The Path to the Prize:
Two Perspectives on Counter-Factual Evaluation

Michael Wiseman
George Washington Institute of Public Policy
The George Washington University
Washington, D.C., USA

15 November 2012

Paper prepared for the 8th Evaluation Conference
Ministry of Regional Development
Polish Agency for Enterprise Development
Evaluation in the system of public policies
12-13 November 2012
Warsaw, Poland

Draft: 30 June 2013

Abstract

The European Commission Directorates-General responsible for the coordination of Cohesion policy are engaged in promotion of more rigorous, "counterfactual" evaluation techniques. Attention to credible counterfactuals in evaluation is a step in improving the quality of evidence produced, but it is important not to lose sight of other objectives of evaluation policy, including fostering and nurturing evaluation–orientation in government and increasing utility of evaluation results. This paper reviews concepts of internal and external validity and moves on to what gives good counterfactual evaluation external *utility*. It proposes establishing a system of prizes for good evaluation plans that measure up to criteria to be developed collectively by the Managing Authorities of Member States. Meeting the need for external utility of Member State evaluation efforts is served, it is argued, by thinking of Managing Authorities as members of a club in which the dues are planning and execution of credible and generally useful evaluations and the benefits are the gain from having access to such results and the coordination of research effort. Polish authorities should continue to lead in evaluation planning and collaborative policy research.

The Path to the Prize:
Two Perspectives on Counter-Factual Evaluation

Michael Wiseman*

The Polish Ministry of Regional Development is widely recognized as being exceptionally active in using funding from the European Union to promote Cohesion Policy. This reputation derives in part from the Ministry's efforts at program evaluation. Given financial and other developments, the demand for evaluation will grow. These concerns are not confined to systems funded through the Structural Funds and Cohesion Fund. Indeed, it is unlikely that an agency not committed to evaluation generally can design or deliver successful assessments of the impacts of the national activities the European Union supports. As a result, progress in improving evaluation of European Union (EU)-related activities is contingent upon and interrelated with progress in development of evaluation-oriented culture and skills within Managing Authorities. It also requires growth in appreciation for evaluation by political leadership.

This paper develops two perspectives on evaluation development. One starts with Managing Authorities. The second involves cross-national engagement. I argue that certain problems with evaluation lead to too little being produced in a context like that of the European Union or, for that matter, any political entity in which budgeting for evaluation is devolved. I end up with a suggestion for a means of drawing attention to what is needed for promoting improved evaluation in the context of decentralized decision-making.

My thoughts are rooted in three sources. One is my one-time mentor, Aaron Wildavsky who years ago in Berkeley taught this (then) young professor of economics the importance of thinking about program management as an essential element of policy analysis (see Wildavsky 1987). Professor Wildavsky considered the central problem of public management to be development of an agency culture that continuously fosters effort at improvement and rewards those personnel who contribute to achieving this end. Evaluation is an obsession of such a culture. The second source is long experience working with American states and observing the failures of what one Supreme Court justice famously termed (in 1932) the "laboratories of democracy." There is much to be learned in the U.S. and elsewhere from European experience with collaborative evaluation. The third is time working over the past 18 months with analysts in the Office of the Directorate-General for Employment, Social Affairs and Inclusion to develop guidelines for counter-factual impact evaluation (CIE) of activities subsidized through the European Social Fund. It is in that third context that I have learned of and come to appreciate the evaluation efforts of Poland and other EU Member States.

* Research Professor of Public Policy, Public Administration, and Economics, The George Washington University, Washington, D.C. 22202 USA. This paper was prepared for the Ministry of Regional Development in connection with the MRD 8th Evaluation Conference, "Evaluation in the system of public policies," Warsaw, Poland, 12-13 November 2012. I have benefitted from discussions of evaluation promotion strategy with Veronica Gaffney and Herta Schönhofer. Like the mistakes, the opinions expressed here are mine and do not represent or reflect the opinion or policy of the agency. The author's email address is WisemanM@gwu.edu.

Evaluation at Home

My argument starts with evaluation as done by Wildavsky-style Managing Authorities as standard policy regardless of funding source.

