HOW COUNTERFACTUALS GOT LOST ON THE WAY TO BRUSSELS

Alberto Martini
Università del Piemonte Orientale, Italy
email: amartini@prova.org

Prepared for the Symposium
"Policy and programme evaluation in Europe: cultures and prospects"
Strasbourg, July 3-4, 2008

Summary

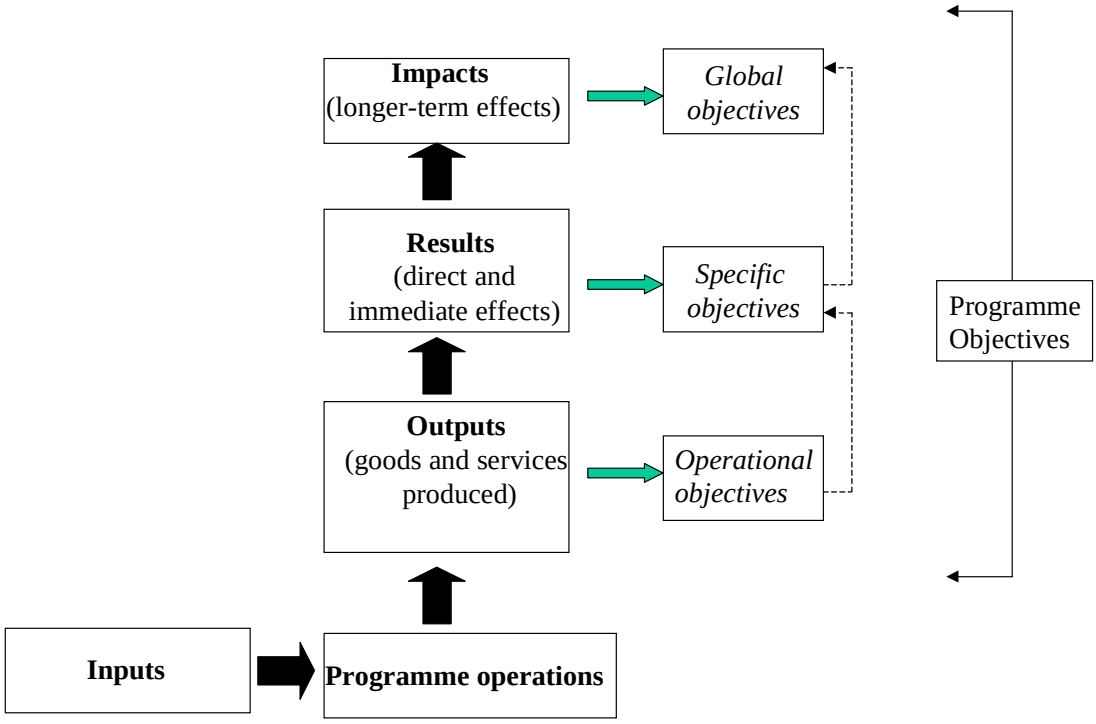
In this paper we address one of the issues raised by the Call for Papers: "Are European Commission evaluation practices giving birth to a European standardisation?" We are mainly concerned with the approach adopted by the European Commission (EC) for the evaluation of the impact of Structural Funds. The EC evaluation guidelines largely ignore the counterfactual methods that the social science community has produced to deal with issues of causal attribution. Counterfactual analysis has become the standard approach for most research institutions and international organizations over the last two decades, with the notable exception of the EC. We offer two main arguments to support the claim that the EC standard approach cannot deal satisfactorily with the estimation of impacts. First, EC evaluation guidelines widely recommend the use of impact indicators: *we contend that indicators alone do not identify nor estimate any impact in a meaningful way. Only a properly conducted counterfactual analysis allows the quantification of impacts, provided that suitable data are available and some (often stringent) conditions are met. Second, we argue that the emphasis on indicators is a symptom of an overriding concern with* accountability for progress toward objectives*, which is different than estimation of causal impacts. We make the case for a partial shift of attention, away from measuring progress toward objectives, and in favour of learning "what works"— that is, gathering evidence on whether the Structural Funds do produce the changes they hold as objectives.*

Introduction

The Conference's Call for Papers includes an important question: "Are European Commission evaluation practices giving birth to a European standardisation?" In this paper we address this question with reference to a specific, although crucial, area of evaluation: the estimation of the impacts of public policies in general and of Structural Funds in particular. To be sure, evaluating impacts represents only one of the many possible tasks facing evaluators. However, it is an essential part of any effort to understand to which extent the Funds are "money well spent".

