estimate the intrinsic value of public policies. There are many reasons why a programme might be perceived as a success even though it had no actual impact or vice-versa. For example it could be that the implementation of the programme coincided with favourable economic conditions, in which case the situation would have improved even without the new programme. Or, in two-group comparison, it could be that those who benefited from the new intervention were somewhat different from those in the control group, artificially boosting or impeding the intervention.

To take into account effects that have nothing to do with the intervention, impact evaluations measure its observed outcome against an estimate of what would have happened in its absence. This estimate is known as the counterfactual[6].

There are different ways of estimating the counterfactual. This section briefly outlines the difference between individual-level and population-level counterfactuals. It then presents the most frequently used techniques. Each of these methods relies on one or several assumptions that might be more or less credible depending on the context of the evaluation and of the data available. It is important that both the evaluator and the policymaker are aware of these assumptions, and interpret the results with the necessary caveats.

Individual-level vs. population-level counterfactuals

Whereas some reforms aim to modify the size and composition of the inflow into a scheme (population-level outcomes), other seek to influence the behaviour of participants to the scheme (individual-level outcomes). It is very difficult to measure population-level outcomes and individual-level outcomes simultaneously.

To estimate the impact of an intervention on individual-level outcomes (e.g., labour-market participation, net income, benefit duration), one needs to build individual-level counterfactuals. Individual-level counterfactuals amount to comparing beneficiaries of a new intervention with beneficiaries of existing provisions.

Individual-level counterfactuals can also be used to estimate the impact of different aspects of the reform and in this way, to elicit the most cost-effective ones. This is highly valuable from a policy perspective, as public expenditure can be optimised by focusing on the policy options that do have an impact on the desired outcomes. Besides, individual-level counterfactuals help estimate the heterogeneity of impact on different sub-populations in order to envision targeting. This can be highly relevant to some policies, such as those aiming to activate minimum-income recipients. Indeed, in addition to constituting a very heterogeneous group, they generally face greater employment difficulties (for example in comparison to Unemployment Insurance beneficiaries).

In order to estimate the impact of the reform on the inflow into the scheme, one needs to build population-level counterfactuals. In this case, two similar groups are compared: one group is given access to the new scheme, whereas the other continues benefiting from the existing scheme. The difference in application rates, enrolment and characteristics individuals entering the two schemes provides a measure of the change in inflow size and composition. However, this type of analysis can be difficult

[6] *See ESF guide on counterfactual evaluation: Design and Commissioning Of Counterfactual Impact Evaluations. A Practical Guidance for ESF Managing Authorities, European Commission, 2013.*

to undertake, especially when service provision is fragmented between different organisations or departments.

Method 1: Randomized controlled trials (RCTs)

The credibility of an impact evaluation depends on the degree of similarity between the control and the intervention group, both in terms of observable and unobservable characteristics. Random assignment to intervention and control groups constitutes the most reliable method; if the sample is sufficiently large, it guarantees that the control group has the same characteristics as the group receiving the programme[7].

Often, the reasons determining participation are correlated with outcomes. For example, if there is a job counselling programme open to everyone, it is very likely that the most motivated job seekers enrol. Motivation in turn can be correlated with the probability of finding a job. Comparing participants to non-participants will give an overestimated measure of the impact of the intervention, as participants can be more motivated (in average) and therefore, more likely to find a job, even in the absence of the programme. The opposite is also possible: if the job counselling programme is only available to unqualified job seekers that have been unemployed for a considerable length of time, comparing participants with non-participants will provide an underestimated impact, as participants would have fared relatively worse than non-participants, even in the absence of the programme.

These examples show that selection matters. Randomly assigning entities from a sufficiently large sample to the control and to the intervention group helps us control for selection, as it ensures that the groups are statistically identical on observable and unobservable characteristics. The only difference is that one group gets the intervention, and the other one does not (at least temporarily). As a result, it is easier to establish causal relationships between the intervention and the observed difference in outcomes of participants and non-participants.

RCTs are seen as a rigorous method for constructing a valid counterfactual. However, randomized evaluations of social programmes require time and resources, and must be designed before the intervention is implemented.

RCT are not always applicable. Social interventions are often constrained by laws and administrative procedures. For instance, a country's constitution may forbid applying to a subset of the population an intervention that included reducing or increasing benefits; or, some interventions may apply to whole communities, e.g., those that foster parallel mutual-help economies. Generally, however, a reform plan will include provisions that can be tested via RCT and others that may need to use other methods.

Example: Integrated healthcare in the US

A US study published in 2002 recruited chronically disabled older people receiving in-home services, who were at risk of using a high amount of acute services. Half of the patients were assigned at random to a clinical nurse care manager (NCMs), who was tasked to improve the linkage between the acute and long-term care services used by programme enrolees. The aim of the intervention was primarily to reduce hospital utilizations.

[7] We distinguish between prospective (experimental) methods and retrospective (quasi-experimental) methods. In the latter case there is no random assignment but manipulation of the independent variable in order to create a comparison group using matching or reflexive comparisons. When it is not practical or ethical to randomly assign participants (e.g. male or female, specific categories…) quasi-experiments are designed to nevertheless maximize internal validity even though it will tend to be lower than with RCTs. The rest of the section provides an overview of the main methods available.

Although there was some variation in health use and cost across treatment and control groups over the 18 month time period, the authors concluded that there were no differences between groups on any of the outcome variables examined. Efforts to integrate the acute and long-term care systems proved more difficult than anticipated. The intervention, which attempted to create integration through high intensity care managers, but without financial or regulatory incentives, was simply not strong enough to produce significant change for the clients served. The programme was also affected by various organisational changes, such as changes in the management of the hospitals involved in the studies, with repercussions on the way they communicated with NCMs (Applebaum et al., 2002).

When randomisation is used to assign participants to the intervention and control groups, there is a high probability that the two groups are identical. This assumption can be tested empirically through a balance test. Researchers can measure the distributions of baseline characteristics for each group in the evaluation to verify that there are no significant differences on key variables that might influence outcomes. Lack of balance can occur even if the process of random assignment was properly undertaken, but this risk is minimized as the sample size increases.

The same caveats that apply to statistical matching (see below) apply to this test: there is never a guarantee that the balance test includes all the relevant variables and the two groups may be skewed along an unobserved variable. However, not finding significant differences along observed variables is reassuring.

Typical use of Randomised Controlled Trials (RCTs)

Random assignment is a viable research design when the following conditions are met:

1. Ethics of research on human subjects are well defined and do not prevent applying different interventions to different people. For instance, it would be unethical to deny an intervention whose benefits have already been documented to some clients for the sake of an experiment if there are no resource constraints.
2. The sample size is large enough. If there are too few subjects participating in the pilot, even if the programme were successful, there may not be enough observations to statistically detect an impact.

Method 2: Regression Discontinuity Design

This method can be implemented when there is a clear, quantitative eligibility criterion or threshold (a cut-off score), that separate a group of people under intervention from another (control) group. It compares people just above the cut-off score (who qualify for the new policy or programme), with those just beneath it (who do not).

As an example, assume that under a new disability benefit provision, those with a level below a certain threshold are reduced benefit and offered active labour market measures. The method would compare disabled people just above (still under old benefits) and just below (new provision applies) the threshold.

This method relies on the assumption that the intervention strictly implements a clearly quantifiable selection criterion, and that participants are unable to anticipate and manipulate the scoring close to the cut-off point (in the example above, ‘tweak their disability level’). Also, it assumes the individuals just below and just above the threshold are not significantly different. This usually entails that the score around the threshold is on a continuum. Significant differences may arise, if for instance, the same cut-off point is used to deliver different services, which might mean that the two groups (above and below the threshold) face qualitatively different conditions even without considering the intervention.

The main weakness of this method is that it measures the effect of the intervention only on the people lying close to the eligibility threshold. If policymakers are interested in evaluating the impact of a policy or programme on the entire population, this method is not appropriate; its results, for instance, could not be used to tell the impact of raising or lowering the threshold. Another practical problem is deciding the 'bandwidth' that will be used to determine the sample. On the one hand, if the bandwidth is narrow, the two groups (just above and just below) will be similar, but the effect will be measured on few people, with greater uncertainty. On the other hand, if it is wide and include many people, to yield a more precise estimate of the effect, it will end up comparing more different groups.

Example: Reform of the Disability Insurance in Norway

Kostol and Mogstad (2014) used this method to assess the impact of a change in work incentives in DI in Norway. Individuals who had been awarded DI before January 1st 2004 were exposed to more generous rules when earning benefits and wages jointly (*new work incentives*) than individuals awarded DI after that date. The authors hypothesised that individuals entering just before and just after the date were very similar, so that differences in outcomes (e.g. work while in DI, exit rates) could be attributed to the change in work incentives. This is not a before and after comparison because all the individuals are observed simultaneously, in the same economic environment. The authors found that financial incentives induced a substantial part of DI beneficiaries to return to work, but only for younger beneficiaries. This supports the claim that some DI beneficiaries can work and that incentives are effective in encouraging them to do so.

Typical use of Regression Discontinuity

A study qualifies as a RDD study if assignment to conditions can be based on a 'forcing variable'. Units with scores at or above/below a cut-off value are assigned to the intervention group while units with scores on the other side of the cut-off are assigned to the control group. Such forcing variable must fulfil two criteria:

1. It must be continuous or ordinal with a sufficient number of unique values. The forcing variable should never be based on non-ordinal categories (like gender).
2. There must be no factor confounded with the forcing variable. The cut-off value for the forcing variable must not be used to assign individuals to interventions other than the one being tested. For example, eligibility to free school meals (FSM) cannot be the basis of an RDD, because FSM is used as the eligibility criteria for a wide variety of services. This criterion is necessary to ensure that the study can isolate the causal effects of the tested intervention from the effects of other interventions.
3. The value of the forcing variable cannot be manipulated by individuals: for instance, one may misreport incomes to become eligible, or decide to remain below some threshold to ensure eligibility to a valuable programme.

Method 3: Differences-in-Differences (DiD)

This method compares the change in outcomes over time among participants and non-participants. More specifically, it measures the change in outcomes for the control group to get an idea of what would the 'natural change' had been in the absence of the programme, and tracks the change in outcome for the intervention group to obtain a measure of the 'natural change' plus the change caused by the programme. By subtracting the difference in outcomes from the control group, to those from the intervention group, the evaluator can obtain a measure of the change caused by the programme.

One of the advantages of this method is that it provides a measure of the impact for the whole population of participants, while controlling for changing environmental conditions. However, it relies on the 'parallel trends assumption'. That is, in order to determine that the difference in outcomes is due to the programme, the trends in outcomes of the participants and of the non-participants should be approximately the same in the absence of the programme. One way to validate the credibility of the 'parallel trends assumption' is to verify if both groups witnessed parallel changes before the introduction of the intervention. This exercise requires many periods of data prior to the intervention, both for the intervention and for the control group.

It is also important to verify that there are no local changes – other than the programme – that might affect the trends while the intervention is being implemented. For instance, the implementation of another programme in the intervention or control region, or a shock affecting just one of the two groups could affect the outcome, biasing the estimated impact of the programme.

Example: Pathways to Work
The British evaluation of *Pathways to Work* used a DiD approach to test the impact of the programme on the outflow from DI and employment (Adam, Bozio and Emmerson, 2010). This pilot reform included stronger financial incentives to return to work, monitoring (mandatory interviews) and activation (voluntary counselling schemes – including the Choices program referred to above). The reform was phased-in experimentally by district. The Department for Work and Pensions (DWP) officials selected pilot districts before the intervention and let evaluators choose suitable control districts based on a set of observed aggregate characteristics. The level of variation was an entire district, therefore this evaluation built up population-level counterfactuals. The people who entered DI before and after the start of the pilot were followed in both districts. The divergence in outcomes between pilot and control areas after implementation of the policy was interpreted as an impact of *Pathways*.

Typical use of Difference-in-Difference

In its simplest version, difference-in-difference can be used to estimate the impact of an intervention if data from two periods can be provided. In the first period – the pre-intervention period – no individual is exposed to the new policy. In the second period – the post-intervention period – those assigned to the intervention group have already been exposed to the policy while those assigned to the control group have not. More general versions of this method allows for partial take-up of the program in the target population.

In order to use this method to identify the impact of an intervention, one must assume that the two groups would have experienced similar trends in the outcome of interest in the absence of the intervention. Arguments in favour of this intervention can be based on information from several periods before the start of the intervention: if trends have been parallel previously, they can be more likely expected to have been potentially so later on.

Method 4: Statistical Matching

This is a collective term for statistical techniques which construct a control group by matching each of the participants with one similar non-participant, on the basis of observed characteristics. The aim is to match participants pairwise with non-participants using as many variables as possible, to ensure that the sole major difference between the two groups is the intervention. The matched non-participants provide the counterfactual.

Successful matching requires a thorough preliminary research, in order to identify the different variables that could be statistically related to the likelihood of participating in the programme and to the outcome of interest. In addition to this, a large sample is needed to create sufficient matches.