CIE essentials

Seeking improvement is to look for changes in what an agency does that are cost effective, meaning the benefits exceed the costs. Counter-factual impact evaluation (CIE) is the foundation of benefit cost analysis. CIE assesses the consequences of program introduction or change for outcomes of interest compared to what is accomplished with an alternative. In management, the alternative is commonly business as usual, but it can be another program strategy intended to achieve equivalent ends.

Impact is an unfortunate word for the difference in outcomes between what is recorded for those who are engaged by a program as introduced or modified and what is predicted to have occurred with the alternative. *Effect* would be better. Benefit-cost analysis looks for change that doesn't cost too much to be worthwhile, but benefits and costs can't be estimated without an estimate of what would have occurred in the absence of the initiative.

The prediction, the counter-factual, can be developed in many ways. What is termed the "internal validity" of the evaluation hinges on the credibility of the counter-factual construct. In many situations the most credible counterfactual is developed by random assignment of potential program participants between "treatment" and "control" status. Those in the treatment group are given opportunity to engage in the new activity and the controls are not. Because if random assignment is done correctly no systematic differences exist between treatment and control groups, the outcome for the control becomes a reliable forecast for what would have happened otherwise.

But whether or not assessment is done formally, good managers are always thinking about counterfactuals—"Suppose we change from A to B. Would the difference in outcome be worthwhile?" Thinking about the counterfactual requires two predictions, one for what will happen with no change, i.e. operation *A* retained, the other for the consequence of change. In some circumstances sufficient information may be available from other trials of doing B that the consequences can be predicted—and expected benefits and costs assessed—with sufficient reliability to support action. If this is the case, the information from other sources is said to have *external validity*. If the evidence is suspect or uncertain, experimentation is in order, or at least plans should be laid for assessment of impact once change occurs. This brings us back to CIE.

Three things should be noted about CIE as done by good (in the Wildavsky sense) management. The first has to do with process, the second with external validity, and the third with costs.

Look back at the "Suppose we change from A to B" statement. It has three parts: the change part, the difference part, and the worthwhile part. Change comes first: Innovation changes input or process. Managers are interested in the changes innovation brings about in the character of services or other activities that are the agency's responsibility. Innovations are typically formulated as an ideal, a new model of how something should be done.

This change in how things are done is the domain of process analysis. Process analysis normally has not one but *two* counterfactuals. The first counterfactual is the process as depicted in the plan or ideal for the change under consideration—the model. The other process counterfactual is what is done for the control, often business as usual. Process evaluation looks at both—how close an implementing authority comes to what was intended, and how great is the difference between the treatment process accomplished and the process counterfactual.

Process analysis is sometimes called "monitoring," and indeed since process assessment can hardly be done without monitoring, that can be appropriate. But in my experience much of what is called monitoring is not process evaluation at all, because neither of the process counterfactuals—the ideal or the control—is well identified. This prevents measurement. Process evaluation requires measuring in some way the difference between what happens to the treatment and control groups as a result of an innovation AND measuring in some way the difference between what happens to the treatment group and the intent of management.

Process analysis is essential to good management, a vital precursor to impact evaluation, and the source of the "costs" side of benefit-cost assessment—the "worthwhile" part of the "suppose we change" statement. This connection with impact arises because process analysis is essential to confirming just what happened as an innovation occurred. If we know what really happened to process, we have clues to what produced whatever impact is observed. If we end up with no effect, process analysis can tell us whether the failure was one of theory (the idea was bad) or delivery (we did not manage to deliver the program planners intended).

I make a side excursion here to "logic models". A logic model is the theory that connects inputs to output and explains why some change in activity is expected to have a particular impact on outcomes of interest. Logic models are an essential part of planning for innovation, and they can guide attention to essential features to be monitored in process evaluation. Combined with evidence on efficacy of the various elements in the presumed chain of connection from program to outcome, logic models can justify policy changes; if the evidence is strong enough it may not be necessary to add additional impact assessment. But no matter how convincing the logic, process monitoring is essential to understanding, and assessing, implementation. As important as clear statement of the logic of change can be, logic models do not substitute for counterfactual process or impact evaluation.