The evaluation approach sponsored by the European Commission over the past 20 years is aptly symbolized by the logical framework shown in Figure 1. The logical framework consists mainly of a "result-chain", linking inputs to outputs to results to impacts, and a hierarchy of objectives that runs parallel to the result-chain. The arrows in the graph should make explicit how the program or policy is expected to bring about the desired changes (the *theory of change*). Having a well-specified logic model does help in clarifying what effects to focus on, and once estimates of impacts were available, it would help in understanding *why* the program worked – or didn't.

Figure 1. The standard EC logical framework for the evaluation of Structural Funds[1]



Our criticism moves from the realization that the above logical framework is (mis)used to support a mechanical interpretation of the result-chain, one that treats the assessments of outputs, results *and* impacts as similar cognitive tasks. We consider this a major shortcoming of the whole approach to evaluation of the Structural Funds. According to this approach, impacts and effects should be measured the same way outputs are, *using indicators.* We start from this point to develop our argument, then move on to some related issues, ending with some suggestions for a partially different focus for the evaluation of the Structural Funds.

Indicators vs. counterfactuals

The assertion that impacts can be estimated directly by indicators is ubiquitous in the EC evaluation guidelines. Examples abound:

> *"The ultimate objective of Structural Funds and Cohesion Fund assistance is a certain impact, measured as far as possible by impact indicators"* [2]
>
> *"Indicators, for both the policy fields and for measures, are defined according to the "logical framework of intervention" as follows: input indicators (financial), physical output indicators ("volume" of what is produced by the operations), outcome indicators (direct and immediate effects of the action) and impact indicators (medium or long-term effects)".*[3]

[1] Reproduced from : DG-Regio, "*Indicative Guidelines on Evaluation Methods: Monitoring and Evaluation Indicators*", Working Document No. 2, August 2006
[2] *Ibidem*, page 10.
[3] DG Employment, "*Guidelines for systems of monitoring and evaluation of ESF assistance in the period 2000-2006*", 1999.

The very definition of "indicator" frequently encountered contains a direct reference to the concept of "effect":

> *"An indicator can be defined as the measurement of an objective to be met, a resource mobilised,* an effect obtained*, a gauge of quality, a context variable"* [4]

This definition leaves little doubt: allegedly the effects of policies can be observed and hence measured using indicators. We argue the opposite, namely that impacts and effects cannot be meaningfully defined, let alone measured, by indicators. While simple descriptions of observed quantities are suited for quantifying outputs, they are *not* for evaluating effects and impacts. We hold the view that effects (and impacts) should be *defined* as differences between an observed outcome and the outcome that would have been observed, for the same individuals/firms/areas, had the intervention not taken place (a.k.a. the counterfactual situation); and that the estimation of impacts (and effects) consists essentially in the recovery of plausible counterfactuals from the available data. Such recovery involves much more than indicators. It requires what is called an "identification strategy".

Harvard sociologist Chris Winship and his colleague Steven Morgan outline rather clearly the importance of counterfactual analysis:

> *"Simple cause-and-effect questions are the motivation for much empirical work in the social sciences, even though definitive answers to cause-and-effect questions may not always be possible to formulate...In the past three decades, a counterfactual model of causality has been developed, and a unified framework for the prosecution of causal questions is now available"* [5]

What does keep the EC from adopting a counterfactual approach, as does a significant majority of the social science research community, together with the OECD, the World Bank and several European research institutions? [6] EC methodological guidelines hardly ever mention the word *counterfactual.* Instead, they constantly invoke *impact indicators*.

The very term "impact indicator" is conceptually ambiguous: the term "outcome indicator" should be used instead, to indicate on which dimension(s) the impact has to be sought by the program and estimated by the evaluator. Impact is the change in the outcome caused by the intervention, a change relative to the counterfactual situation, not to the situation observed before the intervention. An indicator can *describe the change* in the outcome, but in no way it can attribute a causal interpretation to the observed change.

