Applicability of Impact Evaluation to Cohesion Policy[1]

Report Working Paper of

Matthew H. Morton

Department of Social Policy & Social Work
University of Oxford

January 2009

[1] This working paper has been written in the context of the report "An Agenda for a reformed Cohesion Policy". It represents only the opinion of the expert and does not necessarily reflect the views of the European Commission.

TABLE OF CONTENTS

Abstract

As the European Union (EU) has grown so too have the challenges for territorial cohesion. The EU is in the process of making its largest investments yet in social and economic policies intended to improve territorial cohesion and lift-up struggling regions. In light of major expansions of cohesion policy programmes, it is increasingly important to understand whether investments achieve the intended outcomes. An impact evaluation is defined as an outcomes study that seeks to understand the effects of policies by comparing them with 'counterfactual' situations—what would have happened if the policies were not implemented. Randomisation and quasi-experimental designs that establish 'control groups' are described and promoted for the purposes of assessing outcomes. Qualitative methods are highlighted as complementary instruments to impact evaluation, particularly with respect to formative questions seeking to better understand programme processes, implementation, and stakeholder perspectives. Finally, institutional considerations and evidence uptake are explored for successfully integrating impact evaluation within the efforts of the European Commission and subsidiary governments.

Introduction

In January, 2008, the European Commission assembled a working group consisting of policy evaluation experts and European public officials to help the European Union (EU) strengthen its strategies to develop quality evaluation of the impacts of cohesion policy programmes. This paper intends to inform the working group's proceedings with respect to the applicability of impact evaluation to cohesion policy.

Over the last two decades, the EU's regional policy in the form of strategic community investments has grown tremendously. In the mid-1980's, the European Regional Development Fund (ERDF) accounted for only 7.5 percent of the EU's sum budget (Bachtler et al, 2006). The latest allocation for the 2007-13 period constitutes 35 percent of the total EU budget (approximately €347 billion) (European Commission, 2007). This extraordinary investment aligned with the priorities set by the renewed Lisbon strategy offers an unprecedented opportunity to improve territorial cohesion and, in doing so, the vibrancy of the European community and the quality of life for its members.

In spite of the investments made and opportunities ahead, the effectiveness of cohesion policy has been questioned since its inception (Armstrong, 2007; Sapir et al, 2003). Making a clear case for cohesion policy's effectiveness is not easy. Many evaluations have had to rely on poor quality of monitoring data (Bachtler et al, 2006). Multi-level governance and 'additionality' translate to various parties converging for the purpose of evaluation, all of which may carry different characteristics and agendas. This, Baslé (2006) has argued, combined with the heterogeneity of local areas and interventions, has led to the "near-impossibility of obtaining…a reliable and credible assessment of impact, and true legitimisation of EU interventions" (Baslé, 2006).

Even so, the evaluation of EU regional policy has already made considerable progress: evaluation results are actively employed to inform policy-making, infrastructural development for evaluation has been stimulated at the member state and local levels, and the EU's own evaluation capacity and techniques are continually evolving (Bachtler et al, 2006; Bachtler et al, 2003; Baslé, 2006; Molle, 2006). Indeed, whilst further major advance is called for, there is much positive progress and plentiful assets to build on.

It is important to recognise that questions on cohesion policy effectiveness as mentioned above can be allocated to one of two categories. On one hand, one can question whether

specific cohesion policy programmes cause the anticipated effects on intended outcomes. Alternatively, one can ask whether the larger strategy of cohesion policy works. Policy-makers and stakeholders should distinguish between the two kinds of questions and avoid jumping to conclusions about the latter from results of the former.

This paper will promote impact evaluation as the best way to confidently measure the effects of specific policy interventions on clear outcomes. Impact evaluation as discussed in this paper is generally inappropriate for applying to the broader, macro-level questions of effectiveness typically reserved for macro-economic or statistical modelling.

Shifting the EU Evaluation Paradigm

Most EC evaluations of cohesion policy programmes are retrospective, carried out on or after the completion of an intervention by independent evaluators. They are intended to examine accountability for the use of resources and, as far as possible, to assess programme outputs. These ex-post evaluations generally include two preceding stages: the 'ex-ante' evaluation (providing baseline data) and the 'midterm evaluation' (conducted around the middle of the programme period, which assesses early results and allows for mid-programme adjustments). In the case of the URBAN initiative for example, an ex-post evaluation of 118 URBAN programmes included methods such as literature reviews, structured interviews, case studies, street surveys, and 'informed judgements of evaluation teams' (GHK, 2003).

In addition to an inability to isolate intervention effects from other influences, 'ex-post' designs tend to face significant problems with data availability and reliability that inhibit the specification of clearly measurable outcomes (Bachtler et al, 2006). Rowe and Taylor (2005) have criticised EC evaluation reports for being too long and complicated with ambiguous outcomes and discordant, contradictory messages. The EU is not alone in its challenges for producing adequate quantity and quality of impact evaluation. Reviews of reports produced by other major international governmental and non-governmental organisations also estimate very low presence of impact assessments and even fewer that used methodologies sufficiently strong to deduce outcomes (CDG, 2008). Impact evaluation is a natural struggle for anyone investing in changing lives and communities.

Impact evaluation encourages a shift in the focus of policy evaluation from the past to the future and promotes the use of evaluation for policy learning rather than simply for cost containment or control. Most, though not all, impact evaluations are prospective; that is, they seek to establish whether a policy would work, on the basis of designing the evaluation and collecting initial data before the intervention begins for the observed population. In this context, the impact evaluation would inform decisions about the policy's future while the process evaluation might be used to shape its implementation (Nutley et al, 2007).

Such a strategy tends to give perhaps unusual weight to 'evidence' as opposed to other considerations and pressures present in the policy process, a position that is not without its critics (Pawson et al, 2005; Smith and Grimshaw 2005). It is important to recognise that information generated by policy experiments is likely to be imperfect and will always need to be considered alongside expert and stakeholder opinion and theoretically sound judgement. Significantly, 'evidence based policy making' as promoted in the UK from 1997, was subsequently renamed as 'evidence informed policy making' (Walker and Duncan 2007; Newman 2005). Nevertheless, impact evaluation holds out the prospect of coherent and rigorous methodology and possibly more succinct conclusions that may appeal to the policy-making community.

Defining Impact Evaluation: What it Can and Cannot Answer

Within institutions it is not unusual to see a development in policy evaluation from a concern with audit and accountability (making sure that resources have been spent properly), towards an interest in outputs (what has been delivered through the policy) to, and, finally, a focus on outcomes (what has been achieved as a result of the policy). This development represents more than a change in the nature of the questions posed, asking 'what works?' instead of 'what was spent on what?'; it requires a significant change in the method of evaluation towards what is usually referred to as impact evaluation.

The World Bank defines impact evaluation as follows:

An impact evaluation assesses changes in the well-being of individuals, households, communities or firms that can be attributed to a particular project, program or policy. The central impact evaluation question is what would have happened to those receiving the intervention if they had not in fact received the program (World Bank, 2008).

'Well-being' is used in the definition in a very global sense and, of course, includes the possibility that the policy or program being evaluated leads to a reduction in well-being for some or even all individuals or institutions. As importantly, the definition requires the establishment of a counterfactual - that is, a measure of what would have happened in the absence of the policy - against which the policy can be compared. This is different and more challenging than merely establishing a baseline, namely the situation when the policy was introduced. It is usually achieved by creating a comparison or "control" group of individuals, households, communities or firms similar to those targeted by the policy but not themselves affected by it. The impact of the policy is then taken to be the difference in (change in) the well-being experienced by the programme group compared to the control group. The impact of a policy has to be inferred in this way; it cannot be directly measured by noting the change in an indicator applied only to the programme group since some of the change observed may be due to factors other than the policy.