Now, back to Wildavsky's good manager: I have argued that the first feature of good-manager program evaluation is attention to process. The second feature concerns external validity. External validity turns not on the credibility of the counterfactual (that's *internal* validity) but on the relevance of results from one evaluation to situations beyond that in which the impact evaluation occurs. Is it conceivable that a similar innovation, introduced elsewhere in time or place, would have similar impact? My dictionary says *art* involves skills "acquired by experience, study, or observation." I think assessing external validity is an art. National agencies are in good position to assess the validity of what is learned in one place to another within context of a common cultural, legal, and economic environment. Other things equal, we feel more comfortable with evaluation done "nearby", where "nearby" means "in a place like this". Good managers can assess closeness, just what "like" entails.

The third point is that evaluation is costly. I know of no Managing Authorities with unlimited budgets. One simply cannot test all possible innovations or gauge the impact of all features of programs in place at one time. Moreover, within any real-world agency multiplication of evaluations diminishes the attention paid any particular effort and raises the risk of failure. The consequence is that choice of evaluation activity is important and involves trade-offs. Such is life.

Eyes on the Prize

All this considered, suppose key members of the National Parliament have studied Wildavsky and have decided to create incentives for evaluation of current programs and generating and testing ideas for improvement. The incentive is a prize for an outstanding evaluation plan. When should the prize be given? What should the CIE prize committee look for?

Here's a big point: The public interest is likely best served if the prizes are done on the basis of the plan. This does not mean that execution is not important, indeed critical, and the improvement-oriented agency will see evaluation of evaluations as part of the culture of improvement. But plans are the foundation of accomplishment. Postponing contemplation of evaluation until after innovation fielding dramatically lowers the likelihood that much will be learned.

The criteria should themselves be the product of consultation. This helps to ensure a sense of ownership in the competition; it makes the contestants stakeholders in the effort. However expressed, it is likely that the criteria for prize-giving will encompass most of the following. The elements are expressed here for the case in which the proposal calls for an innovation. But it should not be forgotten that in some cases the innovation can be closing a program instead of modifying or initiating one.

- Logic of the intervention

Does the theory underlying the intervention, the change to be evaluated make sense? Is the causal model supported by other assessments?

- Potential

Is there good reason to believe the benefits of the evaluation will exceed the costs?

- Evaluation Methodology

Is the plan feasible? Will the results have internal validity? Is the forecast for outcomes in the absence of the innovation credible?

- Process analysis

Does the evaluation plan include comparison of what the innovation produces both to the model and to process as experienced for the control?

- External utility

Will results of the evaluation have value as input to future decisions, including further deployment of the intervention? External validity is part, but not all, of what makes an evaluation useful. An evaluation is useful when the results have potential to improve subsequent public decision-making.

It is difficult to assign weights to these elements because they are not independent. For example, "potential" is dependent up both the quality of methodology and the external utility of the knowledge the evaluation is expected to produce. The external validity of results of an evaluation of an intervention lacking logical coherence turns on whether or not the evaluation itself uncovers an unsuspected logic. Risky business, that! The list does, however, suggest an order for review. The final consideration should be external utility: Will what a project promises to deliver indeed be useful?

Suppose the Ministry of Regional Development or indeed even broader collections of government agencies were to hold a competition among its constituent units or, for that matter, across its employees for innovation/ evaluation ideas—after the criteria have been developed collectively. Who knows what might turn up? Note that the bar is high in this competition: Not only must contestants have an interesting idea, but they must have at least the outline of a plan for implementation AND evaluation.

Government use of prizes as incentives is not that far-fetched. The British Parliament famously offered the Longitude Prize in 1714 for discovery of a practical means of measuring longitude. They are central to the Obama Administration's "Strategy for American Innovation."[1] (see www.challenge.gov). What is novel about this prizes suggestion is its focus on evaluation planning.