Effects vs. Impacts

A source of confusion in EC methodological documents is the artificial distinction made between "effect" and "impact". EC documents are keen on calling "impacts" only the effects that take place in the long-run, while calling "results" the "immediate and direct" effects (as seen in Figure 1).

> *"Impact indicators refer to the consequences of the programme beyond the immediate effects. Two concepts of impact can be defined: Specific impacts are those effects occurring after a certain lapse of time but which are, nonetheless, directly linked to the action taken and the direct beneficiaries. Global impacts are longer-term effects affecting a wider population."* [7]

[4] DG Regio, *"Indicative Guidelines on Evaluation Methods: Monitoring and Evaluation Indicators*", Working Document No. 2, August 2006, page 5.
[5] Winship C. and S. Morgan, *Counterfactuals and Causal Inference*, Cambridge University Press, 2007, page 3.
[6] For example IFS (UK), IZA (Germany), IFAU (Sweden) and CREST-INSEE (France).
[7] DG Regio, *"Indicative Guidelines on Evaluation Methods: Monitoring and Evaluation Indicators*", Working Document No. 2, August 2006, page 6

The social science literature tends to treat the two terms interchangeably, as synonyms. There is no meaningful distinction between the two terms when dealing with causality. We believe that the crucial distinction should be between *effects* ("changes that can be attributed to a cause") and *accomplishments* (activities, outputs, "things done", progress made toward a target). The first must be *inferred from* data, the second can be *described with* data. Whether effects take place in the short or long run, whether they refer to beneficiaries or to less proximate actors, the issues involved in their estimation remain substantially the same.

The widely popular distinction between "gross" and "net" effects is another source of confusion. One often encounters statements like:

> *"Subsequently, the net effect of the programme can be estimated by subtracting deadweight, substitution and displacement effects from the gross effect."*

Gross effects are just *observed changes* in a given indicator between two points in time. Let's say an intervention takes place between these two points. The effect of such intervention is the difference between what is observed and what would have been observed had the intervention not taken place (the counterfactual situation). The EC manuals prefer the concept of "deadweight effect", defined as:

> *"Deadweight effect: Change in the situation of the beneficiary that would have occurred even without the public funding."*

Thus, the "deadweight effect" is nothing else than the counterfactual, *i.e.* what would have happened without the intervention. Why then give it a different name, why call it an "effect" when it only denotes a "lack of effect" on the part of the intervention which is being evaluated? We offer two complementary explanations. First, the practice derives from the *bad* habit we just discussed, that of calling "gross effects" what are really "observed changes": since the total is (incorrectly) called an effect, its constituent parts will all need to be "effects" too; and "deadweight" seems an apt name for an "effect" that one would not like to see. The second explanation is that deadweight is seen as a special case of counterfactual, one in which beneficiaries receive public resources to change their behaviors but instead behave as they would done anyway. Here is our punch line: *deadweight is the name given to the counterfactual when the Structural Funds have paid for it*.

Accountability vs. learning

The point just made offers some insight on why the EC methodological guidelines are conceptually so distant from counterfactual analysis. The heart of the matter is the overwhelming importance that *accountability* has taken in the design of EU evaluation. Disguised under the language of impacts and effects, the *EC is pursuing a different evaluation question*, which is *not* a causal question.

"Counterfactual evaluators" are motivated primarily by the question "*what works?*", The EC evaluation apparatus is largely focused on the question *"what did the Structural Funds produced with their resources*?" and more specifically "*what progress did the Structural Funds make toward their (measurable) objectives*" Very simply, *progress* is *treated* like it is an impact, but it is not. Establishing progress toward objectives is essentially a descriptive task, albeit an important one, but fundamentally different from the estimation of impacts, which requires establishing causality and drawing causal inferences from data.

The *accountability for progress toward objectives* (APTO) clearly dominates the discourse (and the rhetoric) on the measurement of impacts in the Structural Funds. Here are two examples:

> *"A priority of the new approach to evaluation in the 2007-2013 period is to assess the contribution of cohesion policy to the achievement of the Lisbon goals and to make that contribution more visible."* [8]

> *"The effectiveness of the Structural and Cohesion Fund assistance, which involves the analysis of outputs, results and impacts and the assessment of their compliance with the expected objectives"* [9]

The very idea of assessing *compliance* of impacts with the expected objectives has meaning only in a strict accountability perspective.