It is perhaps worth noting that, whereas some EC documents reserve use of the term 'impacts' to describe policy effects that take place in the long-run, calling 'immediate and direct' effects 'results', different terminology is typically not used in impact evaluation to distinguish between short- and long-term outcomes. It is nevertheless important to recognise that the impact of a policy is not necessarily stable over time. Some policies can manifest immediate impacts that are subsequently reversed. It is more usual, however, for the impact of a policy, measured by the difference between the programme group and the control, to increase for a period and then subsequently to decline (Greenberg et al, 2004).

Impact evaluation is designed to answer whether a policy works. In so doing it is essential to define in advance what effects a programme is likely to have so that appropriate data may be collected. It is also necessary to decide in advance how large an effect is desirable. This is required so as to design an evaluation with the appropriate degree of statistical power to be able to detect an effect of the size expected. It can also engender a degree of discipline in the policy making process; it is wise to avoid the situation in which any change, however miniscule, can be claimed to have been a success.

Impact evaluation, on its own, generally cannot sufficiently address 'why' and 'how' questions (i.e. 'Why did a programme work or not work?' and 'How well was the programme implemented?' or 'How cost-efficient was the programme?'). To some extent,

impact evaluation can help us to understand the 'why a programme worked or did not work' question if we have enough information on programme characteristics to detect patterns by which outcomes are associated with certain programme qualities. For example, suppose we conduct an impact evaluation of a job training centre model, and we evaluate twenty centres that follow this model. Let us assume that only ten out of twenty of the programmes show positive effects on increasing employment. The same ten programmes, however, turn out to be the only ten programmes that included a one-on-one coaching component. In this case, we could reasonably infer that one reason for why some programmes worked and others did not was whether or not they involved an added one-on-one coaching component.

Frequently, however, the answers to 'why' and 'how' questions are not available by analysing immediately known characteristics of programmes. Therefore, impact evaluations are now almost always accompanied by process or formative evaluation, designed to explore and describe how the policy works. Formative evaluations are usually eclectic or qualitative in their methodology. Again it is essential, if the full benefits of policy evaluation are to be achieved, to specify in advance how a policy is thought likely to work; this specification is sometimes called a 'theory of change' (Griggs et al, 2008). A pre-specified theory of change allows the process of evaluation to be designed to focus on policy linkages or pathways through which a policy is postulated to have an effect. It may also permit the impact evaluation to be refined so as to investigate certain of these causal pathways.

The literature sometimes refers to the triangulation of impact and process evaluations. It is important not to interpret this as meaning that results are triangulated to determine which provides the 'correct' answer. Rather, the different kinds of evaluation have different goals and their results are potentially complementary and cumulative. One difficult challenge in the management of evaluation is to ensure that the impact and formative evaluations are integrated throughout the process rather than a simple synthesis being produced at the end (Walker and Wiseman, 2006).

Impact Evaluation Designs and Techniques

The defining feature of an impact evaluation in this context is the establishment of a counterfactual. The use of a counterfactual control group helps us to understand what would happen to a population if a specific policy or programme were not implemented. Without this counterfactual, we cannot be certain how much change in behaviour is actually the result of the policy.

To serve as a reliable counterfactual, members of the 'control'[2] group should be identical to those in the 'programme' group participating in the policy in question. If members of the control group have different characteristics from the programme group, it is difficult to separate the effects of the intervention from the consequences of the differences between groups. These non-programme factors that cause variance between the control group and intervention group are called 'confounding variables' - if they cause some of the difference - or 'spurious variables' - if they cause all of the difference.

---

[2] Strictly speaking, the term 'control group' is sometimes reserved for randomised experiments because quasi-experimental designs are characterized by less control over reducing variation between groups. The term 'comparison groups' can be most appropriate for non-randomised evaluations. This paper, however, uses the two terms interchangeably to reduce confusion. It should also be noted that some scholars believe the term 'control' group is irrelevant even for randomised experiments in social research since they do not take place in a highly controlled laboratory setting.

Frequently the differences between groups are very difficult to detect. We call these unknown factors 'unobservables'. A good impact evaluation design will compare groups that are as similar as possible on both observable and unobservable variables.

Properly conducted, an evaluation that randomly assigns participants to groups before the intervention begins constitutes the best method for establishing a counterfactual due to its ability to make intervention and comparison groups statistically equal (apart from chance alone) on both observable and unobservable characteristics. When randomisation (an 'experimental design') proves to be impractical, evaluators commonly opt for other methods to establish a counterfactual; these are called quasi-experimental designs. Because quasi-experimental designs cannot establish a counterfactual situation with the same level of confidence as randomisation, the challenge is to identify and, as far as possible, to minimise the effect of observable confounding or spurious variables. Little can be done about the effect of unobservable variables.

This paper concentrates on five types of quasi-experimental designs: the regression-discontinuity design; statistical matching; difference-in-difference; natural experiment; and pretest-posttest. The 'designs' are better described as orientations rather than as prescriptive methods, each with many variations that cannot here be discussed in detail. Regardless of the experimental or quasi-experimental design used, certain requisites for high-quality impact evaluation are always advisable. These include:

- *Theory of change* - Making the case for why this particular approach should improve well-being, designers of the programme should illustrate prior to evaluation (and preferably before commencing the programme) how and why specific programme features will have intended effects on specific, measurable outcomes.

- *A clear research question* - The question should clearly state the effect of what (intervention), on what (measurable outcomes), and on whom (defined population). The lack of a thoughtful, explicit research question will result in problems throughout the evaluation process.

- *The question and context should dictate the methodology* (Sackett et al, 1997) - Evaluation design is a reflective art rather than adherence to a blueprint[3].

- *Prospective evaluation* - Constructing a reliable counterfactual, and therefore a better estimate of effects, generally requires a shift from evaluating programmes retrospectively to before the participants begin to receive the intervention. Integrating evaluation design into policy design is ideal.

- *Adequate sample size* - Too large a sample size could waste time and resources whilst too small will jeopardize the results. There is no magic number for sufficient sample size[4]; it depends on a statistical calculation based on what evaluators determine to be a reasonable effect size with a reasonable level of power while allowing for a reasonable margin of error[54].

---

[3] See the visual from a World Bank paper steps (Baker, 2000) as a guide for designing and implementing impact evaluation in Appendices.

[4] Effect sizes are often smaller with social policies and interventions. This does not mean that an intervention is unimportant, but simply that it competes with many other influences on human and organizational behaviour

[5] Software packages have been designed to calculate adequate sample sizes.

- *A plan for measuring implementation fidelity* - Measuring the level of 'fidelity' (adherence) to the way the programme was intended to be implemented in its actual implementation is important for understanding how much impact evaluation results reflect the intervention and not poor implementation.

- *Appropriate balance between rigorousness and feasibility* - 'Rigorousness,' in this context, is defined as the ability to reduce systematic error and bias that diminish the accuracy of conclusions drawn from an evaluation. The most rigorous designs are not necessarily always the most appropriate or feasible, but the optimal balance involves the most rigorous design possible given the question and circumstances (Evans, 2003). It may sometimes be possible to rank designs in order of ability to control for bias—usually starting with an experimental design—to choose the most rigorous one that is achievable.

- *Detailed and transparent reporting* - Reporting of results should be accompanied by transparent description of methods, participant selection and allocation, participant dropouts, implementation fidelity, ethical considerations, and any potential bias. More detailed guides for reporting experimental and quasi-experimental evaluations are provided by the CONSORT Statement (Altman et al, 2001) and the TREND Statement (Des Jarlais et al, 2004) respectively.

Experimental Designs

The most reliable way to ensure similarity between groups is to allocate individuals, organisations, or areas at random to the programme group that receives the policy intervention and to a control group that does not. This process of randomisation guarantees that, on average, the two groups will be statistically equal on all observable and unobservable characteristics. The term experimental evaluation is usually reserved for evaluations in which cases are randomised.

Because randomisation controls for observable and unobservable confounding and spurious variables, theoretically no 'before-programme' study is needed to ensure that the comparison groups are alike (Purdon, 2001). In practice, randomisation only ensures completely identical groups (apart from differences due to chance alone) when very large numbers of cases are assigned and "before programme" measurements are therefore usually still made.