This method provides an estimate of the effect of an intervention for all the participants that where successfully matched to a non-participant, and if there is enough available data, can be applied even if the programme has already ended. However, this method relies on the strong and untestable assumption that all relevant background characteristics can be observed and accounted for. In practice, there is no way to rule out the bias caused by unobserved variables that could influence both participation in the intervention and the ultimate outcome.

Example: Workfare reform in Argentina

Jalan and Ravallion[8](2003) used propensity score matching techniques to test the impact of an Argentinean workfare programme on income. There was no baseline data, and the evaluation was designed after the programme was implemented. For this reason, the researchers opted for a statistical matching technique. Through a survey, they collected information on a set of around 200 characteristics to match each participant to the most similar non-participant. Then, they averaged the difference in income between all these matched groups and verified the robustness of their results through several matching procedures. However, they were not able to rule out any bias caused by unobservable variables.

Typical use of Statistical matching

1. Factors affecting programme participation are known. This may be a problem with innovative programmes adopting novel methods to assist clients. In this instance, pre-evaluation research would be needed to identify the factors involved.
2. Information affecting entry to a programme and the outcome of interest is available. Information might be missing if insufficient time or resources have been devoted to the collection of key pre-programme variables (e.g. work or earning histories).
3. There are no variables that are unobserved and influence both entry and outcomes of interest. This is generally a very strong assumption and one that is not testable.
4. The sample is large enough. With fewer matches available, the evaluator may be prepared to accept more distant matches, resulting in increased bias. In these circumstances, estimated effects may be sensitive to the choice of the type of matching.

Random assignment is not an option. There are circumstances in which random assignment is problematic and matching offers advantages. However, with properly conducted random assignment there is no concern of selection bias given the statistical properties of randomness

[8] *see http://info.worldbank.org/etools/docs/voddocs/172/353/ravallion_antipoverty.pdf*

What's next?

Get more information	The World Bank Handbook at https://openknowledge.worldbank.org/bitstream/handle/10986/2693/520990PUB0EPI1101Official0Use0Only1.pdf?sequence=1 The European Commission Joint Research Centre manual http://bookshop.europa.eu/en/a-note-on-the-impact-evaluation-of-public-policies-pbLBNA25519/ Glennerster R, Takavarasha K (2013). Running Randomized Evaluations: A Practical Guide, Princeton University Press, p. 8-12 and p.73-77

Step 5 Analysing and interpreting the effect of the intervention

Impact evaluation methods estimate the impact of an intervention by comparing the results of the intervention and of the control group. The net effect of an intervention generally amounts to the difference in outcomes in the intervention and in the control group. This chapter presents a general overview of some important aspects to consider when interpreting results, relevant for all of the different evaluation methods.

Choosing the time to Measure Outcomes

Certain outcomes can require some time before they fully materialize and become observable by the researchers. This is typically the case of activation measures, where jobseekers are encouraged to temporarily stop their job search to follow the training offered by the programme (*lock-in period*). It is after they complete the training that they start looking for a job, a process that may in turn require some extra time before it materializes into measurable labour market outcomes. If the outcomes (i.e. employment rates) are measured during the *lock-in* period, the impact can be underestimated. For this reason it is of vital importance to choose appropriately the time periods during which different outcomes will be measured. Similarly, and more generally, social investment policies may require an effort in the short term to reap benefits in the longer term. While in the short term the policy's outcome may be unobservable, upon maturity these become more evident.

Monitoring Compliance, Limiting Attrition and Ensuring Objectivity

Results may be misleading if some of the units assigned to the control group receive the programme, and/or some of the units in the intervention group do not. Partial compliance can potentially reduce the difference in terms of exposure to the intervention of each group. An extreme case would be if the same amount of units in the intervention group and in the control group receives the programme. In this scenario it would be impossible to estimate the impact of the intervention, as both groups will have had the same exposure to the programme. By monitoring compliance while the intervention is being implemented, researchers can take early action to ensure that compliance rates improve. In addition to this, compliance rates must be rigorously reported so they can be taken into account in the analysis of results.

Another important aspect to consider when interpreting results is attrition. Attrition occurs when researchers are unable to measure the outcomes for some of the units included in the evaluation. If the type and rate of attrition is different in the intervention and in the control group, the results can be biased. Take for instance a successful activation policy, where a group of jobseekers are offered an intensive job-counselling programme (the intervention group) and another group of jobseekers can only access the standard counselling programme (the control group). The less employable jobseekers in the intervention group are more likely to improve their employment prospects and not drop out, but the less employable jobseekers in the control group may be discouraged and leave the programme. The result is that the less employable job-seekers are overrepresented in the intervention group, so the two groups are no longer comparable. In this case, the impact of the programme will be underestimated.

Finally, it is of utmost importance that an objective and independent third party undertakes the evaluation. The evaluation design and implementation must be carefully reported, and when possible, data should be made public to enable replications.

Reporting Results

Independent and impartial evaluations, with sufficient power and adequate design and implementation can also yield nil effects. A nil impact can be as informative as a large positive or negative impact. For this reason, researchers and policymakers should acknowledge that the aim of an impact evaluation is to measure whether an intervention is effective in reaching its intended goals, and that it is possible that the intervention is proven to have no effects or negative effects. In order to enhance mutual learning amongst the relevant policy community, the evaluation team should report and disseminate the results in a transparent and comprehensive way.

What's next?

Get more information	*World Bank Evaluation Toolkit*, Module 7: Analyzing Data and Disseminating Results: http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTHEALTHNUTRITIONANDPOPULATION/EXTHSD/EXTIMPEVALTK/0,,contentMDK:23262154~pagePK:64168427~piPK:64168435~theSitePK:8811876,00.html Gertler, PJ, Martinez, S, Premand, P, Rawlings, LB, Vermeersch, CM. (2010). *Impact evaluation in practice*. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/2011/01/13871146/impact-evaluation-practice

Step 6 Disseminating findings

When the evaluation results have important policy implications, research needs to be translated into policy. In addition, giving other policy-makers the opportunity to build on results, be they negative or positive, can further enhance their impact. The following section gives some advice on how it can be done.

Understanding the policy relevance of an evaluation

Policy relevance is very much time-dependent: a topic might be 'hot' one day and 'ice-cold' the following week. Thus, it is important to keep an eye on the policy agenda. A 'window of opportunity' may arise, for example, in the course of budget discussions, when policy-makers set up their priorities and allocate resources. Ensuring that evidence is provided at the right time can help make it more likely that the evidence will be examined and acted on.

Disseminating results in an accessible format

Beyond the research findings, it is also important to communicate the policy implications of the policy evaluation/testing. The results of impact evaluations are most often presented in academic working papers or journals, and these papers tend to be written in a very technical way which can limit their potential audience among policymakers. A key responsibility is to make research more accessible by extracting the most compelling results from longer papers and reports and presenting them in non-technical language.

Disseminating the full details

While most end-users will be best-served with the policy brief, it is useful to disseminate a full documentation of the project, including the data that was collected, after it is rendered anonymous. Other researchers may be interested in this work and their attentive reading will lend reliability and width to the findings. Even when findings are dutifully peer-reviewed, allowing others to scour the results may yield insights that were missed earlier. Finally full dissemination helps systematic reviews and meta-evaluations.

Publishing results in evaluation registries

Social policy-makers and practitioners can sometimes find it difficult to know where to find evidence, and even then, results are often reported in gated academic journals. A number of organisations have made efforts to make rigorous evidence centrally available. They include:

- The EU-funded European Platform for Investing in Children (EPIC)[9];
- J-PAL's Evaluation Database[10];
- The Evaluation Database of the Coalition for Evidence-Based Policy[11];
- The Evaluation Database of the Network of Networks in Impact Evaluation (NONIE)[12].

[9] *http://europa.eu/epic/index_en.htm*

[10] *http://www.povertyactionlab.org/search/apachesolr_search?filters=type:evaluation*

[11] *http://toptierevidence.org/*

[12] *www.worldbank.org/ieg/nonie*

What’s next?

Get more information	Stachowiak, Sarah. Pathways for Change: 6 Theories about How Policy Change Happens: http://goo.gl/ym94f1 Dhaliwal Iqbal and Caitlin Tulloch, From Research to Policy: Using Evidence from Impact Evaluations to Inform Development Policy, J-PAL, Department of Economics, MIT http://www.povertyactionlab.org/publication/research-policy DFID Research Uptake Guidance: https://www.gov.uk/government/publications/research-uptake-guidance Policy Impact Toolkit: http://policyimpacttoolkit.squarespace.com/

Step 7 From local to global

How does one know whether a programme that is effective on a pilot scale has the same impact when scaled up, extended or replicated in a different location? This is a very important question, and it relates to the external validity of an evaluation. External validity, also known as 'generalizability', is the degree to which one can be confident that the results found in a specific context will apply to other contexts. The following section explains how effective interventions can be scaled up so that the new approach "makes a real impact and becomes part of the norm"[13].

The challenge of transferring results

Strong internal validity is an important prerequisite (although not sufficient) to generalize results. If one cannot be confident that the evaluation measures the true impact of the programme in a specific context, then it will be more difficult to generalize conclusions to another context. One of the advantages of randomized evaluations is that they have a strong internal validity. Random assignment guarantees that the sole difference between the intervention and control groups is the fact of receiving the intervention. Any changes in outcomes can be confidently attributed to the intervention being tested, without needing to make additional inferences on the comparability of the groups.

There are four major factors that affect the generalizability of an evaluation, including the quality of implementation, the scale of implementation, the context and the content of the programme.

1. The quality of implementation: Pilot programmes are often implemented with great care, and with well-trained staff. It may be difficult to keep the same standards at a wider scale. Researchers should implement interventions in representative locations with representative partners, and representative samples.

2. The scale of implementation: a programme that is implemented on a small scale may have different effects when scaled up (general equilibrium effects). Researchers can adapt the design of the evaluation to capture these effects by using a wide enough unit of observation (i.e. Community)[14]. Comparing the outcomes in communities that introduced the programme to the outcomes in communities without the programme can help identify and measure some of these effects.

3. The context of implementation: An intervention that proves to be effective in one context may have a different impact in another institutional and cultural context. Behavioural theory can help us define which aspects of the context are likely to be relevant to a particular programme.

4. The content of the programme: The effects of a given programme may vary if some of the components of the programme are modified.

[13] *http://ec.europa.eu/regional_policy/sources/docgener/presenta/social_innovation/social_innovation_2013.pdf*

[14] *J-PAL affiliates undertook an evaluation on intensive job counselling placement and displacement effects. For more information: http://www.povertyactionlab.org/publication/job-placement-and-displacement*

What’s next?

Get more information	Allcott Hunt, Sendhil Mullainathan (2012). *External validity and partner selection bias*, Working Paper 18373, NBER Dhaliwal Iqbal, Esther Duflo, Rachel Glennerster, Caitlin Tulloch (2012), *Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education*, Abdul Latif Jameel Poverty Action Lab (J-PAL), MIT: http://www.povertyactionlab.org/publication/cost-effectiveness Cooley Larry and Richard Kohl (2006) Scaling Up—From Vision to Large-scale Change, A Management Framework for Practitioners , http://www.msiworldwide.com/files/scalingup-framework.pdf

PART II - CASE STUDIES

The following case study will illustrate the role of the methodology in evaluating the outcomes of concrete examples of systemic reforms. The logical steps presented earlier are put into practice with hands-on examples in order to help the reader to understand better how social policy innovations can be supported through research evidence.

The case studies below aim to anchor the methodology above to concrete and realistic social policy changes. These changes are plausible and even desirable, and while their purpose and principles are sound, whether their potential outcome will be realized may lie on implementation detail. Accompanying policy reforms such as the ones illustrated below with reliable outcome measures will help make the most out of them. As such, the following examples illustrate the value in accompanying reforms with suitable evaluations.

Example 1: How to evaluate a Reform of Incapacity-for-Work Benefits?

1. Introduction

The following note discusses how to evaluate the impact of a disability insurance reform. It presents the main research designs available and how these designs have been used in the past. The note shows that these designs vary essentially in:

1. Their methodological requirements; and
2. The assumptions that must be made regarding the comparability of intervention and comparison groups.

This discussion is illustrated by examples taken from Estonia, Norway, the United Kingdom and Denmark; they planned or carried out reforms intending to activate people in DI and accompanied the reforms with assessment plans. The note presents some of the contextual constraints that conditioned the resort to a particular design in those particular instances.

The note is organised as follows: Section 2 briefly describes the features of a potential reform, Section 3 explains how to build counterfactual situations, Section 4 discusses the different possible methods that could be applied, illustrated by concrete examples and section 5 provides a case-study of the contextual constraints that may arise during the evaluation process.