Progress toward objectives is actually defined in two different ways. Progress from a *baseline*, and progress toward pre-defined *targets*. For example:

> *"There is a specific focus on* quantification of impact *in the rural development regulation, particularly in relation to the* baseline *situation."* [10]

> *"Indicators need* quantified targets *because otherwise the extent to which the original objectives are being met cannot be measured."*[11]

In neither case progress is an impact, for the disarmingly simple reason that progress can occur without there being any impact of the policy, while a lack of progress can mask an actual impact of the policy. The two maxims "*Things might have improved anyway"* and "*Things could have gone even worse"* say it all.

When progress is defined with respect to a baseline, the observed change from the baseline is often (and riskily) given a causal interpretation. Before-after comparisons are the weakest form of causal inference.

> *"Administering a before-and-after evaluation design is relatively easy, but causal inference tends to be quite weak. There is always the possibility that something else besides the programme may account for all or part of the observed change over time"* [12]

Baselines must be observed through actual data, and are naturally defined with respect to indicators that are "external" to the intervention being evaluated. On the other hand, targets can be simply made up, and can be "internal" – that is, be defined as "things to be done". Any reference to causality is obviously lost when progress is measured in terms of "things done". Output indicators, or the more ambiguous "result indicators", are used in this case. But targets are set with respect to external outcomes, invariably EC documents refer to impact indicators as the empirical strategy to assess progress toward these targets. The following quotation in emblematic:

> *"Impact indicators should play a decisive role at certain stages of the programming cycle: the ex ante quantification of impacts is an instrument for the strategic orientation of a programme; and only the impacts of a programme found ex post allow a final judgement to be made on the success or failure of a programme*."[13]

---

[8] DG Regio, "*Indicative Guidelines on Evaluation Methods: Evaluation during the Programming Period*", Working Document No. 5, April 2007, page 10.
[9]. *ibidem,* page 10
[10] Directorate General for Agriculture and Rural Development, *Handbook on Common Monitoring and Evaluation Framework* – Guidance document, September 2006, page 14.
[11] DG Regio, *"Indicative Guidelines on Evaluation Methods: Monitoring and Evaluation Indicators*", Working Document No. 2, August 2006, page 6
[12] Nagarajan, N. e Vanheukelen, M *Evaluating EU Expenditure Programmes: A guide to intermediate and ex post evaluation,* 1999.
[13] DG Regio, *"Indicative Guidelines on Evaluation Methods: Monitoring and Evaluation Indicators*", Working Document No. 2, August 2006, page 9.

It is clear that what is really meant by "*ex ante quantification of impacts"* is setting a target with respect to each indicator. More problematic is the second statement, in which "impacts found ex-post" allow a final judgement to be made on the success or failure of a programme. Most likely what is referred to as "impact" is in reality progress made toward a target.

> *"The Regulation encourages the quantification of objectives. This is not always possible. Either the Member State is in a position to announce a quantified objective (e.g. to reduce by half long-term unemployment), or it announces quantified tendencies (e.g. to reduce the level of long term unemployment). In the latter case, the indicators that make it possible to monitor the objectives and the development of the context are identified in the programme as part of a quantified baseline."*[14]

A distinction is made between quantified objectives (*i.e.* targets) and quantified tendencies (*i.e.* direction of change). The role of indicators in the latter case is almost incomprehensible; which is not a rare case in these manuals, in which an "air of magic" surrounds the cognitive potential of indicators.

And yet, causality looms

Occasionally, one finds statements claiming that establishing causality is an important part of the mission of the EC evaluation. For example:

> *"Because a causal analysis of effects is the most important question in ex post evaluations, the method used to analyse these causal relations is the priority in this type of evaluation."*[15]

Sometimes counterfactuals pop up in odd places, as in the following example, in which they would be *identifiable* using indicators, which is patently wrong:

> *"Indicators can be used to identify what would have happened in the absence of the initiative, policy or legislation (the counterfactual)."* [16]

The idea of counterfactual, often without being mentioned by name, surfaces (more seriously) in relation to the concept of *additionality*. One example among the many available:

> *"An assessment of additionality involves establishing a causal relationship between Structural Fund interventions, projects and employment effects ('attribution'). The key question to be asked is: what would have happened to the project if Structural Fund assistance had not been available?"* [17]

Despite these occasional forays into causality, the *accountability* perspective is the outright winner in the EC-led evaluation enterprise. Prevailing is the need to *show progress toward objectives*; while only marginal attention is paid to make sure that what is patently *not due* to the policy is excluded from what is counted as progress (*e.g.*, the deadweight). Still succumbing in the evaluation of Structural Funds is the desire to *understand whether a specific policy tool* is able to produce the desired effects, and how this effect depends on the characteristics of the beneficiaries, and which underlying mechanisms can explain its presence (or absence).