There are sometimes political and practical considerations that make randomisation difficult. It may be difficult to devise a workable method for randomly allocating individuals or organisations into the two groups, or it may be too costly to do so especially when the units to be randomly allocated comprise regions, neighbourhoods or labour markets. Sufficient monitoring systems may also not be in place to ensure that persons assigned to the control group do not find ways to participate in the programme.

Additionally, critics frequently caution that it is unethical to deny some people an intervention for the purpose of evaluation. There are at least three responses to this concern. First, if it is already known that the intervention will have positive effects and therefore those in the control group will be 'unethically' deprived of a positive intervention, there is no point in conducting the evaluation in the first place (Resen et al, 2006). Secondly, policies and interventions are often not universal and resources rarely match the scale of need. Consequently, a denial of services often takes place regardless of randomisation; randomisation simply systematises the allocation process and allows one to learn from those not receiving services. Finally, evaluators often use techniques to

ensure that a control group is established without ultimately denying the group policy services.

One such technique that is commonly used is the waiting list method. In this method, participants are randomised to receive the intervention immediately or to a waiting list to receive it later. Those on the waiting list serve as a control group until such time as they receive the intervention. The most important shortcoming of the waiting list method is that long-term outcomes cannot be assessed; although, quasi-experimental methods can be coupled with randomised waiting list designs in order to create a long-term counterfactual.

	Randomised Controlled Trial (RCT) (a.k.a. random assignment, randomisation or experimental design)	
Description	Random assignment ensures that each study participant has an equal chance of being assigned to the intervention or control group. On average, random assignment makes the personal characteristics of participants in each comparison group statistically equivalent. This design can uniquely neutralise potentially spurious and confounding variables - even those that the evaluator has not thought of or believes s/he cannot measure (Langbein, 1980). It is not only possible to randomise individuals, but also entire neighbourhoods, areas, schools, organisations, etc. Indicators of interest are measured for both groups before and after the intervention, and net effects are compared for a reliable estimate of the outcomes attributable to the policy or intervention.	
Resources	The CONSORT statement (Altman et al, 2001) provides guidance for reporting RCT's. In addition to making evaluation reports more consistent in format and thus more easily comparable, these standards for reporting help readers better understand the appropriateness and reliability of methods used. Among other things, reporting standards are suggested for how participants were determined eligible, research objectives, participant attrition, and methods used to randomise, prevent researchers from knowing who was in which group for data collection, and statistical analyses. All of these aspects of reporting relate to different sources of bias.	
Pros & Cons	– Greatest ability to control for selection bias and systematic error, given adequate sample size – Can be a transparent rationing mechanism when there are not enough resources for the intervention to reach all of those in need – Requires less data and background knowledge of study participants	– May be considered unethical to deny services – Political barriers – May not always be practically feasible given constraints of expertise, resources or time – May be difficult to ensure that selection is truly random – By itself, an RCT is often unable to detect unintended consequences – 'Blinding' (especially 'double-blinding')[6] to prevent 'Hawthorne Effects' and 'substitution bias' is very difficult, if not impossible, for many social experiments (GSR, 2005)[76]

[6] 'Blinding' referrers to keeping group assignment secret to either the participants or the investigators. 'Double blinding' means keeping group assignment secret to both.

[7] 'Hawthorne Effects' occur when a study participant behaves differently because s/he is aware of being studied as a member of an intervention or control group. 'Substitution bias' occurs when a member of the control group seeks a 'substitution' treatment to compensate for not receiving treatment from the studied intervention, thereby diminishing the achievement of a 'counterfactual' or 'no-treatment' comparison.

	Randomised Controlled Trial (RCT) (a.k.a. random assignment, randomisation or experimental design)
Variations & Techniques	(a) Waiting list design. (b) Randomised encouragement design. (c) area/place-level randomisation. (d) Random assignment with random selection. (e) Longitudinal/long-term follow-ups. (f) 'Stratifying' randomisation if evaluators wish to ensure that a population with a certain characteristic is equally allocated between the two groups.
Example	The Positive Parenting Program ('Triple P') provides examples of randomised evaluations of a multi-component, multi-level intervention in Australia and Hong Kong (Leung et al, 2003; Sanders, 1999). The unit of randomisation was individuals, and the study used waiting list control groups.

Quasi-Experimental Designs

When randomisation proves impractical, evaluators commonly opt for other methods to establish a counterfactual; these are called quasi-experimental designs. Because quasi-experimental designs cannot establish a counterfactual situation with the same level of confidence as randomisation, the challenge is to identify and, as far as possible, to minimise the effect of observable confounding or spurious variables. Little can be done about the effect of unobservable variables.

The term 'quasi-experimental designs' encompasses a wide variety of methodologies. The variations of these designs differ substantially with respect to their ability to construct a reliable counterfactual. This paper concentrates on five types of quasi-experimental designs: the regression-discontinuity design; statistical matching; difference-in-difference; natural experiment; and pretest-posttest. The 'designs' are better described as orientations than as prescriptive methods, each with many variations that cannot here be discussed in detail.

Regression-Discontinuity Designs (RDDs) take advantage of quantitative measures frequently used to determine participant eligibility for a policy programme for the purpose of establishing a control group. For example, evaluators may need to measure the impact of loans for small enterprises on research and innovation production. The evaluators could use the cut-off of fewer than 50 employees for a business to qualify as a "small enterprise". Businesses with 50 employees or more are thus ineligible for the loan. Assuming that there would be little systematic difference between businesses just below and above the cut-off point, evaluators would assess the difference in outcomes measurements for businesses barely eligible (say, 45-49 employees) before and after the intervention against a control group of businesses barely ineligible (50-54 employees) before and after the intervention; the difference would provide an estimate of effects. Problematically, however, RDD still cannot control for unobservable variables (such as motivation to take up the loan), it does not measure impacts of populations on the extreme sides of the eligibility spectrums (such as businesses with only 10 employees), and not all interventions lend themselves to clearly quantifiable selection criteria.

For Difference-in-Difference (DID), we can use the example of teacher training funded by cohesion policy. A 2007 report on the human capital programme plans for 2007-13 from the Polish Ministry of Regional Development explains the shortcomings of its existing model for teacher training in preparing teachers with adequate practice, skill, and knowledge breadth to meet vocational education needs (Ministry of Development, 2007). Evaluation of the impacts of a new EU-funded teacher training programme on student achievement outcomes could draw a sample of teachers not receiving the new training that are similar on certain characteristics to a group of teachers receiving the training (e.g. age, teaching experience, or supervisor assessment). The evaluators could measure performance outcomes for students of both groups of teachers prior to the training and one year after the training. The difference in performance outcomes of students over the measured time period for the intervention group subtracting the difference in performance outcomes of students for the control teachers (hence the name, difference-in-difference) provides an estimate of the new teacher training's impact on student outcomes.

If the evaluation team has access to greater background data, it may opt for constructing a prospective control group through statistical matching. The sophistication and confidence of statistical matching methods vary with data availability. Methods such as Propensity Score Matching (PSM) allow evaluators to use many characteristics that are believed to influence changes in behaviour for an intervention and control group outside of the programme, calculate a score with these variables, and create intervention and control groups that appear to be as similar as possible on all observable characteristics in order to reduce the likelihood of bias.

Like DID, A Pretest-Posttest (PP) evaluation would measure the difference in student achievement outcomes before the teacher training and one year after the training. Importantly, however, evaluators would not establish a control group. In this case, the counterfactual is assumed to simply be the same intervention group before the intervention. This approach is generally the simplest, least resource and expertise intensive form of impact evaluation discussed here. However, the lack of a control group leaves impact estimates very suspect because it becomes virtually impossible to disentangle differences in changes in behaviour attributable to the programme from other influences.