The Big Problems

Thinking about external utility brings us to evaluation in EU context. Emphasis on impact evaluation is growing across most areas of Structural Funds and Cohesion Fund activity, especially as planning is underway for the 2014-2020 funding cycle. "Guidance" multiplies as the various Directorates-General struggle to make the case that the Funds have effects beyond transfer of resources. What's different when we scale up to thinking about evaluation across member states?

There are two big problems: The first is incentives, and so is the second.

The first incentives problem arises because to some resources devoted to satisfying Fund requirements for evaluation appear to be deadweight loss, nothing more than a ticket that must be purchased to gain and sustain access to the Funds. In the Wildavsky world, calling for evaluation would be superfluous because the supported agencies would already be doing it and resources received from Brussels would be treated no differently from those collected by taxes from the good citizens of, say, Nowy Tomyśl. In reality, as Wildavsky recognized, developing and sustain a culture of improvement and evaluation is not easy, agencies often don't manage it, and politicians rarely support it. Even when the will exists, Managing Authorities without long

[1] The White House Office of Management and Budget provides useful overview in http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-11.pdf

experience in evaluation may simply lack the capacity to do it well. Nurturing such capacity can be difficult business, for both political and tactical reasons. It is hard to do from Warsaw, let alone from Brussels (or Washington).

The second incentives problem is more complicated. It involves spillovers, benefits gained by one agency or country as a result of the activities of another. Economists have long recognized that spillovers can lead to market "failure" in the sense that too little of a good is produced when spillovers are good and too much is produced when spillovers are bad. Evaluation may be a useful means of regulation and ex post justification for grants. But the external utility of evaluation may be the source of greatest benefit from evaluation effort. If an evaluation has convincing external validity and what is learned is useful, then this knowledge is a spillover. If the external utility is not considered when evaluation plans are drawn, the benefits will be understated. If costs fall exclusively on the evaluating agency, then some evaluations will be avoided or conducted on a less-than-optimal scale even if, when viewed from a perspective that recognizes the value of the knowledge gained, they should be undertaken.

Put differently, there may be more "stakeholders" for some evaluations than local authorities recognize. The classic solution to the externalities problem is to "internalize" them by assigning evaluation to a level of government that includes all stakeholders. Evaluation may be done locally, but if spillovers are present, evaluation subsidy, drawn from the entire community of beneficiaries, is justified. The object of the subsidies is to encourage evaluators to recognize the benefits of producing knowledge that has both external benefits and external utility. Like most other problems in public policy, the design of such subsidies is not easy. This is why God gave us economists (in case you wondered).

Externalities are certainly present in evaluation of local delivery of national policy, and at least in principal the spillovers from such evaluations can be internalized by a national government that bears some or all of the local evaluation costs. But EU agencies lack (often for good reason) many of the instruments of policy available to national government. Nevertheless, if effort is not made to promote attention to external benefit in designing and evaluating initiatives funded through the Funds, EU citizens in general will lose. More than guidance is required. Thought should be given to prizes and clubs.

"I'm from Brussels, and I'm Here to Award You"

Consider prizes first.

Suppose the Directorate-General for Regional and Urban Policy was to initiate a competition for plans for evaluation of activities conducted using DG-REGIO funds. Setting aside what the prizes should be, how should the criteria list be altered when applied internationally?

Again, this should be a matter of collective decision. On first consideration, it seems likely that the same list would apply, but emphasis might change. The important thing is to pay even more attention to external validity and utility. It helps to shift from the perspective of an evaluation producer to that of consumer. Suppose you are in Member State A and you are looking at an evaluation produced in Member State B. How does your perspective change from the way in which you might look at an evaluation done in your own country?

Certainly some things don't change at all. Internal validity is still an essential element of the evaluation scheme, and you want to know the methodology. What may be more important are the details of process. You need to know what was delivered, not in theory but on the ground. You are interested in what might be called the "production function" for what was delivered. And you certainly want to know how much of what was intended was actually achieved.

The treatment is, however, only one side of the story. It is important to understand the control. What occurred in the control situation? How comparable is this to the base in your own country. It is difference in control that makes validity across borders so problematic. What difference does it make to you if the innovation introduced in some other country increased job retention within group X if what would have happened to group X in the absence of the innovation in that other country is completely different from policy as practiced in your own? Surely the art of determining external utility requires process, as well as impact, detail. At the EU level, prize evaluation plans should include provision for collection of that information.