2. The Reform of Disability Insurance at a Glance

Over the past few decades the number of recipients of disability insurance (*DI hereafter*) has increased, particularly in Northern Europe.[15] This progression could be the result of a deterioration of labour market opportunities, coupled with policy frameworks combining generous disability benefits with lenient screening and monitoring. In fact, DI is more generous than unemployment insurance in the long run, making this type of social benefits comparatively more attractive. [16]

The constant increase in the number of incapacity for-work pensioners can hinder the sustainability of the pension system and increase manpower shortages. Policy reforms aiming to reduce the inflow and stock of incapacity beneficiaries have so far included:

[15] *http://www.oecd.org/els/emp/45219540.pdf*

[16] *This evolution has also been documented in the United States by Autor and Duggan, 2003.*

- Reinforced screening into the scheme,
- Reduction in the level of benefits, and
- Raising exit out of the DI scheme, either through incentives (implicit marginal tax rate) or activation.

The model reform contemplates all of these policy options with a variety of possible implementing provisions.

3. Criteria to Evaluate the Reform

A current reform objective would to be two-fold: it aims to curb the current increase in the number of beneficiaries and to activate the partially disabled. The proportion of disabled people in employment and the transition from DI to jobs are two important indicators of the success of activation policies. The participation of enterprises is a key determinant of this success. In this respect, it is interesting to estimate "*creaming*", which occurs when enterprises select the most employable among the disabled and leave the less employable for government jobs and subsidised jobs. This requires measuring the number of disabled people employed in the private sector (as opposed to in the public administration) and the proportion of those on subsidised jobs.

Generally, the main **outcomes** on which to evaluate such a reform are:

- Size and composition of inflow into the scheme
- Participation, hours of work, earnings, public or private jobs (and other measures of job quality), while in the scheme
- Exit from the scheme and nature of exit (towards work, towards higher work capacity)
- DI benefits, income and individual exposure to poverty
- Health of participants
- Budgetary cost

Data collection, both from administrative sources or from surveys must be planned in advance, before the implementation of the new scheme, and must be adapted to the evaluation protocol.

One general impact evaluation question could be: *How does the scheme, including all of its different features, make a difference?* Other important questions are: *Which of the policy option is most effective? Are all of the above options needed?*

The **mechanisms** of the reform depend on several aspects, the most important of which are:

- How different will the new screening process be?
- How sensitive are potential DI recipients to the level of transfers compared to potential wages or other benefits? More precisely, how important is the incentive profile for re-employment?
- How efficient are activation policies? For how long are they offered? What is the take-up of these policies?

None of the evaluation programmes referred to in the following section addresses all of these questions or considers all the outcomes at once. Some of the evaluation protocols provide an estimate of the impact of the scheme as a bundle of different policies. Other protocols disentangle the effect of the isolated components of the scheme, and in this way, help to determine which policy options are more effective. Similarly, some protocols help estimate the impact on the inflow into the scheme, while others are better suited to measure the effect of the scheme on individual outcomes. The analysis below considers evaluations that compare the new scheme with a previously existing scheme.

4. How to Build Counterfactuals

To assess whether the reform makes a difference, by how much and with what cost and benefit, one needs to establish a counterfactual situation[17], which is an estimate of what the outcome(s) of interest would have been in the absence of the new scheme.

Let's assume that, without reform, the number of people on DI woyld be expected to increase, and an estimated increase in – say – the next ten years is available.

It is not possible to gauge the impact of the reform by simply measuring the evolution of the number of DI beneficiaries before and after the implementation of the reform, because these figures can change over time for a large number of reasons unrelated to the reform. For example, a decrease in unemployment may affect the inflow into DI, which would be wrongly attributed to the reform. Inversely, unexpected adverse economic shocks can increase the number of people applying for DI, even if the reform proves to be highly cost-effective. In short, missing or attaining a target number of people benefiting from DI tells little about the desirability of the reform: the change in the number of DI beneficiaries needs to be compared with a well-defined counterfactual situation. The challenge is therefore to identify or to build such a counterfactual.

There are two possible levels of analysis, the individual-level counterfactual, and the population-level counterfactual. Each of these levels allows for the measurement of the impact on different outcomes.

4.1. Individual-level Counterfactuals

The introduction of a new scheme creates new conditions that differ from those faced by individuals under the initial scheme. Changes can include new rates, activation measures and incentives for re-employment. This may have an impact on most of the above listed outcomes, including new labour market participation rates, DI exit rates, income level and the health of beneficiaries. *To estimate the impact of the reform for each of these outcomes, one needs to compare beneficiaries under the new scheme conditions with similar beneficiaries under the initial scheme conditions.*

Individual-level counterfactuals can also be used to estimate the impact of different aspects of the reform and in this way, to elicit the most cost-effective ones. This is highly valuable from a policy perspective, as public expenditure can be optimised by focusing on the policy options that do have an impact on the desired outcomes. Besides, individual-level counterfactuals help estimate the heterogeneity of impact on different sub-populations in order to envision targeting. For instance, addressing "creaming" from employers requires focussing on individuals who may be more (or less) employable; their re-entry into the labour market may have differed between the current provisions and those under the envisaged reform.

4.2. Population-level Counterfactuals

The policy may also affect the number and the characteristics of the people joining the new scheme (inflow), through two mechanisms simultaneously: the change in the screening process –supply side– and the change in the perceived value of the scheme to potential applicants –demand side. In order to measure the change of the inflow, the relevant counterfactual cannot be based on comparing similar individuals in the

[17]See ESF guide on counterfactual evaluation: DESIGN AND COMMISSIONING OF COUNTERFACTUAL IMPACT EVALUATIONS. A Practical Guidance for ESF Managing Authorities, European Commission, 2013.

new and old scheme: it must be computed at the level of a **population** potentially eligible under each scheme, because inflow is measured as a share from a particular population. O*ne thus needs to observe two arguably similar populations in similar contexts, the only difference being that in one population, eligible individuals have access to the new scheme, whereas in the other population, eligible individuals have access to the initial scheme.*[18]Given that the aim is to measure how the new scheme affects the inflow, the populations to compare are not beneficiaries, but a well-defined set of people that could enter the scheme and that, depending on the rules, will or will not apply (potential applicants).

The difference between population-level counterfactuals and individual-level counterfactuals is that for the first, one must find two similar populations of *potential beneficiaries*, one *facing* the initial scheme, and the other facing the new scheme, whereas for the latter, one must find similar *individual beneficiaries*, some *participating* in the initial scheme and others in the new scheme.

With population-level counterfactuals it is possible to measure the new labour market participation rates, the number of people terminating their benefit claim (DI exit rate), and income level of beneficiaries that happen to enter the new scheme, and compare these outcomes with those of the beneficiaries that happen to enter the initial scheme. However, these measures combine composition effects (*different people may react differently to a given scheme*) and scheme effects (*same people behave differently under the new rules*).In other words, if the new scheme affects the inflow of DI beneficiaries, the individuals participating in the new scheme are no longer comparable to those in the initial scheme, both because they have different characteristics and because they face different rules.

In order to make a valid comparison between individuals in the new and in the initial scheme, one would need to assess what happens with the beneficiaries under the initial scheme that remain out of the new scheme and the opposite group of those who are only eligible under the new scheme.[19] However, it is not possible to identify the individuals that would enter both schemes and individuals that enter the initial scheme but would remain out of the new scheme because many of the characteristics determining entry into DI are unobservable. Unobservable characteristics can include motivation to go back to work and sensitivity to the level of transfers.

In short, individual and population counterfactuals measure the impact on different outcomes. Individual-level counterfactuals are better suited to gauge the impact on outcomes such as new labour market participation rates, DI exit rates, income level and health of beneficiaries, whereas population-level counterfactuals allow to determine the impact of the reform on the size and composition of the inflow and stock of beneficiaries. It is particularly difficult to evaluate all of the implications of the reform considered here, because it has many elements that affect both the entry into the scheme (inflow) and conditions faced while in the scheme.

It must be noted that regardless of the level of analysis, a reform of this type must be accompanied by both an administrative follow-up that registers people in and out of the system, and samples that study smaller groups more closely, asking them questions that are not documented in the administrative data.

[18] *The difference between population-level counterfactuals and individual-level counterfactuals is that for the first, one must find two similar populations of potential beneficiaries, one facing the initial scheme, and the other facing the new scheme, whereas for the latter, one must find similar individual beneficiaries, some participating in the initial scheme and others in the new scheme.*

[19] *The full welfare analysis of such a reform would be a formidable undertaking. It would require following a very wide population (all people that have not an ex ante zero probability of joining DI). This is feasible in principle using DiD or similar experimental approaches but will not be pursued here. We only consider important sets of outcomes separately.*

5. Potential for the Different Counterfactual Impact Evaluation Methods

5.1. Matching

Description
This method constructs a comparison group by matching each of the beneficiaries in the new scheme with one similar beneficiary in the initial scheme using a set of observable characteristics. Successful matching requires a preliminary research to identify the different variables that could be statistically related to the likelihood of participating in the programme and to the outcomes of interest. Large samples are required in order to create sufficient matches. This method provides an estimate of the effect of an intervention for all the new scheme participants that can be successfully matched to an initial scheme participant.

Matching could be implemented to compare individuals in the new scheme with those in the initial scheme. Notice however, that this method is of very limited interest if the initial scheme is entirely stopped when the new scheme is introduced, and the comparison is based on retrospective data. In this scenario, participants in the new scheme will inevitably differ from individuals in the initial scheme at least in one major aspect: the economic environment they face.

Assumptions
The main assumption is that all relevant background characteristics influencing participation and the outcomes of interest can be observed and accounted for.

Example
In the context of the evaluation of Pathways to Work run by the Institute for Fiscal Studies (see below), Adam, Bozio and Emmerson (2009) discussed the implementation of this method to evaluate *Choices*, one of the components of the Pathways to Work programme. *Choices* included a variety of **voluntary** schemes designed to improve the employability and job prospects of applicants. The researchers concluded that matching participants to non-participants in *Choices* on a large set of observable characteristics was not a rigorous evaluation strategy. It was their opinion that many important unobserved differences between matched participants had remained, making it impossible to know how much of the impact was determined by the unobserved differences between individuals who happened to choose different programmes and how much by *Choices*.

Applicability to the Reform in section 2
If applied to the reform above, this method would amount to comparing individuals with similar or very close observed characteristics in the former and new versions of the scheme respectively. It would estimate the impact of the new features of DI on individuals in the new scheme. To estimate the effect on the inflow, one would need to match similarly disabled people, some facing the initial scheme, and others the new scheme, and compare their DI entry and exit probability.

In theory, matching could be implemented to evaluate the reform at a full scale, provided that data on a cohort before the implementation date and a cohort after the implementation date is available, either because administrative information is recorded or because a survey has been set up in time. In this case, the required information would include the degree and type of disability, age, gender, family status, study and work experience. The cause of the disability (e.g., work) could be important. This information would have to be collected for every disabled individual, including those who did not apply for benefits.

However, as mentioned earlier, this method requires that all the relevant characteristics determining participation and influencing the outcomes of interest can be observed and accounted for. This is in general a strong hypothesis, and in this particular case, a highly improbable one. In fact, many of the relevant characteristics that could influence final outcomes are related to employment capacity and the level of disability. Even if the assessment of capacity to work recorded in administrative data would provide a relevant proxy of work capacity, one of the central elements of the reform is a change in the screening process. This implies that the two measures of disability (assessment of work capacity under the initial and under the new screening administration) would be hard to compare. If the new reform did not include a change in the screening process, then this approach could be more relevant.

Finally, by matching on observed variables that are measured in a uniform way in the initial or in the new scheme (such as age, gender, education, etc.) it is possible to decompose the observed changes in overall outcomes of beneficiaries (i.e. employment rates, etc.) into a composition effect (due to all the variables used in this matching) and other effects *altogether*, including impact of the reform, but also characteristics that are not taken into account in this matching.[20]

An alternative use of matching could be to compare the effect of various forms of activation measures (or the lack of thereof) on individuals participating in the new scheme, instead of comparing the initial scheme to new scheme.

5.2. Regression discontinuity design (RDD)

Description
Regression Discontinuity Design compares individuals just above a given continuous eligibility threshold, with those just below. Those individuals are arguably very similar, and the threshold determines if they are exposed or not to the intervention. The bandwidth between the lower and the upper limit containing the threshold determines the sample size.

Assumptions
This method relies on the assumption that the intervention implements a clearly quantifiable selection criterion based on some continuous score, and that participants are unable to anticipate and manipulate the scoring close to the cut-off point. Also, it assumes the individuals just below and just above the threshold are not significantly different.

Example
Kostol and Mogstad (2014)[21] used this method to assess the impact of a change in work incentives in DI in Norway. Individuals who had been awarded DI before January 1st 2004 were exposed to more generous rules when earning benefits and wages jointly (*new work incentives*) than individuals awarded DI after that date. The authors hypothesised that individuals entering just before and just after the date were very similar, so that differences in outcomes (e.g. work while in DI, exit rates.) could be attributed to the change in work incentives. This is not a before and after comparison because all the individuals are observed simultaneously, in the same economic environment. The authors found that financial incentives induced a substantial part of DI beneficiaries to return to work, but only for younger beneficiaries. This supports the claim that some DI beneficiaries can work and that incentives are effective in encouraging them to do so.