Sometimes, natural conditions or policy changes not orchestrated by evaluators automatically create intervention groups and control groups. These so-called 'natural experiments' can operate like a DID design with a control group or a PP design without a control group. For example, some local governments within a Member State might pass legislation to implement certain waste recycling measures whilst other local governments do not. Evaluators could take advantage of the occurrence to measure the effects of the waste recycling measures on changes in citizens' recycling behaviour by using the areas enacting the policy as intervention groups and the others as control groups. Cases such as this leave concerning opportunity for uncontrolled confounding variables, but they may be useful when evaluator-constructed intervention and control groups are unrealistic.

Each design carries different strengths and weaknesses for validity and feasibility under different circumstances. They should be carefully weighed against a given evaluation question and context in order to determine the most appropriate approach.

	Regression-Discontinuity Design (RDD)	
Description	RDD requires that the selection process rank individuals (or organizations, neighbourhoods, etc.) on a quantitative scale of criterion (e.g. income level or test scores). Applicants above a certain cutoff score are eligible for the intervention; those below are not. The comparison groups are identified from just below and above the cutoff line. Evaluators then measure specific indicators for both groups at roughly the same time before and after the intervention. The net difference in outcomes between the two groups can provide a reliable estimate of the impact of the policy.	
Pros & Cons	– Establishes a potentially strong counterfactual – Does not require assigning needy individuals/areas to control group – Can be seen as more politically realistic option for assigning comparison groups – Growing in popularity in the policy evaluation arena	– Threats to internal validity – Appropriate only when the distinction between treated and untreated groups rests on a clear, quantitative eligibility criterion – If the intervention is not universally used by (or mandated to) all eligible individuals above the cut-off score, or if samples from both sides of the cutoff line are not randomised, the evaluators cannot control for motivation as a potentially confounding variable – The assumption of comparability is challenged if the cut-off point represents a significant difference between eligible and ineligible groups (e.g. if the same cut-off score is used for other types of services)
Variations & Techniques	a) To eliminate the effects of 'motivation' or 'volunteerism,' experimental methods can be nested within an RDD by setting a range on either side of the cut-off point in which eligibles and ineligibles may be randomised to control or intervention groups. (b) Waiting list control groups. (c) Longitudinal/long-term follow-ups.	
Example	The Italian government currently uses RDD to measure the outcomes of subsidies to SME's and micro-enterprises funded by the EU in Southern Italy	

	Statistical Matching Design
Statistical Matching Design	Matching constructs a comparison group through a one-to-one matching basis of known, shared characteristics. In principle, matching attempts to match the comparison group and intervention group closely enough so as limit the difference between the two groups to whether or not they received the intervention. Successful matching necessitates strong preliminary research, which will inform investigators of the wide range of variables known to be statistically related to both the likelihood that a unit will choose treatment and the outcome measure (GSR, 2005).

	Statistical Matching Design	
Pros & Cons	– Can be performed after a programme has already been implemented if there is sufficient data – If the investigator can identify and measure the spurious and confounding variables, and locate untreated and treated participants similar on these factors, then reasonable internal validity can be attained	– Potential for selection bias in which the nonparticipant group is systematically different from the participant group due to unobservable variables – Requires substantial knowledge of the characteristics of the participants and nonparticipants – Can require advanced statistical expertise – Large samples are needed to create sufficient matches
Variations & Techniques	– (a) Propensity score matching (PSM). (b) Cell matching. (PSM is generally more amenable to matching a large number of variables than cell matching. Consequently, in many cases PSM is preferable over cell matching.)	
Example	(1) A Czech labour market programme evaluation attempted to determine whether participants in five different programs were more successful in re-entering the labour market than were nonparticipants and whether this varied across subgroups and with labour market conditions. The evaluation identified several characteristics by which it matched members in the participant pool with members from the nonparticipant pool who shared the most characteristics (Baker, 2000). (2) Although described as a 'conditional difference-in-difference design,' an impact evaluation to evaluate whether EU-funded research and development (R&D) business investments crowd out private R&D investments in Flanders and Germany used a control group determined by statistical matching (Aerts, 2006). This example illustrates both the advantages as well as the complexities involved with applying matched comparison group designs to cohesion policy programmes.	

	Difference-in-Difference (DiD) (a.k.a. non-equivalent comparison group design (NECG, pretest-posttest with comparison group (in some uses))
Description	In this case, a control group can be selected from individuals (groups, areas, etc.) who share some characteristics with the intervention group (e.g. eligibility for the programme, income level, or annual budget). The two groups are measured before and after the programme. The control group that did not participate in the programme is measured for 'natural change', and the intervention group is measured for 'natural change' plus change due to the programme. Subtracting the difference for the control group from the difference for the intervention group gives an estimate of the change due to the introduction of the programme or policy.

	Difference-in-Difference (DiD) (a.k.a. non-equivalent comparison group design (NECG, pretest-posttest with comparison group (in some uses))	
Pros & Cons	– Can be an easier selection process than other designs with more rigorous selection standards	– Difficult to attribute results to the intervention and not confounding or spurious variables – Lack of longitudinal information
Variations & Techniques	Establish several standardised posttest follow up points so as to collect longitudinal data on outcomes.	
Example	An evaluation using the DiD design evaluated the impact of the 'workfare' programmes in Norway on enhancing self-sufficiency in the labour market. The study compared 300 programme recipients with approximately 150 non-participants social assistance recipients. The design was considered a practical alternative to randomisation (Dahl, 2003).	

	Natural Experiment	
Description	Natural experiments define comparison groups from naturally occurring, not artificial, events. It can be seen as a version of the difference-in-difference design. Natural experiment designs can occur in policy, for example, when a policy measure is universally applied to a population and a comparison group is chosen from a different population that is as similar as possible.	
Pros & Cons	– Perhaps the most politically benign way to establish a comparison group because the investigators take no part in the denial or postponement of services	– Shares the problems associated with the NECG design plus increased difficulty to establishing a similar comparison group (given that natural experiments often respond to universal interventions forcing them to look to very different populations)
Variations & Techniques	(a) With comparison group or without comparison group. (b) Occasionally, natural experiments can even be 'randomised' if policy naturally uses a randomised system for allocating limited resources (e.g. Colombia's private school vouchers distributed through national lottery process (Angrist, 2002))	
Example	An evaluation of the Earned Income Tax Credit (EITC) reforms on the employment of single mothers in the U.S. The study uses single women without children as a natural control group (Eissa, 1996).	

	Pretest-Posttest (PP) (a.k.a. reflexive comparisons or before-and-after design)	
Description	With PP, the population, or a sample of the population, that is exposed to an intervention is measured on certain characteristics before the intervention and then again after the intervention. The baseline provides the comparison and impact is measured by change in outcome indicators before and after intervention (e.g. neighbourhood crime rate before and after intervention). It may be preferable that the posttest data collection time is stated from the start so as to ensure a more objective data collection period	
Pros & Cons	– may be particularly useful when interventions are truly universal – more reliable than simple posttest – does not deny or postpone services	– Does not provide a counterfactual, therefore cannot confidently attribute behaviour changes to the intervention (observation group could have changed for reasons other than the programme over the time period)
Variations & Techniques	(a) Conducting multiple follow-up data collections makes this a single interrupted time-series design (SITS). SITS has the added advantage of detecting only short-term responses to extreme scores over time. (b) If baseline data is available, the investigator could create a 'retrospective pretest' as an alternative to directly collecting data for outcomes measures before the intervention started (Pratt, 2000).	
Example	The U.S.-funded 'Favela-Bairro Project' in Brazil was a $180 million, multi-component project, which aimed to improve the state of urban slums and integrate them into "the fabric of the city." The evaluation assessed outputs and outcomes using the population census, particularly to assess impacts on rents, investment, and mortality (Soares, 2005).	

Qualitative Methods

The techniques and methods of qualitative evaluation are too many to outline here. Moreover, though important, qualitative methods are not the focus of this paper as they cannot credibly measure outcomes by establishing a counterfactual. Nevertheless, it is worth including qualitative evaluation in this section in order not to lose sight of its role in the larger picture of evaluation. Qualitative methods can play a powerful role in conjunction with quantitative methods in order to tell a more complete story of programme challenges and achievements.