Just what is needed to enhance external utility, and how to balance the benefits of the effort to collect and provide such data against the costs, likely varies with the type of innovation undertaken. My point is that when discussing evaluations within a national environment, the nature and generality of the control experience is often understood and can be presumed when a policy maker ponders whether results in one place/ time carry over to another. Extended across borders, the peril of such presumption grows.

A prize may serve to draw attention to the need for process detail and attention to the gains to outsiders from doing evaluation well, but it's hardly clear that the spotlight of competition will be adequate to bring real growth in the evidence base for policy over the long run, which in our case means the stretch from 2014 to 2020. More is needed. Again, it helps to change perspective.

The All-Europe Evaluation Club

The argument to this point has been cast vertically, beginning with evaluations done within nations and moving upward to evaluation of innovation funded at the EU level. Return to the horizontal level, the relations between Country A and Country B. I have argued that failure to acknowledge the benefits to Country B of evaluation-based knowledge about an innovation's impact causes Country A to under-invest in evaluation, and vice-versa. Moreover, not only does Country A under-invest, but it likely fails to collect or report the kinds of information that would give the results "legs" for travel across the border.

There would appear to be gains from trade here. Suppose our two countries formed an evaluation union, a "club" if you will. Membership in this club comes with an obligation and a benefit. The obligation is to acknowledge the interests of the other member in both determining what will be evaluated and the kinds of detail that will be collected. The benefit is that the partner does the same. The easiest way to ensure that those interests of the partner-stakeholder are incorporated in evaluation planning is to involve them directly. Does this sound like a sort of "Open Method of Coordination" between A and B? You bet!

The A&B Evaluation Club has an additional positive consequence. As pointed out earlier, national capacity for evaluation is limited, no matter how great the enthusiasm for Wildavsky-style governance. With coordination, countries can focus on one set of issues and rely on partner work for others. But to gain such economies, the evaluation commitment must be serious on both sides of the exchange. Open coordination and frequent communication is essential. President Reagan's comment about the Strategic Arms Reduction Treaty seems appropriate. To paraphrase, "trust your partner's evaluation, but verify."

We are a long way from an All-Europe Evaluation Club, to be sure. And looking to lateral coordination in program evaluation among countries within the Union is several steps beyond the current challenge of developing evaluation cultures within Managing Authorities. But a leader like Poland should begin to think this way, possibly by forming bilateral coordinating evaluation partnerships in particular policy areas. This ground-up approach could be a very useful complement to efforts emanating from Brussels to promote CIE.

Summary

Much attention is currently devoted to encouraging counter-factual evaluation of Member State initiatives subsidized from the Structural Funds and the Cohesion Fund. While evaluation is central to good governance, it is important longer-term objective to foster and support within Managing Authorities a general culture that continuously fosters effort at improvement and rewards those personnel who contribute to achieving this end. Moving toward this objective is in part a matter of incentives. One approach to incentives is to establish prizes for plans for initiative assessment that recognize the importance of both process and impact analysis and the role of evaluation in the collaborative European effort to build the evidence base for cohesion and regional development policy. To this end it may be useful to think of Managing Authorities as a sort of club for which the dues are delivery of externally valid and useful innovation assessments. The rigor of these evaluations might best be ensured by the development of intra-union exchange along the lines originally proposed for the Open Method of Coordination. Those countries most engaged in evaluation development—Poland, for example—might take the lead.

The author has benefited from conversations on these topics with Veronica Gaffey, Herta Schönhofer, and Alberto Martini. None should be held responsible.

References

Rossi, Peter H., Mark W. Lipsey, and Howard E. Freeman. 2004. *Evaluation: A Systematic Approach*. 7th ed. Thousand Oaks, California, USA: SAGE Publications.

Wildavsky, Aaron. 1987. *Speaking Truth to Power: The Art and Craft of Policy Analysis*. Piscataway, New Jersey, USA: Transaction Publishers.