A major aspect of this method, strongly emphasized by Kostol and Mogstad (2014) is that **beneficiaries were awarded DI before the change in rules applied to them**. This means that individuals entering DI before and after the 1st of January 2004 were not aware that a change of rules would happen, and were admitted under the same screening mechanism. The change was implemented retroactively, so individuals were unable to manipulate their entry into the programme.

[20] *This is called "Oaxaca-Blinder" decomposition in the statistical literature.*

Had the screening process and the value of DI changed at the discontinuity date along with the content of the programme, the RDD hypothesis would not have held any more: people just before and just after the date would have been very different and their comparison would have therefore been meaningless.

Applicability to the Reform under section 2

The main requirement to implement this method to evaluate the reform is that there is some continuous variable that determines entry into the initial or into the new scheme. As in Kostol and Mogstad (2014), a natural candidate is the date of implementation of the scheme if one looks at individuals entering very close to that date. RDD could be used to estimate jointly the impact of financial incentives and activation, but will not be able to disentangle the impact of each separate component. Besides, two conditions must hold: financial incentives and activation measures must be implemented separately from a change in the screening process and the reform must be introduced retroactively on a set of beneficiaries. It is not clear that this would be applicable to the reform.

If the new DI scheme is threshold-based, i.e. offers benefits depending on the levels of disability, and disability is measured and reported on a continuum, the impact of the scheme on individual people could be assessed by comparing the behaviour and labour market outcomes of those just below and just above the threshold(s). Notice however, that this would not compare the new with the initial scheme, but the impact of being in DI under the new scheme relative to not being in DI.

Some of the limits of RDD are that it cannot help estimate impacts on the inflow. Besides, it requires a substantial amount of inflow, so that there can be at least a few hundred individuals close to the date or disability-level discontinuity point.

5.3. Difference-in-difference (DiD)

Description
Differences-in-Differences (DiD) requires that an experimental dimension is introduced in the form of pilot and control areas. This method compares the change in outcomes before and after the start of the programme, in pilot and control areas. DiD provides a measure of the impact for the whole population of participants, while controlling for constant conditions (observed and unobserved) that may be correlated with both the final outcomes and whether the individual is in the control group.

Assumptions
DiD relies on the "parallel trends" assumption. That is, in order to determine that the difference in outcomes is due to the programme, the trends in outcomes of the participants and of the non-participants should have been the same in the absence of the programme. One way to validate the credibility of the parallel trends assumption is to check if both groups witnessed parallel changes before the introduction of the programme. Other alternatives to support this assumption include performing "placebo" tests on fake treatment groups (groups out of which none is affected by the start of the reform) or on fake outcomes (outcomes that should not be influenced by the reform).

Example
The British evaluation of Pathways to Work used a DiD approach to test the impact of the programme on the outflow from DI and employment (Adam, Bozio and Emmerson, 2010). This experimental reform included stronger financial incentives to return to work, monitoring (mandatory interviews) and activation (voluntary counselling schemes – including the *Choices* program referred to above). The Department for Work and Pensions (DWP) officials selected the pilot districts before the intervention and let evaluators choose suitable control districts based on a set of observed aggregate characteristics. The level of variation was an entire area, therefore this

evaluation built up *population-level counterfactuals*. The individuals that entered DI before and after the start of the pilot were followed in both districts. The divergence in outcomes between pilot and control areas after implementation of the policy was interpreted as an impact of Pathways.

From a managerial viewpoint, it is conceivable that pilot areas are selected precisely because they are ‘atypical’. In the present case, pilot areas were chosen where Jobcentre Plus had been in action for some time. All other things being equal, selecting a high-performing area as pilot site increases the chance of observing a positive outcome. Conversely, a pilot could be seen as a way of challenging or reforming a low-performing area. Both decisions compromise comparability. In addition to this, selecting the pilot areas either because they have high or low performance reduces the external validity of the evaluation. If the policy is rolled-out to cover new areas with different performance records, the impact of the programme at scale will probably be different.

The “parallel trends” hypothesis becomes hard to justify if pilot and control areas are different (for instance because social services are performing better in pilot). This hypothesis is untestable, but the authors provide some supporting evidence. They selected two groups of pilot districts that entered the experiment in two separate waves. They found that before the introduction of the programme, control and pilot areas had similar exit rates out of incapacity benefits. The exit rates in pilot and control areas diverged only when pilot areas entered the scheme, a pattern that was observed separately during the first and the second wave. This suggests that that outflow from DI was changing precisely at the time Pathways was introduced, providing support to the chosen methodology and to an impact of the “pathways”.

The evaluators found that the reform accelerated exits from DI, but only for those who would exit within a year anyway. However, they observed lasting effects on employment. They interpreted these two results as implying that the effect was driven by married women that would have exited the scheme anyway and rely on their partner’s resources, but instead returned to employment because of Pathways. This was (weakly) supported by subgroup data analysis.

One of the advantages of DiD is that it controls for time-invariant differences, both in observable and unobservable characteristics. If the inflow into the new scheme remains unchanged, then the differences of the change in the outcomes of interest in the pilot and control areas can be interpreted as the impact of the new scheme. However, if the programme influences the individual’s decision to enter the scheme or modifies the screening rules and changes the size and the composition of the population before and after the introduction of new scheme, DiD can measure the change in inflow size and composition, but cannot determine the impact on other outcomes; these will be influenced also by the different composition itself.

For example, if the new incentives and activation rules of Pathways were well known and influenced the decision to enter the new scheme, the cohorts entering DI in the pilot and control areas would be of a different composition. To make things salient, imagine that only men enter DI under the initial scheme and that only women enter DI when Pathways is implemented. The different behaviour of men and women would mix with Pathways’ impact. In this case, it would be impossible to separate the observed differences in behaviour by gender from those caused by Pathways. In practice, the blurring is subtler than this extreme example, but also more complex and insidious because pilot and control populations can also differ in terms of unobservable characteristics, such as personal motivation to find work, family support and sensitivity to benefit changes. For this reason, if the evaluation aims to test the impact of the reform on similar people, researchers must verify that the composition of the inflow is not affected by the introduction of the programme.

In the Pathways example, neither the size nor the observable composition of the inflow rate was affected by the policy. In this case, the differences in the change of outcomes between pilot and control areas can be interpreted as the impact of the new rules on the population that was initially in the scheme (and remains).

Finally, Pathways to Work did not include new screening rules. However, had the policy changed the screening rules, the populations before and after the introduction of the scheme would have been systematically different. As in the previous example, such a simple DiD would only measure the effect on the size and composition of the inflow, but would be unable to measure the impact on other outcomes. However, in as far as the eligibility for DI under the reform could be assessed for people in DI under the older scheme, a comparison could be made between similar groups, eligible for DI under both old and new rules.

Applicability to the Reform under section 2

To summarise, the DiD approach generates population-level counterfactuals (the population of the pilot and control areas). If applied to the reform, the evaluators should verify whether the policy has a potential impact on inflow size or composition and, if possible, make the necessary adjustments.

On the one hand, if the policy does not change the inflow into the scheme, then this design can evaluate the impact of the components of the scheme on the behaviour of similar beneficiaries. This is always under the hypothesis that trends in outcomes would be parallel in pilot and control areas in the absence of the programme.

On the other hand, if the policy affects the inflow, DiD's measure will be influenced by differences in inflow size and composition. As with matching, it is possible to decompose the different outcomes of beneficiaries under the two schemes (e.g. employment rates) into the effects of changing the composition of measured characteristics (age, gender, education) and a residual effect that encompasses scheme impact and all remaining unobserved characteristics. Compared to the matching case, the DiD setup allows to neutralise the effect of different economic environments. The DiD hypothesis can also be weakened when supplemented with matching.

In practice, in this case it may pay to separate the population into groups with homogenous disability levels and carry out the analysis separately by group (as determined in the reform, from "fit to work" to "no capacity") and comparing the flows or stock of people in DI within each group. This would require assigning each current disabled person to one of the new three groups, defined under the reform.

5.4. Randomised Controlled Trials (RCTs)

Description

The two examples above have evaluated the impact of the reforms implemented in Norway and in the United Kingdom without disentangling the individual effect of their components (the DiD case further estimated that there was no effect on the inflow). In order to determine which components of the reform are more effective (i.e. incentives vs. activation and different variations of each), as well as to identify which individuals benefit the most, a set of individuals exposed to each of the components must be compared with a set of similar individuals that are not exposed.

Planning randomised experiments rather than relying on given features of the scheme provides the option of choosing what questions to answer. For example, an experimental design where individuals are randomly assigned into different groups, and then each group is offered different elements of the programme can help disentangle the impact of different policies comprised in the reform. This can be done in two ways, either comparing people in different places (by randomising districts that offer different variants) or different people in the same place (by randomising individuals within districts, for instance based on randomised birth dates).

This comparison can be made by having pilot areas, each implementing a variant of the programme (a generalization of the Pathways protocol). For instance, some areas could implement incentives only, and others activation measures only.

It must be noted that when uptake of the variants of the programme is voluntary, there may be self-selection of participants into each of the different variants. In principle, matching could be implemented to improve comparability. However, as previously discussed, in this case matching is not a very reliable method because volunteers probably differ in many unobserved ways.

Assumptions
We do not need to rely on strong assumptions as with other protocols because random assignment from sufficiently large samples ensures that individuals are similar on average, both in terms of observable and unobservable characteristics. Still, one must assume that people do not behave differently because they are aware of being in an experiment. This assumption applies to all types of experiments, including randomised controlled trials and natural experiments, such as phased-in programs.[22]

It is also important that the scheme being evaluated is well defined and mature, and therefore similar to what would be implemented if the scheme is generalised onto a broader scale. Again, this is not specific to randomised controlled trials, as it applies to any sort of evaluation that aims to estimate the impact of an intervention that is gradually phased-in.

Example
The National Labour Market Authority in Denmark launched a randomised controlled trial in early 2009 to test on a small scale some of the provisions in a planned reform of DI. Rehwald, Rosholm and Rouland (2013) randomly assigned sick-listed workers in Danish job-centres into a treatment and a control group. The treatment group was offered a series of activation services (graded return to work, preventive health care action) but faced otherwise similar conditions as the control group. They found that the activation services had no impact overall, in spite of its cost.[23]

Applicability to the Reform under section 2
A randomised experiment must be planned in advance, before the reform is implemented. The two main elements that could be tested are activation and financial incentives.

Activation can be tested following the above example (Rehwald, Rosholm and Rouland, 2013). There have been a large number of randomised controlled trials on activation policies in different European countries including Denmark, France and Germany.

It is important to note that the impact on labour market outcomes, which are the most important here, must be observed over an extended period, implying that the separate provisions (treatment and control) must remain separate for extended periods. These outcomes can be measured once the increased motivation to work is translated into an actual job, a process that might take some time. This requires that

[22] Evaluation-driven effects occur when the subjects change their behaviour because they know they are part of a study, and not because of the intervention being studied. In addition, scientists carrying out the experiments and its measures could be biased if they know who is part of the study: for that reason, data collection must strictly follow the same protocol in the treatment and control groups (as they must with any other method anyway).

[23] This evaluation is the only one of the three discussed here that measures health outcomes, finding no impact of the interventions tested.

the experiment remains in place and that generalisation is withheld until outcomes are observable.[24]

Financial incentives to work can also be tested in a similar manner. There have been several randomised experiments in Canada and in the United States on various populations (*see the chapter on Guaranteed Minimum Income Reform in this Guide).*

The reform's implementing provisions could be tested. Subsidies to employment could be modified, for example by increasing the amount of subsidy provided and by shortening its duration, or by focusing on the job that the disabled person occupied prior to her/his disability. Subsidies could also promote the adaptation of the work-place (i.e., part of the subsidies could be earmarked for adapting the workplace). One could randomly assign people who are offered the different features, or areas where those elements would be available. Such experiments could move on while the basic scheme is already implemented.

5.5. Pilot and RCTs

If subjecting different individual people to the current provision and the reform at the same time and in the same places is not feasible for legal or administrative reasons, the reform might be phased in by geographical areas, with all residents in an area under the same rules and incentives. Under this scheme, the treatment (reform) and control (current provisions) group would consist of areas. This would amount to running exactly the same evaluation as in the DiD *Pathways to Work*, but the pilot areas where the reform is implemented first would be chosen at random (the remaining serving as controls under the current provisions), rather than by the administration. Random assignment increases the comparability of treatment and control areas and dispense with the need of relying on the "parallel trends assumption". Such area level randomisation was implemented, for instance, in France to evaluate job-search assistance (Crépon et al., 2013) and can be employed once it is decided to phase-in the programme.

Random assignment can sometimes face political constraints. For instance, governments may want to select high performing areas as pilots to promote a reform, or inversely, focus on low performing areas to challenge it. In Pathways, the government selected the pilot areas where the Jobcentre Plus were running well and for a sufficient amount of time. In practice, the definition of the areas may have to follow administrative boundaries.