Notably, qualitative methods are sometimes associated with an inherent lack of consistent structure or standardisation. Researchers, however, have developed standards for conducting and reporting qualitative research and tools for appraising studies against such standards. While some qualitative researchers criticise set standards as inappropriate for qualitative studies, others posit that they produce more useful, rigorous, and interpretable research (Patton, 2003).

	Qualitative Evaluation	
Description	Qualitative methods include four kinds of data collection: (1) in-depth, open-ended interviews; (2) focus groups; (3) direct observation; and (4) review of written documents (Patton, 2002). Qualitative research can be conducted in ways that ensure extra levels of rigour and systematic approaches to reduce bias (Mays and Pope, 1995).	
Resources	Researchers have developed tools for conducting, reporting, and appraising qualitative research to improve its usefulness. See CASP (2001) for a concise list of questions for appraisal of qualitative research and Patton (2003) for a more detailed checklist for qualitative evaluation	
Pros & Cons	– Helpful for explaining why programmes work or not for certain groups or individuals – Calls attention to details and unmeasured outcomes – Assesses process – Qualitative methods can often help in the design of interventions, further research questions, and quantitative evaluations – Can help to determine outcomes – Often lends itself better to community participation in research	– Flexibility can lend itself to haphazard methods and outcomes – Naturally subjective and therefore characterised by personal biases – Special risk of researcher and reporting biases as investigators make decisions about what to ask, what material to include, and what inferences to make – Conclusions can be especially complex and contradictory, leaving difficult messages to communicate to practitioners and policy-makers – Unable to establish reliable programme effects – Can require significant time and interpersonal communication expertise (e.g. for interviews and focus groups)
Variations & Techniques	Focus groups. Interviews. Case studies. Field observations. Literature reviews	
Example	The Neighbourhood Nurseries Initiative is an example of a policy-level evaluation that integrated quantitative and qualitative methodologies. Qualitative methods were used in tandem with quantitative impact evaluations to assess child centre quality. The qualitative study collected observational data with a standardised assessment form (Smith et al, 2007). See below for more details.	

Considering Designs within the Context of Cohesion Policy Characteristics

In all cases, strong impact evaluation requires proficiency in evaluative methods and thoughtful, creative planning to tackle the unique challenges of each situation. This section concentrates on additional considerations for addressing evaluation challenges associated with two common characteristics of cohesion policy programmes—and, indeed, many other large-scale public policies and programmes. These include a tendency to have multiple components packaged within an intervention (and varied components among interventions packaged under the same policy), and the reality that, for many cohesion policy programmes, the intervention targets not individuals, but whole areas, groups, or institutions.

Multi-component programmes:

Applying impact evaluation designs to multi-component interventions will provide an estimate of effects just as they would for simpler interventions. The challenge with evaluating multi-component interventions is that it is difficult to know which component(s) are the 'active ingredients'—the features that are responsible for positively changing behaviour—and which do not contribute to positive outcomes. In other words, if the research question simply asks whether or not a policy works, then the evaluation considerations for impact evaluation remain constant. If the evaluation aims to answer other useful questions, however, such 'which aspects of the intervention work?', or 'are we spending excessively on additional components that add little to intended outcomes?', then more complex evaluation designs are required.

To identify the active ingredients, the evaluators could compare the effects of programmes that are similar with the exception of variations in specific components. For example, a Member State may offer financial incentives to small and medium enterprises (SME's) to encourage research and innovation through the EU-funded JEREMIE initiative. The JEREMIE intervention might include several financial instruments such as loans, equities, venture capital, and guarantees. The question of how the different financial instruments (or combinations of instruments) affect outcomes requires additional steps to the evaluation process.

One approach the evaluation team might recommend involves a stratified evaluation, which creates more than one comparison group. In addition to a control group, other comparison groups could include SME's with access to just grants, just equities, just venture capital, or a combination of instruments. The evaluators could also create comparison groups based on the size of loans, venture capital, or equities to assess how the scale of financial support impacts outcomes. Alternatively, instead of creating various comparison groups prospectively, evaluators might conduct a 'sub-group' analysis, which retrospectively looks for trends in the statistical analyses after evaluation data has already been collected.

There is an important caveat to establishing more than one comparison groups or retrospective subgroup analyses: they reduce statistical power. Subgroups should be used sparingly in order to avoid 'fishing' for outcomes and drawing from insufficient subgroup samples (Assman et al, 200). The more comparison groups established or sub-groupings analysed, the larger the sample size that is required. This obstacle can sometimes render such strategies unfeasible.

Some posit the use of qualitative methods to 'distill' the active ingredients of multi-component programmes (Miller et al, 2003). In this case, the investigators will conduct interviews or focus groups with questions intended to explore practitioners' and/or participants' perspectives with respect to each of the programme components so as to get a better idea of what they found useful, and how each component affected their behaviour. Significantly, however, qualitative methods are limited by the fact that they are dependent on the subjective views of individuals and cannot reliably establish causal relationships.

Impact Nevertheless, they may be helpful for making better-informed inferences about which components were more and less helpful, and qualitative findings can advise further quantitative research.

Place-level programmes:

Cohesion policy often supports group and area-level interventions in which the unit of evaluation is larger than individuals or households. These play a valuable role because they recognise the need for improving the larger context, institutions, and conditions in which individuals are situated. Social science continually grows with better methods for developing rigorous evaluations with groups and places.

Randomised evaluation for groups and places is often described as 'cluster random assignment' or a 'cluster-based experiment.' Cluster-based experiments have been successfully implemented and may prove usefulness for assessing the impacts of cohesion policy programmes including neighbourhood-level interventions such as URBAN, or potentially school-level interventions in countries such as Hungary that aim to reduce educational disparities among ethnic groups and improve school efficiency.

Sometimes, cluster-based evaluation is used as a technique to mitigate 'contamination'—a problem that occurs when control group members are impacted by the intervention and therefore diminish the reliability of the counterfactual. If, for example, the government provides job training to individuals in small communities, it may be difficult to prevent non-recipients from participating in the training or being influenced by participants of the intervention group. This would bias the results. Thus, the government could provide the job training to whole communities and withhold or postpone job training from other 'control' communities (preferably through random allocation). In such cases, contamination is less likely to occur across area units than within them.

Even in evaluating the effects of a policy on groups or places, it is advisable for investigators to identify evaluation units at the lowest level possible whilst still attaining the benefits of cluster evaluation. For instance, if schools can be evaluated rather than school districts, or neighbourhoods (or blocks) instead of towns, the clusters will often be more manageable and effects more detectable. The smaller the outreach of the intervention and the larger the area, the less pronounced the effects will likely be, and thus more difficult to measure.

Again, sample size can be a substantial obstacle for cluster-based evaluation (Donner and Klar, 1996). Achieving an adequate sample size of places, groups, or institutions can pose greater difficulty than individuals. Navigating the politics and decision-making processes to evaluate areas, groups, or institutions can often be more cumbersome than for individuals as well. Finally, in cluster-based programmes, intervention heterogeneity—variation—is common, because the differences between groups and places are likely to lead to differences in programme implementation. For this reason, implementation and process evaluations can be particularly important complements to cluster-level impact evaluations.

Making Impact Evaluation Explicit, Relevant, and Useful for Stakeholders

The simple fact of knowledge production does not necessarily translate to knowledge use. This challenge speaks to the demand side for evaluation; policy makers both within Brussels and the Member States need policy evaluation to be transparent, intelligible, unambiguous, and non-threatening. The following three issues expand on priorities for fostering a 'culture of evaluation' around cohesion policy.

Clarity: A mission and message of learning

Impact evaluation, properly conducted, should indicate whether or not a policy works. Complemented by stakeholder input, explicitly stated theories of change and process evaluation, it can also reveal in what respects a policy works well or less well and the likely reasons. Taken together, this evaluative package provides a secure basis for the assessment of policy and a foundation for policy learning. It may also counter the fears sometimes precipitated by the 'judgemental', 'all or nothing' application of impact evaluations set in the context of accountability and potential blame (Cousin et al, 2005; McCartney et al, 207; Sanderson, 2002; Weiss, 1998).