[14] DG Employment, "*Guidelines for systems of monitoring and evaluation of ESF assistance in the period 2000-2006*", 1999
[15] DG Regio, "The *Evaluation of Socio-Economic Development: The Guide*", page 71.
[16] *ibidem*, page 130.
[17] DG Regio, "*Measuring Structural Funds Employment Effects*", Working Document No. 6, March 2007, page 10

But things might be changing, and counterfactuals might find their way to Brussels, after all. At the same time, we recognize that *indicators will always be with us*, with targets and baselines, because the need for APTO will not go away.

Some suggestions

The two perspectives, *accountability for progress* and *learning what works*, can indeed coexist and should both be pursued, as long as it is *recognized that they require largely different analytic tools, time frames and levels of aggregation.* In the accountability perspective, *aggregation* and a *fixed time frame* are essential features of the analysis. In the learning what works perspective, what drives the evaluation design is the *identification strategy* chosen to recover the counterfactual. *Progress toward objectives* will remain the major concern of EC evaluation effort, particularly at aggregate levels of policy. We only advocate the recognition of the fact that this task has little to do with impacts in any meaningful way. We only advocate a partial shift of attention, not the abandonment of current practices. We also are fully aware that counterfactual analysis does not always guarantee a plausible identification of impacts, particularly when it deals with complex programs, not a rare event in the Structural Funds.

The application of counterfactual analysis would also need a repository of its findings, in order to favour dissemination and eventually utilization. The EC Evaluation Units, besides coordinating the production of evaluation for accountability, could become such a repository, a clearinghouse of findings on what works and what doesn't in the Structural Funds.[18]

Finally, it must be kept in mind that the counterfactual approach is not the only paradigm to tackle the issue of causality. The *successionist* view of causation – from whom the counterfactual approach derives – is vehemently opposed by the Realists, who are proponents of a *generative* view of causation, based on the idea of *discovery of the underlying mechanisms generating the effects*.[19] The controversy between these two epistemologically distant positions is crucial for the future of evaluation in Europe, and it has some bearing on the issue at hand. If it moves partially away from accountability, should the EC focus primarily on the "what works?" question or jump immediately to answer the far more challenging "why it works?" question, as the Realists would advocate? Our first suggestion is to keep in mind that answering to "why it works?" can be crucially helped by knowing *in which direction* it worked – whether an effect was actually there or not. A second crucial reason for keeping focused on a "did it work?" question, is the fact that "it" – that is, the Structural Funds – implies a massive use of taxpayer's money. An exclusive attention on the "why" question would risk missing the main point – that is, is "it" money well spent ?

Bibliography

DG Agriculture and Rural Development, *Handbook on Common Monitoring and Evaluation Framework* – Guidance document, 2006.

DG Employment, "*Guidelines for systems of monitoring and evaluation of ESF assistance in the period 2000-2006*", 1999

DG Regio, "*Evaluation of Socio-Economic Development: The Guide*", 2003

DG Regio, "*Indicative Guidelines on Evaluation Methods: Evaluation during the Programming Period*", WD No. 5, 2007

DG Regio, "*Measuring Structural Funds Employment Effects*", WD No. 6, 2007

Nagarajan, N. e Vanheukelen, M, *Evaluating EU Expenditure Programmes,* EC document, 1999.

Pawson R. and N. Tilley, *Realistic evaluation,* Sage, London, 1997

Winship C.,S. Morgan, *Counterfactuals and Causal inference*, CUP, 2007

[18] A good example of such an effort in the What Works Clearinghouse of the US Department of Education (http://ies.ed.gov/ncee/wwc/).

[19] Pawson R. and N. Tilley, *Realistic Evaluation,* Sage, London, 1997.