That being said, random assignment of pilot areas provides more reliable results than a deliberate selection when a reform is being phased in. When assignment to pilot and control areas is not random, the pilot areas may not be directly comparable to control areas by construction. In this case, the DiD "parallel-trends" hypothesis can be invoked, but this remains an untestable hypothesis. Besides, when assignment is not random and typical areas are chosen as pilots (the very best, or the very worst), the external validity of the evaluation is limited. There can be a balance between defining a large set of mature enough areas and testing the programme on a randomised subset of them, and choosing the very best ones; however, any results here may be applicable solely to "mature" areas.

6. Institutional, organisational and political requirements

Institutional, organizational and political requirements can also influence methodological decisions, as evidenced by the evaluation of Pathways.

[24] *This is in fact required by any evaluation protocol that measures labour market outcomes, if one wants to wait until the evaluation is available before making a policy decision.*

The Pathways to Work programme was to be implemented by Jobcentre Plus, which was a new type of public employment service, resulting from the merger of the Employment Service and the Department for Social Security. When the first Pathways Pilots went live in October 2003, only a third of employment offices operated the Jobcentre Plus model of delivery, which significantly constrained the selection of pilot areas, and thus the design of the evaluation. The fact that, in October 2003, the Department for Work and Pensions (DWP) was already running six welfare-to-work pilots across the UK (putting frontline agents under strain), further complicated the selection of pilot sites.

Cost and capacity issues

The evaluation of Pathways considered the impact of the programme as a whole. In other words, it did not shed light on whether any particular component of the package (e.g. work-focused interviews, return-to-work credit) was primarily responsible for the overall impact. Investigators regretted that the evaluation was not designed to give a more complete picture of the effectiveness of the policy (Adam, Emmerson, Frayne and Goodman 2006: 4)[25].

However, DWP officials emphasized that designing an evaluation which allowed for the impact of the different components of Pathways to be estimated individually would have required a more complex, larger and more expensive pilot and evaluation, or run a substantial risk of delivering inconclusive results (Boa, Johnson, King, 2010: 22)[26].

References

Autor David H. and Mark G. Duggan, *The Rise in the Disability Rolls and the Decline inUnemployment,* The Quarterly Journal of Economics, Vol. 118(1): 157-205, 2003.

Adam Stuart, Antoine Bozio and Carl Emmerson, *Can we estimate the impact of the Choices package in Pathways to Work?*, Working paper no. 60 Department for Work and Pensions, 2009.

Adam Stuart, Antoine Bozio and Carl Emmerson, *Reforming Disability Insurance in the UK: Evaluation of the Pathways to Work Programme,* Institute for Fiscal Studies, 2010.

Bruno Crépon, Esther Duflo, Marc Gurgand, Roland Rathelot and Philippe Zamora, *Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment*, The Quarterly Journal of Economics, Oxford University Press, vol. 128(2), pages 531-580, 2013.

Kostol Andreas Ravndal and Magne Mogstad, How *Financial Incentives Induce Disability Insurance Recipients to Return to Work*, American Economic Review 104(2): 624–655, 2014.

Rehwald Kai, Michael Rosholm and Bénédicte Rouland*, Does Activating Sick-Listed Workers Work? Evidence from a Randomized Experiment, (work in progress, 2013).*

[25] *http://cep.lse.ac.uk/seminarpapers/06-06-06-EMM.pdf*
[26] *https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/207544/wp82.pdf*

Example 2: How to evaluate a Reform on a Guaranteed Minimum Income?

1. Introduction

This case study discusses how to evaluate the impact of minimum income schemes. It presents the main research designs available and how these designs have been used in the past. The case study shows that these designs vary essentially in:

1. Their methodological requirements; and
2. The assumptions that must be made regarding the comparability of the intervention and control groups.

This discussion is illustrated by examples taken from the United States, Canada, Cyprus, France and the United Kingdom.

The case study is organised as follows: Section 2 provides an overview of minimum income benefit schemes; Section 3 briefly describes the features of a likely reform, for reference; Section 4 explains how to build counterfactual situations, and discusses the different possible methods that could be applied to accompany the reform with an assessment framework. This discussion is illustrated by concrete examples.

2. Guaranteed Minimum Income (GMI) at a Glance

Almost all European countries have established minimum-income schemes aiming primarily at alleviating poverty. These cash or in-kind transfers intend to provide an adequate standard of living for families unable to earn enough income. They most often function as last-resort safety nets along with unemployment benefits, but in some countries they can constitute the main vehicle for delivering social protection.

Minimum income schemes typically include assistance benefits that do not depend on employment status or on past contributions, means-tested lone-parent benefits, housing benefits and tax credits.

An important policy challenge is to design a scheme that ensures a minimum income for those unable to afford an acceptable standard of living while safeguarding the incentives to work. If the amount of benefits provided are higher than the expected earnings from work, minimum income benefits can act as a disincentive to work, both at the extensive margin (whether to work or to rely on benefits) and at the intensive margin (the number of hours worked). If on the contrary, the amount of benefits is too low, or the activation measures are not properly targeted, eligible families that are willing but unable to work would not be protected against poverty and destitution. Time limits and conditions sometimes address this balance. Active labour market policies (ALMPs) and earnings supplements are two policy alternatives that have been implemented to encourage welfare recipients to work while preserving an adequate safety net. Whether these policies create the appropriate incentives to return to work while guaranteeing an adequate level of living is an empirical question.

3. A Guaranteed Minimum Income (GMI) Reform

Minimum Income Schemes can vary in:

- The generosity of benefits and their shape;
- The allowances that are available (i.e. childcare support, health-care services, housing support, educational allowances);
- Eligibility criteria;
- Behavioural conditionality (i.e. need to comply with activation measures);
- Whether the unit is the individual or the family.

A GMI reform aims to:

- Avoid duplication by replacing the existing public assistance benefits by a centralised GMI scheme operated;
- Enhance work incentives by implementing tightened requirements to comply with active labour market programmes;
- Create additional fiscal space by reinforcing the screening into the supplementary need-driven benefits (family, disability, health benefits and educational grants).

The GMI can include a basic allowance, a housing allowance, a tax allowance, as well as a one-time allowance provided in case of extraordinary needs. The unit of eligibility is usually the family, and the main criteria of eligibility are that the family's basic needs exceed the family's income and the fulfilment of activation conditions. The scheme would cover a heterogeneous group of beneficiaries, including families that have exhausted their unemployment benefits, families that do not qualify for unemployment benefits, and working families whose earnings cannot cover their basic needs.

Generally, the main outcomes on which to evaluate such a reform are:

- Poverty levels: change in net income and consumption;
- Labour supply both in terms of labour market participation and of hours worked (of main beneficiary and/or of spouse);
- Wage progression;
- Intra-household distribution of income (i.e. impact on spouse welfare, impact on child welfare).

The **mechanisms** of the reform depend on several aspects, the most important of which are:

- How different will the new means-testing process be?
- How sensitive are potential GMI recipients to the level of transfers compared to potential wages?
- How do the supplementary needs-driven benefits affect family incomes and employment incentives?
- How do activation policies affect the take up of the GMI scheme? How efficient are they? For how long are they offered?

4. How to Build Counterfactuals

To assess whether the reform makes a difference, by how much and with what cost and benefit, one needs to establish a counterfactual situation[27], which is an estimate of what the outcome(s) of interest would have been in the absence of the new scheme.

The GMI reform changes the eligibility criteria to access the scheme, as well as the level of entitlements and activation measures faced once in the scheme. As a result, the reform can potentially change *the size and composition of the inflow into the scheme (population-level outcomes)*, as well as *the labour-market participation rates and income levels of benefit recipients (individual-level outcomes).*

In order to gauge the impact of the new GMI on individual-level outcomes (i.e. labour-market participation, net income, benefit duration) one needs to build **individual-level counterfactuals**. In this case, individual level counterfactuals would amount to comparing beneficiaries of the new GMI scheme with similar beneficiaries of the initial social assistance programme.

Individual-level counterfactuals can also be used to estimate the impact of different aspects of the reform and in this way, to elicit the most cost-effective ones. This is

[27] *See ESF guide on counterfactual evaluation: Design and Commissioning of Counterfactual Impact Evaluations. A Practical Guidance for ESF Managing Authorities, European Commission, 2013.*

highly valuable from a policy perspective, as public expenditure can be optimised by focusing on the policy options that do have an impact on the desired outcomes. Besides, individual-level counterfactuals help estimate the heterogeneity of impact on different sub-populations in order to envision targeting. This is highly relevant for policies aiming to activate minimum-income recipients, as in addition to constituting a very heterogeneous group, they generally face greater employment difficulties (for example in comparison to Unemployment Insurance beneficiaries).

In order to estimate the impact of the reform on the inflow into the scheme, one would need to build **population-level counterfactuals.** In this case, two similar populations are compared: one population has access to the initial public assistance scheme, and the other has access to the new GMI scheme. The difference in application rates, enrolment and characteristics of families entering the initial social assistance programme and the new scheme would provide a measure of the change in inflow size and composition. However, this type of analysis might be difficult to undertake for the present reform, as the initial welfare system is fragmented into different social assistance programmes, administered by different ministries and departments. Duplication of the decentralised information can thus provide conflicting data on the participation in the different social assistance programmes.

It must be noted that when the reform affects the inflow into the new scheme, individual-level outcomes can combine composition and scheme effects. In fact, if the reform changes the eligibility criteria, the individuals that decide to apply to the new GMI scheme and that are eligible will probably differ from those that decide to apply and enrol in the initial regime, hindering comparability. More precisely, when the inflow composition is modified, any change in individual-level outcomes (labour-market participation, net income and benefit duration) could be induced simultaneously by specific characteristics of the GMI beneficiaries that qualified for benefits under the new eligibility rules, and by the new conditions they face while in the scheme. In order to disentangle how much of the impact is due to the change in individual characteristics, and how much is induced by mechanisms introduced by the new scheme, one can identify and compare similar individuals in the scheme under the two regimes. This can be done only if one is ready to admit that observed characteristics are enough to make individuals comparable (see the discussion of matching below).

As in the Disability Insurance Reform, it is very difficult to measure changes in inflow (population-level outcomes) and changes in employment status and in incomes (individual-level outcomes) simultaneously.

5. Potential for the Different Counterfactual Impact Evaluation Methods

5.1. Matching

Description
This method constructs a counterfactual by matching participants to non-participants using a set of observable characteristics. Successful matching requires a preliminary research to identify the different variables that could be statistically related to the likelihood of participating in the programme and to the outcomes of interest. Large samples are required in order to create sufficient matches. This method provides an estimate of the effect of an intervention for all the participants that can be successfully matched to a non-participant.

Assumptions
The main assumption is that all relevant background characteristics influencing participation and the outcomes of interest can be observed and accounted for.
Applicability to the Reform above

In theory, this method could be used to gauge individual-level outcomes as well as changes in the inflow into the new scheme.

In order to estimate the impact of the new scheme at the *individual-level*, one could match beneficiaries in the former scheme to individuals in the new scheme based on a set of observed characteristics, and then compare their different outcomes (employment status, income level, etc.). The characteristics should include eligibility criteria into the old and new schemes; this would allow stranding out families who would not qualify under the new scheme and who would be assessed separately. An alternative use of matching could be to compare the effect of various forms of work incentives (or the lack of thereof) on individuals participating in the new scheme, instead of comparing the initial scheme to new scheme.

In order to estimate the effect on the *inflow*, one would need to match potential GMI recipients, some facing the initial scheme, and others the new scheme, and compare their entry probability into the GMI scheme.

Notice however, that in this context, matching presents important shortcomings:

- As mentioned above, this method requires that all the relevant characteristics determining participation and influencing the outcomes of interest can be observed and accounted for. This is in general a strong hypothesis, and in this particular case, a highly improbable one. It is important to note that benefit take-up is voluntary, thus it is difficult to ascertain how much of the impact is determined by the new GMI scheme, and how much is driven by pre-existing differences in the observable and unobservable characteristics of individuals that chose to participate.
- Besides, if the new GMI scheme completely substitutes the previous social assistance programmes, and the comparison is based on retrospective data, participants in the new scheme will no longer be comparable to participants in the initial scheme, as they will be probably facing a different economic environment.

5.2. Regression discontinuity design (RDD)

Description
Regression Discontinuity Design compares individuals just above a given continuous eligibility threshold, with those just below. Those individuals are arguably very similar, and the threshold determines if they are exposed or not to the intervention being evaluated. The bandwidth between the lower and the upper limit containing the threshold determines the sample size.

Assumptions
This method relies on the assumption that the intervention implements a clearly quantifiable selection criterion based on some continuous score, and that participants are unable to anticipate and manipulate the scoring close to the cut-off point. Also, it assumes the individuals just below and just above the threshold are not significantly different.

Example
Jones (2013) evaluated the impact of the Earned Income Tax Credit (EITC) on the number of hours worked using a Regression Kink Design, a variant of the Regression Discontinuity Design (RDD)[28]. The EITC was first established in 1975 in the United States as a refundable credit for low income individuals and couples. It aims to transfer income to low income families and at the same time, to encourage and support those who choose to work. Eligibility depends on three main criteria: the taxpayer must have a positive earned income, this income must be below a specific

[28] *Card et al. (2012) introduced a variant of the Regression Discontinuity Design (RDD) which they call RKD.*

threshold, and although childless taxpayers are eligible for a small EITC, the most significant EITC is provided to taxpayers with resident children.