It is important that policy-makers, stakeholders, and the general public properly understand the intended aims of evaluation. At heart it is about better governance, about learning whether policies (and their implementation) are effective and whether money is well or poorly spent. It is, of course, 'inherently political' (Taylor and Balloch, 2005) especially insofar as it is likely to have an effect on resource allocation. The better job the EC does of explicitly promoting its intentions, the better it will be able to minimise any unintended and unhelpful politicising of the evaluation process (Grosee, 2006).

In facilitating the growth of impact evaluation, the EC should stress the primary aim of impact evaluation as being a vehicle for policy learning and for improving policy and policymaking. Frequently in the public discourse of evaluation, failure is stigmatised as the problem to be fixed. Failure, however, can be viewed as a primary - vital even - part of the process to innovation and ever-growing solutions.

The problem is not failure; the problem is the lack of awareness of failure and the lack of support, opportunities, and - when needed - pressure to respond to knowledge of failure. The EC should encourage a culture of learning from evaluation that does not demonstrate positive results, rather than a culture that fears and incentivises the avoidance of any negative feedback through over-politicising impact evaluation results.

Usefulness: Making impact evaluation practical for learning

It is important to ensure that the ownership of impact evaluations is appropriately located. Normally this means locating responsibility as close as possible to the implementation of the policy subject to the need for independence. Evaluation usually requires the involvement of all kinds of actors and levels of staff engaged in the design and delivery of the policy as well as that of the intended recipients.

Recent communications from the EC on negotiations for cohesion policy strategies reinforce the importance of multi-level ownership and partnerships to achieve Lisbon-related objectives (Commission,2008). The same is crucial to ensure the usefulness of impact evaluation within the EU. Cooperation and commitment are generally fostered by offering as many people as possible a stake and a say in evaluation design, implementation, and findings (Minkler and Wallerstein, 2003; Strand et al, 2003).

Impact evaluation complemented by other methods—such as qualitative process evaluations, user satisfaction surveys, and cost-benefit analyses—can provide more information to explain the outcomes of an impact evaluation, which provides instruments for learning and improvement. Evaluation units and stakeholders should arrange venues such as special meetings, working groups, or online exchanges to serve as feedback loops to discuss the implications of evaluation findings.

Relevance: Engaging the complex networks of actors involved with evidence uptake

A growing body of research in 'knowledge translation' and 'diffusion of innovations' in the public sector emphasises the complex, interdependent interactions between proponents and resistors of innovations that will determine the fate of evidence uptake (Dopson, 2005). The successful uptake of impact evaluation and its findings within the EU will likely depend on recognition of its own complex interdependencies and finding creative ways to make impact evaluation relevant for the actors involved.

Establishing opportunities for evaluation findings to be shared can democratise the use of research findings and increase their utility across the European community (and beyond). An internet-based clearinghouse for data and evaluation research for the EU could serve as one possible venue for such sharing, if an adequate resource does not already exist for the EU's needs. Clearinghouses have been constructed by various governments, NGO's, and foundations to address similar data-sharing and evidence-sharing needs. Furthermore, the clearinghouse concept fits well with the EU's current promotions of “e-governance” (European Union, 2004).

Symposiums, evaluation unit exchange visits, shared trainings, and regional reviews of research could also serve as meaningful venues for improved sharing of evaluation designs and outcomes between the EC, member states, regions, and localities so as to promote learning and a healthier 'knowledge society’.

Institutionalising Impact Evaluation for post-2013

The successful pursuit of more and better impact evaluation in relation to cohesion policy requires the development of a supportive infrastructure. At minimum, of course, this requires the commissioning and undertaking of impact evaluation and the administrative systems necessary to do this. Moreover, it necessitates the fostering of an evaluation community with the skills to undertake such evaluations effectively.

Cohesion policies are such that they do not readily lend themselves to ‘recipe’ evaluations; rather, the development of appropriate designs and their implementation are likely to require world-class evaluation expertise. While there is a strong case for DG-Regio and DG-Employment to reach out for expertise, it is likely to be a quite limited supply, at least until impact evaluation becomes better established in Europe.

The working group for which this paper is written has consistently championed a modest approach to instigating impact evaluation in the EU. This first involves recognition that impact evaluation is intended to complement, not supplant, other forms of evaluation. Secondly, initial steps towards integrating impact evaluation should not aim too broadly. The EC should work with select programmes in Member States particularly interested in advancing assessment of outcomes to experiment with just a few pilot impact evaluations. The experience of these pilots will help to inform future attention to improving the institutional setting for successful impact evaluation to occur.

Looking ahead, it is advisable for the EU to establish the improvement of institutional setting for impact evaluation as a priority within its larger ambition to strengthen the public "institutional capacity" at all levels in Convergence regions and Cohesion countries (Commission, 2008). Substantial amounts of funds are already allocated to Member States for the purpose of evaluating cohesion policy programmes. The EC could be more intentional about guiding the use of these funds towards strategic capacity building around evaluation with attention to impact evaluation. The focus on internal infrastructure

development for evaluation has the advantages of improved sustainability, ownership, consistency, knowledge of evaluated programmes, and development of an institutionalised culture of evaluation. Disadvantages may include the threats to objectivity if internal evaluation units have a less independent character than external evaluators[8].

Some evaluation units may need to develop their infrastructure in such a way that better integrates qualitative and quantitative evaluation. This integration could be achieved by improving the resources of and coordination between separate teams of experts that specialise in quantitative and qualitative analysis, or by constructing hybrid evaluation teams which consist of individual experts for various types of methodology.

Finally, it seems that quality impact evaluations may occur more than we realise at a Member State and Regional level. In addition to investigating these examples, the EC might also develop a registry of evaluators with whom subsidiary governments are contracting on impact evaluation needs. In partial response to DG-Regio's expressed challenges with a lack of quality external evaluation help, this list of names and institutions, if one does not exist already, could provide valuable information on evaluation resources categorised by thematic and methodological expertise. A shared resource of this nature could provide value to the EC as well as Member State, Regional, and Local governments.

Conclusion

There is no magic prescription for mapping impact evaluation methods onto cohesion policy programmes. Advances will require significant time, ingenuity, resources, and cooperation from all levels of governance.

The contents of this paper aspire to establish some common language and provoke discussion to this end.

Impact evaluation cannot answer the question of whether cohesion policy as a macro-level strategy 'works'. It can, however, play a critical role in estimating the effects of specific cohesion policy programmes on specific outcomes. The confidence one can place in such estimates depends largely on the ability of the impact evaluation design to construct a reliable counterfactual that allows investigators to assess the difference in outcomes for individuals, groups, or places before and after an intervention, accounting for natural changes in behaviour that would have occurred had the programme not existed.

Random assignment of participants to an intervention and control group is the most reliable way to ensure that any differences in outcomes between the intervention and control group, apart from those caused by the intervention, occurred by chance alone. Evaluators, however, may not find randomised designs to be achievable for all situations; in these cases, several quasi-experimental designs can be employed with varying degrees of ability to establish a credible counterfactual.

---

[8] There is reason to doubt that this would be the case (e.g. comments from DG-Regio of examples in the past in which external evaluators sought to provide DG-Regio with the outcomes it "wanted to see" even in spite of explicit directions from DG-Regio for an independent assessment.) External evaluators could, in some cases, be motivated by a perceived advantage of producing favorable outcomes in order to receive further contracts.

Many cohesion policy programmes, like other public policy schemes, present additional challenges for the task of impact evaluation given their scope and complexities. Place- and group-level units of evaluation have been previously pioneered with rigorous experimental methods, and multiple comparison groups and subgroup analyses constitute some of the approaches for separating the effects of multiple programme components. Nevertheless, these approaches can be more resource-intensive and reduce statistical power. Greater industriousness, resource investments, and, occasionally, a reworking of the intervention design or evaluation question altogether may be essential to properly measure programme impacts.