Advocates of the EITC argue that this credit transfers income to the neediest while incentivising work, because the credit is only accessible to taxpayers. However, it is not clear how does the credit's structure incentivises work at the intensive margin. The credit initially increases with income (phase in), but then reaches a constant region followed by a gradual phasing-out. Taxpayers in the phase-in income range face a positive substitution effect as the EITC increases with the number of hours worked. Taxpayers in the plateau region face a negative income effect because the EITC offers a steady amount of credit regardless of the number of hours worked. In this scenario, the taxpayer can reach a given level of utility working fewer hours than what would be required in the absence of the EITC. This means that as income increases, taxpayers "buy" more leisure by reducing the number of hours worked (leisure being a normal good). Finally, taxpayers in the phase-out region face a negative substitution effect and a negative income effect. The amount of credit diminishes as the number of working hours increases, making an extra hour of leisure relatively less expensive than an extra hour of work (negative substitution effect). Besides, as income increases, taxpayers can "buy" more leisure by reducing the number of hours worked (negative income effect).

Similarly to RDD method, RKD method relies on a kink (change in slope) in a policy rule to identify the causal effect of the policy. In this case, the author takes advantage of the discontinuities or "kinks" in the EITC benefit function to examine how they affect the number of hours worked by single mothers (intensive margin of labour supply). In fact the amount of benefits received is a function of earnings. This function is continuous except at two points or "kinks" (just before entering the plateau region, and when the plateau region ends and the phasing-out begins). The author compared the number of hours worked by those just before a kink to those just after the kink. The results showed that single mothers adjusted their behaviour to maximise their benefits. That is, single mothers with more than one child reduced the number of hours worked when their income in the preceding period fell just after the kink in the EITC benefit function where the benefit begins to decrease.

The first assumption is that two groups of women near the "kink" are similar on observed and unobserved characteristics. Significant differences could arise, if for instance, the same cut-off point was used to deliver other types of services, which could influence the outcome of interest. This would be the case if other tax and transfer programmes changed close to the EITC kinks. The author highlights that women with one child face different federal income tax burdens and Child Tax Credits depending on which side of the kink they fall. Similarly, women with one child or more face different state marginal tax rates depending on the side of the kink they fall in. For these groups, the RKD method would not allow to disentangle the effect of the change in other taxes and credits from that of the EITC incentives.

The second assumption is that, even if women can modify the number of hours worked once they find out where their earnings of the preceding year had placed them, they are unable to precisely assign themselves to their preferred position of the benefit function at a given tax year. The author argues that this foreknowledge is unlikely, as the "kink" points change every year. She provides some supporting evidence by showing that the concentration of earnings is not "lumped" around the kink points.

Applicability to the Reform above

The RDD method can be applied to the reform if there is some continuous variable that determines entry into the new GMI scheme. One possible candidate could be the

date of implementation of the scheme. Applicants entering minimum income benefits after a certain date will be assigned to the new GMI scheme and those before would remain under the initial social assistance programme conditions. The cut-off date must be determined retroactively so that minimum income beneficiaries cannot manipulate their entry into the new scheme. Also, the initial social assistance programme must keep running in parallel to the new GMI scheme (at least during the evaluation), as individuals in each scheme are observed during the same period of time.
Another approach would be based on an earnings threshold that determines eligibility, since the fact that individuals on both sides of an eligibility threshold can be considered very similar. However, this approach is not valid if individuals can manipulate their income to be on the eligible side. Unfortunately, this is likely to be a possibility in most institutional contexts.

With this method it is possible to measure the impact of the new scheme as a bundle, but not to disentangle the effect of the new benefit levels from that of activation measures.

Besides, RDD cannot help estimate impacts on the inflow, and it requires a substantial amount of inflow so that there can be at least a few hundred individuals close to the date discontinuity point ensuring the necessary statistical power to detect an impact.

5.3. Difference-in-difference (DiD)

Description
Differences-in-Differences (DiD) compares the change in outcomes before and after the start of the programme, over time for participants and for non-participants. DiD provides a measure of the impact for the whole population of participants, while controlling for constant conditions (observed and unobserved) that may be correlated with both the final outcomes and with the fact of being part of the control group.

DiD can either compare a group of individuals that is eligible to receive the intervention, to an arguably similar group that is not eligible, or can compare pilot areas where the programme is introduced to comparison areas that do not receive the programme.

Assumptions
DiD relies on the "parallel trends" assumption. That is, in order to determine that the difference in outcomes is due to the programme, the trends in outcomes of the participants and of the non-participants should have been the same in the absence of the programme, and the composition of each group must remain unchanged. One way to validate the credibility of the parallel trends assumption is to check if both groups witnessed parallel changes before the introduction of the programme. Other alternatives to support this assumption include performing "placebo" tests on fake intervention groups or on fake outcomes, as well as undertaking the DiD analysis using different control groups.

Example
The impact of the EITC reforms in the US have been extensively studied using DiD. Eissa and Liebman (1996) implemented this method to evaluate the impact of the EITC expansion of 1987[29]on labour force participation and hours of work on women with children. The authors focused on single women with children, as they constitute the largest group of taxpayers eligible for the EITC. At the time of the evaluation, one of the eligibility criteria was having at least one resident child. Using a difference-in-difference strategy, the authors compared the change in labour supply of single

[29] *The EITC expansion resulted from the Tax Reform Act of 1986 and consisted in a higher subsidy rate in the phase-in region, higher maximum credit and a lower phase-out region.*

women with children (intervention group, potentially eligible for the EITC) before and after the EITC expansion, to that of single women without children (control group, not eligible for the EITC). They found that the labour supply (extensive margin) of single women with children increased more than that of women without children. They found no impact on the number of hours worked (intensive margin).

Differences-in-Differences helps to control for environmental factors (new policies, economic conjuncture) that could induce a change in labour supply. The authors include different control groups to provide support for their evaluation strategy. However, they rely on two major assumptions. First, they hypothesise that single women with children would have behaved similarly to single women without children in the absence of the EITC expansion. They provide some supporting evidence, showing that the long-run trends of labour force participation do not follow very different paths, although the labour force participation of single mothers seems to be more sensitive to the business cycle.

The second major assumption is that there are no unknown shocks—other than the EITC expansion—that could have affected differently the outcomes of the intervention and the control group. This would be the case if there was a change in other tax and credit policies, business cycle fluctuations or other economic shocks that affect differently single mothers and single women without children. This is a very strong hypothesis, as it is very difficult to rule out the existence of such unknown shocks.

Blundell, Brewer and Shephard (2005) used a similar strategy to measure the impact of the Working Families' Tax Credit (WFTC), introduced in the UK in October 1999. The aim of this scheme was to support low-income working families with children. The authors compared employment rates of parents to those of non-parents, assuming that the underlying employment trends would have followed similar paths in the absence of WFTC. They found that WFTC and other contemporaneous taxes and benefits reforms resulted in higher employment rates for lone mothers, and lower employment rates for fathers in couples.

Another example is the evaluation of the Programme "Revenu de Solidarité Active" (RSA) in France. The RSA was first introduced in 2009 to replace several welfare schemes. The new scheme provided higher incentives to return to work through cash-grants conditional on employment. The amount of benefits increased after one year of work in order to encourage job stability. In addition to this, the duration of the benefits was extended and job counselling was reinforced.

The evaluation introduced an experimental dimension, by establishing pilot and control areas. The main outcomes considered were employment rates and job quality. Pilot areas where chosen by the Government, and the evaluators suggested a list of control areas, matched to the pilots on a set of socio-demographic criteria.

Here it is important to note that even if from a managerial point of view it is conceivable to select "atypical" pilot areas either because they perform better than average (to champion a reform for example) or worse than average (to challenge a reform), this non-random choice can undermine comparability, and potentially yield biased results. All other things being equal, selecting a high-performing area as pilot site increases the chance of observing a positive outcome. Conversely, low-performing areas increase the chances of observing negative outcomes. In addition to this, selecting the pilot areas either because they have high or low performance reduces the external validity of the evaluation. If the policy is rolled-out to cover new areas with different performance records, the impact of the programme at scale will probably be different.

Both the EITC and the RSA evaluation face the challenge of low statistical power. It is very difficult to identify with certainty potential beneficiaries of minimum income, so both evaluations had to measure the relevant outcomes (i.e. employment, income) on very large samples, without knowing precisely which individuals were in an employment situation that made them likely to enter the scheme or sensitive to changes of its features. The EITC sample comprised all single women, and in the RSA sample included all individuals initially benefiting from two welfare schemes, RMI[30]or API[31].

Applicability to the Reform above
DiD can either measure the impact of the GMI scheme on inflow or on other outcomes such as labour participation and income levels. On the one hand, *if the policy does not change the inflow into the scheme*, this design can evaluate the impact of the components of the scheme on the behaviour of similar beneficiaries. This is always under the hypothesis that trends in outcomes would be parallel for intervention and control groups in the absence of the programme.

On the other hand, *if the policy affects the inflow*, DiD's estimate will be influenced by differences in inflow size and composition. As with matching, it is possible to decompose the different outcomes of beneficiaries under the two schemes (e.g. employment rates) into the effects of changing the composition of measured characteristics (age, gender, education) and a residual effect that encompasses scheme impact and all remaining unobserved characteristics. Compared to the matching case, the DiD setup allows to neutralise the effect of different economic environments.

For this particular reform, DiD can be implemented mainly in two cases. One is if some *predetermined* groups are excluded from the scheme or face different versions of the scheme. These groups should be selected based on objective observed variables, such as age, family size, and previous employment status. In that case, one can follow those groups over time and assume that, in the absence of the reform, their employment or poverty situation would have evolved in parallel. Another application of DiD can result from a phase-in implementation of the policy, where some areas are chosen as pilot and others as control, as in the French RSA experiment.

5.4. Randomised Controlled Trials (RCTs)

Description
Randomised Controlled Trials measure the average impact of a policy or programme by randomly assigning entities to intervention and control groups, and then comparing the difference in outcomes.

Assumptions
We do not need to rely on strong assumptions as with other protocols because random assignment from sufficiently large samples ensures that individuals are similar on average, both in terms of observable and unobservable characteristics. One must assume that people do not behave differently because they are aware of being in an experiment. This assumption applies to all types of experiments.

It is also important that the scheme being evaluated is well defined and mature, and therefore similar to what would be implemented if the scheme was to be generalised at a broader scale. Again, this is not specific to randomised controlled trials, as it

[30] *Revenu Minimum d´Insertion*
[31] *Allocation de Parent Isolé*

applies to any sort of evaluation that aims to estimate the impact of an intervention that is gradually phased-in or that could be modified following the assessment.

Example

Since the 1970s, randomised experiments have been widely implemented to measure the elasticity of labour supply with respect to financial incentives[32].

The Self-Sufficiency Project (SSP)[33] is a large scale randomised experiment implemented in Canada from 1992 to 1999. SSP offered supplements to the earnings of single-parents who had been income assistance recipients for three or more years, on the condition that they left welfare and return to work within the year following the introduction of the scheme. Single parents were randomly selected from the Assistance Insurance (AI) records. The supplement was very generous: the combination of the supplement and earnings was nearly twice as big as the minimum-income for a full-time jobj.

It is important to note this project did not evaluate the impact of introducing a minimum income scheme, but that of changing the rules of the benefit levels offered to minimum income recipients. This is an important policy question, as the shape of benefit levels can affect welfare recipients´ behaviour by making potential work earnings more or less attractive.

One of the advantages of randomised experiments is that it is possible to measure the impact of different components of a given scheme. For example, the Recipient SSP study measured the impact of financial incentive alone, while the SSP Plus study gauged the effects of financial incentives and employment-related services.

Researchers found that while financial incentives alone had a positive impact in labour market participation and employment earnings, the combination of earnings supplement with job-counselling resulted in even larger effects.

A similar randomised experiment has been implemented in France. Researchers[34] tested whether a guaranteed minimum income extended to the youth (below 25), the *Revenu Contractuel d'Autonomie*, improved participation in a job-placement programme and helped youth secure better-paying, permanent positions. They randomised individuals enrolled in an activation programme[35] and offered them the guaranteed minimum income. Compared with control individuals, the beneficiaries decreased their labour supply, but only in the few months following the introduction of the scheme. Besides, the beneficiaries attended more regularly the activation programme, had higher disposable incomes but no different employment status after only three months.

Applicability to the Reform above

By planning randomised experiments, rather than relying on given features of the scheme researchers and policymakers do not need to tailor their questions to already existing data. Instead, they have the freedom to focus on the most relevant questions,

[32] *Meyer (1995) provides an overview of the main lessons learnt from the U.S. Unemployment Insurance Experiments: http://economics.sas.upenn.edu/~hfang/teaching/socialinsurance/readings/fudan_hsbc/Meyer95(4.13).pdf*

[33] *http://www.srdc.org/what-we-do/demonstration-projects-impact-evaluation-studies/self-sufficiency-project.aspx*

[34] *Researchers are Romain Aeberhardt, Véra Chiodi,Bruno Crépon, Mathilde Gaini and Augustin Vicard.*

[35] *Starting with individuals already in a program is a way to avoid any impact on the inflow and concentrate on the impact of the scheme on identical individuals.*

and to design the data collection strategy that is better suited to answer them. For example, an experimental design where individuals are randomly assigned into different groups, and then each group is offered different elements of the programme can help disentangle the impact of different policies comprised in the reform. To do this, it is possible to compare people in different places (by randomising districts that offer different variants) or people in the same place (by randomising individuals within districts).

In this reform, a possible candidate for such an evaluation is the active labour market programmes. A number of RCTs have been run in several countries to evaluate the impact of such interventions and this does not raise particular difficulties. A natural strategy is to randomly allocate a share of minimum-income entrants into the activation programme. The remaining minimum-income entrants constitute the control group, and do not receive the activation measures for a given period. The feasibility of such a scheme would have to be checked against laws and regulations.

It is also possible to test the impact of different variants of the benefits (or of the ways to deal with supplementary benefits). This could be done by randomly different versions of the scheme to a set of areas included in the experiment. In order to envision this, however, one needs a sufficiently large number of areas.

References

Blundell Richard, Mike Brewer and Andrew Shephard, *Evaluating the Labour Market Impact of Working Families' Tax Credit using difference-in-differences*, HM Revenue and Customs, 2005.

Bourguignon François, *Rapport final sur l'évaluation des expérimentations rSa - Comité d'Evaluation des expérimentations*, Haut commissaire pour la solidarité active contre la pauvreté, 2009.

Eissa *Nada and* Jeffrey B. *Liebman*, *Labor Supply Response to the Earned Income Tax Credit*, The Quarterly Journal of Economics, 1996.

Jones Maggie R, *The EITC and Labor Supply: Evidence from a Regression Kink Design*, Center for Administrative Records Research and Applications U.S. Census Bureau, 2013.

Michalpoulos, Charles, Doug Tattrie, Cynthia Miller, Philip K. Robins, Pamela Morris, David

Gyarmati, Cindy Redcross, Kelly Foley and Reuben Ford, *Making Work Pay: Final Report on the Self-Sufficiency Project for Long-Term Welfare Recipients*, Social Research Demonstration Corporation, 2002.

Michalopoulos, Charles Philip K. Robins, David Card, *When financial work incentives pay for themselves: evidence from a randomized social experiment for welfare recipients*, Journal of Public Economics 89, p. 5-29, 2005.

Ying Lei, Charles Michalopoulos, *SSP Plus at 36 Months: Effects of Adding Employment Services to Financial Work Incentives*, Social Research Demonstration Corporation, 2001.

Example 3: How to evaluate a long-term care reform?

1. Introduction

This case study discusses how to evaluate the impact of long-term care (LTC) provisions. It presents the main research designs available and how these designs have been used in the past. The case study shows that these designs vary essentially in:

1. Their methodological requirements; and
2. The assumptions that must be made regarding the comparability of the beneficiaries and control groups.

This discussion is illustrated by examples taken from the US, the UK and Slovenia.

The case study is organised as follows: Section 2 provides an overview of long-term health policy; Section 3 briefly describes the features of a typical reform; Section 4 explains how to build counterfactual situations, and section 5 discusses the different possible methods that could be applied, illustrated by concrete examples.

2. Long-term care management at a glance

There will be more than twice as many old people aged over 80 years old in 2050 than there are now. The share in the population will rise from 4.7%to 11.3% across 27 EU Member States. Between one quarter and one half of them will need help in their daily lives (OECD/European Commission, 2013). Healthcare systems are often ill-equipped to respond to the rapid rise in patients with multiple health problems, including reduced functional and cognitive capabilities. Care for such people may become fragmented between different professionals and organisations, with attendant risks to quality and safety from duplication or omissions of care. This has led to widespread calls for care to be better integrated (Curry & Ham, 2010).

Case management is a key feature of integration and is increasingly combined with use of tools to identify patients at risk of adverse outcomes (Lewis, Curry & Bardsley, 2011). Case management is defined as a "proactive approach to care that includes case-finding, assessment, care planning and care co-ordination" (Ross, Curry & Goodwin, 2011).

Evidence on the impact of case management is 'promising but mixed' (Purdy 2010). This is mainly because of the difficulty in attributing any tangible impact (e.g. reduction in hospital utilisation) to the case management intervention when there are multiple factors at play. This problem of attribution is common in the evaluation of schemes to reduce hospital utilisation (Steventon et al 2011; Purdy 2010). A further complication when assessing impact is that case management does not refer to a standard intervention; programmes can vary widely, which makes it difficult to make comparisons or generalised conclusions. The impacts of case management can also be difficult to quantify (for example, the impact on the patient experience and health outcomes). Furthermore, impacts may not be measurable in the short term, heightening the difficulties of attributing cause and effect.

There is, however, evidence that appropriately designed and implemented case management can have a positive impact on:

- Patient experiences;
- Health outcomes, including quality of life, independence, functionality and general well-being;
- Service utilisation, including hospital utilisation, length of stay and admissions to long-term care (see Ross, Curry & Goodwin 2011 for a review).

Case management has been found particularly effective when part of a wider programme where the cumulative impact of multiple strategies (as opposed to a single intervention) can be successful in improving care experiences and health outcomes (Powell-Davies et al 2008; Ham 2009). Despite the mixed evidence it is widely accepted that case management is a valid approach for managing individuals with highly complex needs and long-term conditions. For this reason, the approach is now widely used for the management of people with long-term conditions.

3. The reform

A reform in line with the above-mentioned principles would aim to address the inequities and fragmentation of long-term care provision by developing community-based and home-based services and unifying health and social care services.

More specific changes would include:

- A single-entry point for patients;
- A uniform expert procedure for LTC needs assessment;
- A process for preparing individual care plans; and
- Training for informal carers.

The person in need of LTC would then decide on whether to opt for services in kind or cash-benefits. The threshold, the scope and the content of the rights and provisions are important elements.

4. How to build counterfactuals

In order to evaluate the impact of case management, one needs to build a counterfactual[36]; i.e. compare beneficiaries of the new scheme with similar beneficiaries of existing (non-integrated) provisions.

In addition to allow the comparison of a new scheme with existing provisions, counterfactuals can also be used to estimate the impact of different aspects of a same reform and in this way, to elicit the most cost-effective ones. This is highly valuable from a policy perspective, as public expenditure can be optimised by focusing on the policy options that have the greatest impact on the desired outcomes. Besides, counterfactuals help estimate the heterogeneity of impact on different sub-populations (e.g. men vs. women) in order to envision targeting.

5. Potential for the Different Counterfactual Impact Evaluation Methods

5.1. Matching

Description
This method constructs a counterfactual by matching participants to non-participants using a set of observable characteristics. Successful matching requires a preliminary research to identify the different variables that could be statistically related to the likelihood of participating in the programme and to the outcomes of interest. Large samples are required in order to create sufficient matches. This method provides an estimate of the effect of an intervention for all the participants that can be successfully matched to a non-participant.

Assumptions
The main assumption is that all relevant background characteristics influencing participation and the outcomes of interest can be observed and accounted for.

Example
Challis and Davies used matching to evaluate the impact of the Community Care Scheme in Kent (UK). The scheme attempted to tackle both the problem of

[36] *See ESF guide on counterfactual evaluation: Design and Commissioning Of Counterfactual Impact Evaluations. A Practical Guidance for ESF Managing Authorities, European Commission, 2013.*

substandard care provision and to reduce the fragmentation of services using two separate but related strategies. These were: (i) greater flexibility of response to need so as to enhance service context; and (ii) improved case management through the clear responsibility of a key worker for a defined caseload to integrate services into a coherent 'package of care'.

In order to provide a comparative basis for the evaluation, the effects of care for those receiving the scheme were compared with the experience of similar cases from adjacent areas. Individual cases were matched by factors likely to be predictors of survival in the community. These were age, sex, household composition, presence of confusional state, physical disability and receptivity to help. As a result of this process 74 matched pairs receiving the new service and standard provision could be identified for comparison.

The evaluation showed that there were significant improvements both in subjective well-being and quality of care for the recipients of Community Care compared with those elderly people in receipt of standard services.

Applicability to the reform above
In order to estimate the impact of the reform, one could match beneficiaries in the former scheme to individuals in the new scheme based on a set of observed characteristics, as it was done in the evaluation of the Community Care Scheme in Kent. An alternative use of matching could be to compare the effect of various forms of case management and services on individuals participating in the new scheme, instead of comparing the initial scheme to new scheme.

Notice however, that in this context, matching presents an important shortcoming. Indeed, the method requires that all the relevant characteristics determining participation and influencing the outcomes of interest can be observed and accounted for. This is in general a strong hypothesis, and in this particular case a highly improbable one. Here it is important to note that benefit take-up is voluntary, thus it is difficult to ascertain how much of the impact is determined by the Community Care Scheme, and how much is driven by pre-existing differences in the observable and unobservable characteristics of individuals that chose to participate.

5.2. Regression discontinuity design (RDD)

Description
Regression Discontinuity Design compares individuals just above a given continuous eligibility threshold, with those just below. Those individuals are arguably very similar, and the threshold determines if they are exposed or not to the intervention being evaluated. The bandwidth between the lower and the upper limit containing the threshold determines the sample size.

Assumptions
This method relies on the assumption that the intervention implements a clearly quantifiable selection criterion based on some continuous score, and that participants are unable to anticipate and manipulate the scoring close to the cut-off point. Also, it assumes the individuals just below and just above the threshold are not significantly different.

Applicability to the reform above
As the pre-test variable must be on a continuous scale, the selection of instruments available to measure effectiveness is somewhat limited. One possible candidate could be the date of implementation of the scheme. Applicants eligible for integrated care after a certain date will be assigned to the new case management scheme and those before would remain under the existing provisions. The cut-off date must be determined retroactively so that carers cannot manipulate their entry into the new

scheme. Also, the initial care provisions must keep running in parallel to the new, integrated scheme (at least during the evaluation), as individuals in each scheme are observed during the same period of time.

Another approach would require using one of the variables that determines eligibility to the scheme. In principle, this is a viable option when enrolment is based on predictive risk models. Such models use statistical algorithms to predict an individual's level of future risk of hospital admission (Billings et al 2006; Nuffield Trust 2011) In practice however, most programmes use a combination of a predictive model and clinical judgement: the model is used to flag individuals who are at high risk, and the clinician then makes a judgement as to whether a person is likely to benefit from case management. Even in that case, one can use the threshold of the risk index to separate two similar populations close to the threshold, even though it does not entirely determine the intervention (this is called the 'fuzzy' design, see the ESF guide on counterfactual evaluation, (op. cit.)).

Importantly, risk assessment requires good-quality data. The most powerful predictive models require access to an individual's prior hospital admission records, as well as GP records and accident and emergency attendances. Social care data can also add predictive power. Yet, this might not always be available.

A fundamental criterion necessary for obtaining an unbiased estimate of an intervention effect is to have an assignment process that is completely known and perfectly measured (Shadish et al. 2002). The underlying premise of the RD design is that participants located immediately adjacent to the cut-off are the most similar and, therefore, provide the best comparison units for assessing intervention effect.

If the cut-off is strictly adhered to, the RD design controls for most threats to validity simply because any given bias would have to affect the intervention group causing a discontinuity that coincides with the cut-off. While it is theoretically possible, the likelihood of such an occurrence is remote.

5.3. Difference-in-difference (DiD)

Description
Differences-in-Differences (DiD) compares the change in outcomes before and after the start of the programme, over time for participants and for non-participants. DiD provides a measure of the impact for the whole population of participants, while controlling for constant conditions (observed and unobserved) that may be correlated with both the final outcomes and with the fact of being part of the control group.

DiD can either compare a group of individuals that is eligible to receive the intervention, to an arguably similar group that is not eligible, or can compare pilot areas where the programme is introduced to control areas that do not receive the programme.

Assumptions
DiD relies on the 'parallel trends' assumption. That is, in order to determine that the difference in outcomes is due to the programme, the trends in outcomes of the participants and of the non-participants should have been the same in the absence of the programme, and the composition of each group must remain unchanged. One way to validate the credibility of the parallel trends assumption is to check if both groups witnessed parallel changes before the introduction of the programme. Other alternatives to support this assumption include performing "placebo" tests on fake intervention groups or on fake outcomes, as well as undertaking the DiD analysis using different control groups.

Example

In 2008, the English Department of Health invited applications from healthcare organisations offering innovative approaches to providing better integrated care following concerns that, especially for older people, care was becoming more fragmented. The government deliberately gave no guidance on how integration should be achieved, rather encouraging a range of diverse approaches to be developed 'bottom up' by those providing care. Although this produced a diverse range of interventions, a common approach adopted by pilot sites was case management of older people identified as being at risk of emergency hospital admission. In these interventions, the main integrating activities were between surgeries and other community-based health services.