The demands of quality impact evaluation over a complex network of stakeholders will require attention to capacity-building, coordination, and institutional setting across all relevant levels of government. The EC and subsidiary governments should begin modestly by facilitating a handful of pilot impact evaluations in order to experiment with methods as well as navigating institutional challenges.

Finally, Impact evaluation will need to be located high in the EU's priorities as it influences other major priorities, including better governance, better resource allocation, and better results for social and economic cohesion. The task of fostering a culture of impact evaluation in the European community is best served by treating impact evaluation as part of a larger framework of research and evaluation that promotes continued learning and improvement within the context of a vibrant 'knowledge society'.

References

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., et al. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. Annals of Internal Medicine, 134(8), 663-694.

Angrist. (2002). Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. Nashville, Tenn. [etc.]: American Economic Association.

Angrist, J. (2003) Randomized Trials and Quasi-Experiments in Education Research, Cambridge, Massachusetts:NBER Reporter: Research Summary, Summer, http://www.nber.org/reporter/summer03/angrist.html

Armstrong, H. (2007). Regional policy. In A. El-Agraa (Ed.), The European Union (8th ed., pp. 425). Cambridge, UK: Cambridge University Press.

Assmann, S., S. Pocock, L. Enos, L. Kasten. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. The Lancet. 355(9209), 1064-1069.

Bachtler, J. and C. Wren. (2006). Evaluation of European Union cohesion policy: Research questions and policy challenges. Regional Studies, 40(2), 143.

Bachtler, J. and Taylor, S. (2003). Community added value: Definition and evaluation criteria (IQ-Net Report on the Reform of the Structural Funds. Glasgow, UK: European Policies Research Centre.

Baker, J. (2000). Evaluating the impact of development projects on poverty: A handbook for practitioners. Washington, D.C.: The World Bank Group.

Baslé, M. (2006). Strengths and weaknesses of European Union policy evaluation methods: Ex-post evaluation of objective 2, 1994-99. Regional Studies, 40(2), 225.

Campbell Collaboration. (2007). Steps in proposing, preparing, submitting, and editing of campbell collaboration systematic reviews. Retrieved 02/20, 2008, from http://www.campbellcollaboration.org/guidelines.asp

CASP. (2001). 10 questions to help you make sense of qualitative research Critical Appraisal Skills Programme.

CGD (Center for Global Development). (2006). When will we ever lean? Improving lives through impact evaluation. Report of the Evaluation Gap Working Group. May. Center for Global Development; Washington, D.C.

Chalmers, I. (2003). Trying to do more good than harm in policy and practice: The role of rigorous, transparent, up-to-date evaluations. The ANNALS of the American Academy of Political and Social Science, 589; 22

Commission of the European Communities. (2008). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on the results of the negations concerning cohesion policy strategies and programmes for the programming period 2007-2013. COM(2008) 301/4, Brussels.

Cousin, G., J. Cousin and F. Deepwell. (2005). Discovery through dialogue and appreciative inquiry: A participative evaluation framework for project development. In Taylor, D. and S. Balloch (Ed.), The politics of evaluation: Participation and policy implementation (pp. 109). Bristol, UK: The Policy Press.

Dahl, E. (2003). Does 'workfare' work? The norwegian experience. International Journal of Social Welfare, 12(4), 274-288.

Des Jarlais, D., C. Lyles, N. Crepaz, K. Abbasi, et al. (2004). Improving the reporting quality of nonrandomized evaluations of behavioural and public health interventions: the TREND statement. American Journal of Public Health. 94(3), 361-367.

Donner, A. and N. Klar. (1996). Statistical considerations in the analysis of community intervention trials. Journal of Clinical Epidemiology. 49(4), 435-439.

Dopson, S. (2005). The diffusion of medical innovations: can figurational sociology contribute? Organizational Studies. 26, 225, Sage Publications.

European Commission. (May 2007). Growing regions, growing europe: Fourth report on economic and social cohesionEuropean Union. Retrieved 16/01/2008 from http://ec.europa.eu/regional_policy/sources/docoffic/official/ reports/cohesion4/pdf/4cr_en.pdf

European Union. (2004). Fourth European Conference on e-Government Dublin, Ireland.

Evans, D. (2003). Hierarchy of evidence: A framework for ranking evidence evaluating healthcare interventions. Journal of Clinical Nursing, 12(1), 77-84.

GHK. (2003). Ex-post evaluation: URBAN community initiative, 1994-1999.

Glasziou, P., Vandenbroucke, J., & Chalmers, I. (2004). Assessing the quality of research. BMJ, 328(7430), 39-41. doi:10.1136/bmj.328.7430.39

Greenhalgh, T. (1997) How to read a paper: Papers that summarise other papers (systematic reviews and metaanalyses). BMJ, 315(7109), 672-675.

Greenberg, D and Shroder, M. (2004) Digest of Social Experiments, Washington: Urban Institute Press (third edition)

Greenberg, D., Ashworth, K., Cebulla, A., and Walker, R. (2004) 'Do Welfare-To-Work Programmes Work For Long?' Fiscal Studies, 25(1): 27-53.

Grosee, T. (2006). Euro-commentary: An evaluation of the regional policy system in poland: Challenges

and threats emerging from participation in the EU's cohesion policy. European Urban and Regional Studies, 13, 151.

Gruendel, J. and Aber, J.L. (2007). Bridging the gap between research and child policy change: The role of strategic communications in policy advocacy. In J.L. Aber, S.J. Bishop-Josef, S.M. Jones, K.T. McLearn & D.A. Phillips (Ed.), Child development and social policy: Knowledge for action (pp. 43). Washington, D.C.: APA Press.

GSR. (2005). Background paper 7: Why do social experiments? Experiments and quasi-experiments for evaluating government policies and programs. London: Government Social Research Unit, HM Treasury.

Langbein, L. (1980). Discovering whether programs work: A guide to statistical methods for program evaluation. Glenview, IL: Scott, Foresman and Company.

Leung, C., Sanders, M. R., Leung, S., Mak, R., and Lau, J. (2003). An outcome evaluation of the implementation of the triple P-positive parenting program in Hong kong. Family Process, 42(4), 531-544.

Mays, N., and Pope, C. (1995). Qualitative research: Rigour and qualitative research. BMJ, 311(6997), 109-112.

McCartney, K. and Weiss, H.B. (2007). Data for democracy: The evolving role of evaluation in policy and program development. In J.L. Aber, S.J. Bishop-Josef, S.M. Jones, K.T. McLearn and D.A. Phillips (Ed.), Child development and social policy: Knowledge for action (pp. 59). Washington, D.C.: APA Press.

Minkler, M., and Wallerstein, N. (2003). Community based participatory research for health. San Francisco, CA: Jossey-Bass.

Ministry of Development, Poland. (2007). Human Capital Operational Programme. Government Report, September 07, Warsaw.

Moher, D., D. Cook, S. Eastwood, I. Olkin, D. Rennie, and D. Stroup. (1999). Improving the quality of reports of metaanalyses of randomised controlled trials: The QUOROM statement. The Lancet, 354, 1896.

Molle, W. (2006). Evaluating the EU cohesion policy. Unpublished manuscript.

Newman, T., Moseley, A., Tierney, S., and Ellis, A. (2005). Evidence-based social work: A guide for the perplexed . Dorset, UK: Russell House.

Nutley, S., I. Walter, and H. Davies. (207). Using evidence: How research can inform public services. Bristol, UK: The Policy Press.

O'Brien, T. (1995). In the lake of the woods. USA: Penguin Group.

Patton, M. (2003). Qualitative evaluation checklist. Unpublished manuscript. Retrieved 19 Feb 2008, from http:// www.wmich.edu/evalctr/checklists/qec.pdf

Patton, M. Q. (2002). Qualitative research and evaluation methods (3rd ed.). Thousand Oaks, CA: Sage.

Pawson, R., T. Greenhalgh, G. Harvey, and K. Walshe. (2005). Realist review – a new method of systematic review designed for complex policy interventions. Journal of Health Services Research and Policy, (3), 21-34.

PMSU (2003) The Magenta Book: Guidance Notes for Policy Evaluation and Analysis, London: Cabinet Office, Prime Minister's Strategy Unit, Government Chief Social Researcher's Office. http://www.policyhub.gov.uk/magenta_book/ index.asp

Pratt, C. C., McGuigan, W. M., and Katzev, A. R. (2000). Measuring program outcomes: Using retrospective pretest methodology. The American Journal of Evaluation, 21(3), 341-349.

Purdon, S., C. Lessof, K. Woodfield, and C. Bryson. (2001). Research Methods for Policy Evaluation. Research Working Paper No. 2. UK Department for Work and Pensions on behalf of the Controller of Her Majesty"s Stationary Office. http://dwp.gov.uk/asd/asd5/WP2.pdf.

Resen, L., Manor, O., Engelhard, D., and Zucker, D. (2006). In defense of the randomized controlled trial for health promotion research. American Journal of Public Health, 96(7), 1181. doi:1075235191

Rowe, M. and Taylor, M.. (2005). Community-led regeneration: Learning loops or reinvented wheels? In David Taylor and Susan Balloch (Ed.), The politics of evaluation: Participation and policy implementation (pp. 205). Bristol, UK: The Policy Press.

Ryer, J. (2008). The seven stages of FrameWorks learning. Retrieved 02/18, 2008, from http:// www.frameworksinstitute.org/strategicanalysis/sevenstages.shtml

Sackett, D., and Wennberg, J. (1997). Choosing the best research design for each question. BMJ, 315(7123), 1636.

Sanders, M. (1999). Triple P-positive parenting program: Towards an empirically validated multilevel parenting and family support strategy for the prevention of behavior and emotional problems in children. Clinical Child and Family Psychology Review, 2(2), 71.

Sanderson, I. (2002). Evaluation, policy learning and evidence-based policy making. Public Administration, 80(1), 1-22.

Sapir, A., et al. (2003). An agenda for a growing Europe: Making the EU economic system deliver

Smith, I. and L. Grimshaw. (2005). Evaluation and the new deal for communities: Learning what from whom? In Taylor, D. and S. Balloch (Ed.), The politics of evaluation: Participation and policy implementation (pp. 189). Bristol, UK: The Policy Press.

Smith, T., K. Coxon, M. Sigala, K. Sylva, S. Mathers, G. Smith, I. La Valle, R. Smith, and S. Purdon. (2007). National evaluation of the neighbourhood nurseries initiative: integrated report No. SSU/2007/FR/024)Her Majesty's Printer and Controller of HMSO 2007. Retrieved 23 Feb 2008 from http://www.dfes.gov.uk/research/data/uploadfiles/ SSU2007FR024.pdf

Soares, Fabio and Yuri Soares. (2005). The socio-economic impact of favela-bairro: What do the data say? (Working Paper No. OVE/WP-08). Washington, D.C.: Inter-American Development Bank.

Strand, K., Marullo, S., Cutforth, N., Stoecker, R., and Donohue, P. (2003). Community-based research and higher education. San Francisco, CA: Jossey-Bass.

Taylor, D. and S. Balloch. (2005). The politics of evaluation: An overview. In Taylor, D. and S. Balloch (Ed.), The politics of evaluation: Participation and policy implementation (pp. 1). Bristol, UK: The Policy Press.

Thomas, J., Oakley, A., Oliver, A., Sutcliffe, S., Rees, K., Brunton, R., and Kavanagh, J. (2004). Integrating qualitative research with trials in systematic reviews. London: BMJ Pub. Group.

Walker, R. and Duncan. S. (2007). 'Knowing What Works: Policy evaluation in central government', Pp. 169-190 in H. Bochel and S. Duncan (eds.) Making Policy in Theory and Practice, Bristol: Policy Press.

Walker, R. and Wiseman, M. (2006)., 'Managing evaluations' pp. 360-383 in I. Shaw, J. Greene and M. Mark (eds) Handbook of Policy Evaluation, London: Sage.

Walker, R. (2002). Evaluation: Evidence for public policy. Unpublished manuscript.

Weiss, C. H. (1998). Have we learned anything new about the use of evaluation? The American Journal of Evaluation, 19(1), 21-33.

World Bank. (2008). Overview: Impact evaluation. Retrieved Feb/14, 2008, from http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,menuPK:384339~pagePK:162100~piPK:159310~theSitePK:384329,00.html#whatis

Appendices

Appendix I

From World Bank paper (Baker, 2000).

Box 2.1 Main Steps in Designing and Implementing Impact Evaluations

During Project Identification and Preparation

(1) Determining whether or not to carry out an evaluation
(2) Clarifying objectives of the evaluation
(3) Exploring data availability
(4) Designing the evaluation
(5) Forming the evaluation team
(6) If data will be collected:

- (a) Sample design and selection
- (b) Data collection instrument development
- (c) Staffing and training fieldwork personnel
- (d) Pilot testing
- (e) Data collection
- (f) Data management and access

During Project Implementation

(7) Ongoing data collection
(8) Analyzing the data
(9) Writing up the findings and discussing them with policymakers and other stakeholders
(10) Incorporating the findings in project design

Appendix II

Synthesising Evaluation Evidence

The systematic review may not be very relevant to the needs of the EC at this stage, particularly if it is well known that few decent studies exist to review. Nevertheless, it can be helpful to know about. While primary evaluations are critical for measuring a programme's effectiveness, a summary of primary evaluations can help to inform future programme and policy measures. The systematic review is probably the most reliable way to summarise existing evidence.

A common misperception is that a systematic review is synonymous with a quantitative meta-analysis, or even that it necessarily includes a quantitative meta-analysis. This is not accurate. A meta-analysis can be a helpful device for summarising quantitative effects measured via common outcomes. If this data is not available, however, it is not necessary for the completion of a systematic review.

	Systematic Review	
Description	A systematic review is an overview of primary studies, which contains an explicit statement of objectives, materials, and methods and has been conducted according to explicit and reproducible methodology (Greenhalgh, 1997). It is justified on the basic premise that science is a cumulative activity - suggesting that synthesised knowledge from many studies is better than isolated knowledge from one study (Chalmers, 2003). The primary goal of a systematic review is not to conclude whether something 'works' or not. The most important goal is to provide a synthesised statement on the state of the evidence for a specific question. Depending on the state of the evidence, the review may or may not draw conclusions as to whether or not an intervention works and in what context.	
Resources	The Campbell Collaboration (2007) provides detailed guidance for rigorous systematic reviews in social research (Campbell Collaboration, 2007). The QUOROM statement (Moher et al, 1999) provides guidance for reporting the meta-analyses of randomised controlled trials (when meta-analyses are appropriate). Pawson et al (2005) provide steps for a 'realist' version of systematic review, which puts more emphasis on detailed analyses and conclusions with the intent to provide a more context-relevant appraisal of complex policies and interventions.	
Pros & Cons	– Provides a synthesised summary of what all of the available evidence tells us, and doesn't tell us, on a particular question – Can provide a more generalisable statement of effects than a single, isolated primary study – Can highlight where further research is needed	– Often focus on narrow, clearly defined questions or interventions, which may not reflect complex, multi-component policy realities – Methods for qualitative systematic reviews are being explored but still fairly undeveloped – Systematic search requires access to good databases – Language barriers may prevent relevant studies from being included – Primary studies that meet the quality inclusion criteria of systematic reviews are frequently more likely to come from wealthier, Western countries, which raises questions to the generalisability of the results to dissimilar regions
Variations & Techniques	Integrating qualitative research in systematic reviews (Thomas et al, 2004)	
Example	The Campbell Collaboration provides supports and a clearinghouse for rigorous systematic reviews with standardised methods in the areas of Social Welfare, Methods, Education, and Crime and Justice.	