A 2012 evaluation reported the outcome for the six case management sites, including staff reports of changes to their own work and to patient care, changes in patients' experience, and changes in hospital utilisation and costs (Roland et al., 2012). A difference-in-differences analysis was conducted to compare two groups of patients in terms of hospital utilisation in the six months before the intervention and the six months after: patients confirmed to have received the intervention and patients under a different scheme. The analysis showed a significant increase in emergency admissions and significant reductions in both elective admissions and outpatient attendances for intervention patients compared to controls.

A concern in this type of studies is that systematic differences might exist between intervention and control groups that are unobserved and therefore cannot be balanced between groups. The evaluators did suggest that the two groups were not strictly comparable.

There were other challenges in drawing conclusions from this study. For instance, the pilots represented a somewhat heterogeneous group of interventions, and moreover they adapted and changed during the course of the pilot period, reflecting the changing health care environment in which they were operating. Thus, the idea that the evaluation describes a single simple intervention is somewhat far-fetched.

Applicability to the Reform above

Differences-in-Differences helps to control for environmental factors (new policies, economic conjuncture) that could induce a change in labour supply. However, the method relies on the assumption is that there are no unknown shocks—other than the intervention—that could have affected differently the outcomes of the intervention and the control group. This is a strong hypothesis, as it is very difficult to rule out the existence of such unknown shocks.

One way of applying DiD to the LTC reform would be to pilot the intervention in areas that are reasonably representative of the territory as a whole – or at least not different in any essential socio-economic and demographic terms. Control areas would also need to be identified from the same pool of candidate areas. The difference in outcomes between pilot and control areas after implementation of the policy would be interpreted as the impact of the intervention.

The 'parallel trends' hypothesis is hard to justify if pilot and control areas are different, for instance because social services are performing better in pilot areas. This hypothesis can be tested by selecting two groups of areas that entered the pilot at different times. If it can be shown that both intervention and control areas had similar hospital admission or mortality rates before the introduction of the programme, then it can be assumed that controls and pilots are comparable. Any significant difference between the two groups occurring after the introduction of the programme can be considered as an impact of the intervention.

5.4. Randomised Controlled Trials (RCTs)

Description
Randomised Controlled Trials measure the average impact of a policy or programme by randomly assigning entities to intervention and control groups, and then comparing the difference in outcomes.

Assumptions
Assumptions are weaker than with other protocols because random assignment from sufficiently large samples ensures that individuals are similar on average, both in terms of observable and unobservable characteristics. One must assume that people do not behave differently because they are aware of being in an experiment. This assumption applies to all types of experiments, including randomised controlled trials and natural experiments.[37]

It is also important that the scheme being evaluated is well defined and mature, and therefore similar to what would be implemented if the scheme is generalised at a broader scale. Again, this is not specific to randomised controlled trials, as it applies to any sort of evaluation that aims to estimate the impact of an intervention that is gradually phased-in.

Example
The systematic review conducted by You *et al.* (2012) found 10 RCTs. One of them attempted to integrate acute and long-term care services (Applebaum et al., 2002). The intervention relied on targeted staff resources, improved communication, and presumed provider interest in delivering the best service possible.

The study recruited chronically disabled older people receiving in-home services, who were at risk of using a high amount of acute services. Half of the patients were assigned at random to a clinical nurse care manager (NCMs), who, in conjunction with the programme care managers, was tasked to improve the linkage between the acute and long-term care services used by programme enrolees. A geriatrician supervised the NCMs.

Although there was some variation in health use and cost across intervention and control groups over the 18 month time period, the authors concluded that there were no differences between groups on any of the outcome variables examined. Efforts to integrate the acute and long-term care systems proved more difficult than anticipated. The intervention, which attempted to create integration through high intensity care managers, but without financial or regulatory incentives, was simply not strong enough to produce significant change for the clients served. The programme was also affected by various organisational changes, such as changes in the management of the hospitals involved in the studies, with repercussions on the way they communicated with NCMs.

Applicability to the LTC Reform above
The example above is a good guide. An interesting feature of this approach is that the respective impacts of different ‘variations’ of the policy can be evaluated separately. For example, one could imagine an intervention with different ‘arms’, testing the effect of smaller caseloads (e.g., 30 patients per care manager) against larger caseloads (e.g., 100 patients per care manager) or the effect of subsidised care against cash payments to the client (See Yordi et al, 1997 for a real example).

[37] *Evaluation driven effects occur when the subjects change their behaviour because they know they are part of a study, and not because of the intervention being studied.*

References

Applebaum R, Straker J, Mehdizadeh S, Warshwa G, Gothelf E (2002). 'Using high-intensity care management to integrate acute and long-term care services: substitute for large scale system reform?' *Care Management Journal*, 3(3): 113–119. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12632877

Challis D, Davies B (1985). 'Long Term Care for the Elderly: the Community Care Scheme'. *British Journal of Social Work*. 15(6): 563-579. Abstract available at:

http://bjsw.oxfordjournals.org/content/15/6/563.short

Curry N, Ham C (2010). *Clinical and Service Integration: The route to improved outcomes. London: The King's Fund*. Available at: www.kingsfund.org.uk/publications/clinical_and_service.html

Ham C (2009). 'The ten characteristics of a high-performing chronic care system'. *Health Economics, Policy and Law*, vol 5, pp 71–90.

http://www.ncbi.nlm.nih.gov/pubmed/19732475

Lewis G, Curry N, Bardsley M (2011). *Choosing a predictive risk model: a guide for commissioners in England.* London: Nuffield Trust. Available at: www.nuffieldtrust.org.uk/publications/choosing-predictive-risk-model-guide-commissioners-england

OECD/European Commission (2013). *A Good Life in Old Age? Monitoring and Improving Quality in Long-term Care*. Paris: OECD Publishing. Available at:

http://www.oecd.org/els/health-systems/PolicyBrief-Good-Life-in-Old-Age.pdf

Powell-Davies G, Williams A, Larsen K, Perkins D, Roland M, Harris M (2008). 'Coordinating primary health care: an analysis of the outcomes of a systematic review'. *Medical Journal of Australia*, 188(8): S65–S68. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/18429740

Purdy S (2010). *Avoiding Hospital Admissions: What does the research say?* London: The King's Fund. Available at: www.kingsfund.org.uk/publications/avoiding_hospital.html

Steventon A, Bardsley M, Billings J, Georghiou T, Lewis GH (2011). *A Case Study of Eight Partnership for Older People Projects (POPP): an evaluation of the impact of community-based interventions on hospital use.* London: Nuffield Trust. Available at:

http://www.nuffieldtrust.org.uk/sites/files/nuffield/publication/An-evaluation-of-the-impact-of-community-based-interventions-on-hospital-use-summary-Mar11.pdf

Roland M, et al. (2012). 'Case management for at-risk elderly patients in the English integrated care pilots: observational study of staff and patient experience and secondary care utilisation'. International Journal of Integrated Care. Vol. 12. Available at: https://www.ijic.org/index.php/ijic/article/view/URN%3ANBN%3ANL%3AUI%3A10-1-113731/1771

Ross S, Curry N, Goodwin N (2011). *Case management. What is it and how can it best be implemented?* London: Kings Fund. Available at: www.kingsfund.org.uk/publications/case_management.html

Yordi C, DuNah R, Bostrom A, Fox P, Wilkinson A, Newcomer R (1997). 'Caregiver Supports: Outcomes from the Medicare Alzheimer's Disease Demonstration'. Health Care Financial Review, 19(2): 97-116. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/10345408

You EC, et al. (2012). 'Effects of case management in community aged care on client and carer outcomes: a systematic review of randomized trials and comparative observational studies'. *BMC Health Services Research*, 12:395. Available at:

http://www.biomedcentral.com/1472-6963/12/395

Definitions

Assumption Accepted cause and effect relationships, or estimates of the existence of a fact from the known existence of other fact(s).

Baseline The baseline is the standard against which all subsequent changes implemented by an intervention are measured.

Conclusion validity Conclusion validity is the degree to which conclusions reached about relationships in the data are reasonable (Trochim and Donnelly, 2007).

Construct validity Construct validity refers to the degree to which inferences can legitimately be made from the operationalizations in a study to the theoretical constructs on which those operationalizations were based (Trochim and Donnelly, 2007).

Counterfactual A counterfactual is a conditional statement of how the people in a programme would have fared if the programme had never been implemented. This notion is used to understand the causal impact of the programme (Glennerster, Takavarasha 2013).

Effect size An effect size is a measure that describes the magnitude of the difference between two groups.

End-line The end-line is the measure at the end of a study.

Experiment An experiment is an orderly procedure carried out with the aim of verifying, refuting, or establishing the validity of a hypothesis. Experiments provide insight into cause-and-effect by demonstrating what outcome occurs when a particular factor is manipulated.

External validity External validity is the degree to which the conclusions of a study would hold for other persons in other places and at other times (Trochim and Donnelly, 2007).

Input An input is a resource or factor of production (labour, capital) used in the production of an organisation’s output.

Internal validity Internal validity is a property of scientific studies which reflects the extent to which a causal conclusion based on a study is warranted.

Intervention (policy) Action taken to improve a social problem.

Meta-evaluation Meta-evaluation (or meta-analysis) is the use of statistical methods to combine results of individual studies (Cochrane

Collaboration).

Probability sample

A probability sampling method is any method of sampling that utilizes some form of random selection (Trochim, 2007).

Programme

In public policy, a programme refers to a set of combined interventions.

Protocol

A protocol is the the detailed plan of a study. By convention, it is written according to the following format:

- Project title;
- Project summary;
- Project description (Rationale; Objectives; Methodology; Data management and analysis);
- Ethical considerations;
- References.

Systematic review

A systematic review attempts to identify, appraise and synthesize all the empirical evidence that meets pre-specified eligibility criteria to answer a given research question. Researchers conducting systematic reviews use explicit methods aimed at minimizing bias, in order to produce more reliable findings that can be used to inform decision making (Cochrane Handbook for Systematic Reviews of Interventions).

Theory of change

A Theory of Change (ToC) is a specific type of methodology for planning, participation, and evaluation that is used in the philanthropy, not-for-profit and government sectors to promote social change. Theory of Change defines long-term goals and then maps backward to identify necessary preconditions (Brest 2010).

Triangulation

In the social sciences, triangulation is often used to indicate that two (or more) methods are used in a study in order to check the results. The idea is that one can be more confident with a result if different methods lead to the same result.

Bibliography

Brest P. (2010). The Power of Theories of Change. Stanford Social Innovation Review. Spring.

Gertler, Paul J. Sebastian Martinez, Patrick Premand, Laura B. Rawlings, Christel M. J. Vermeersch (2011). Impact Evaluation in Practice, The World Bank. Available at:

http://siteresources.worldbank.org/EXTHDOFFICE/Resources/5485726-1295455628620/Impact_Evaluation_in_Practice.pdf

Glennerster R., Takavarasha K. (2013). Running randomized evaluations: A practical guide. Princeton University Press.

Haynes L., Service O., Goldacre B., Torgerson D. (2012). Test, Learn, Adapt. Developing Public Policy with Randomised Controlled Trials. Cabinet Office Behavioural Insights Team. Available at:

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/62529/TLA-1906126.pdf

HM Treasury, The Magenta Book. Available at:

https://www.gov.uk/government/publications/the-magenta-book

Jalan, J, Ravallion, M. (2003). 'Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching'. Journal of Business & Economic Statistics, 21(1), pp. 19-30. Available at:

http://info.worldbank.org/etools/docs/voddocs/172/353/ravallion_antipoverty.pdf

J-PAL Europe (2011). Social Experimentation: A methodological guide for policy makers. Available at: http://ec.europa.eu/social/BlobServlet?docId=10947&langId=en

Kostøl, AR, Mogstad, M (2014). 'How Financial Incentives Induce Disability Insurance Recipients to Return to Work'. American Economic Review, 104(2): 624-55.

Morris, S, Tödtling-Schönhofer, H, Wiseman, M (2012). Design and Commissioning of Counterfactual Impact Evaluations - A Practical Guidance for ESF Managing Authorities. European Commission, DG Employment. Available at:

http://ec.europa.eu/social/main.jsp?catId=738&langId=en&pubId=7646

Morris S, et al. (2004). Designing a Demonstration Project: An Employment Retention and Advancement Demonstration for Great Britain. Cabinet Office. Available at: http://goo.gl/FyrGni

Torgerson DJ, Torgerson CJ (2008). Designing Randomised Trials in Health Education and the Social Sciences. Palgrave McMillan, Basingstoke.

Trochim, WM. The Research Methods Knowledge Base, 2nd Edition. Internet WWW page, at URL: http://www.socialresearchmethods.net/kb/ (version current as of October 20, 2006).

World Bank: World Bank Evaluation Toolkit. Available at:

http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTHEALTHNUTRITIONANDPOPULATION/EXTHSD/EXTIMPEVALTK/0,,contentMDK:23262154~pagePK:64168427~piPK:64168435~theSitePK:8811876,00.html