Invited Article

What Doesn't Work? Three Failures, Many Answers

Evaluation
16(4) 371–387

DOI: 10.1177/1356389010381914
http://evi.sagepub.com
SAGE

Nicoletta Stame
Universita di Roma 'La Sapienza', Italy

Abstract
Current debates on impact evaluation have addressed the question 'what works and what doesn't?' mainly focussing on methodology failures when providing evidence of impact. In order to answer that question, this article contrasts different approaches to evaluation in terms of the way they address different kinds of possible failures. First, there is more to be debated than simply methodological failures: there are also programme theory failures and implementations failures. Moreover, not all methodological failures are a simple matter of selection bias. Second, the article reviews issues that have recently been raised within different approaches relative to each failure. For programme theory failure, it is a matter of complexity and providing rival explanations; for implementation failure: how to use guidelines, and how to take context into account; and for methodology failure: how to move from internal to external validity, and to syntheses, within the framework of 'situational responsiveness'. All these issues disclose a terrain for potential exchange between the protagonists of different approaches to impact evaluation.

Keywords
external validity; implementation failure; methodology failure; ruling out rival explanations; theory failure

Introduction

The recent debate on impact assessment starts from a recurrent problem in evaluation: 'What works? Nothing works!' Howard White's intervention (White, 2010) in *Evaluation* hinges around what he calls a 'misunderstanding' between advocates of two different definitions of impact evaluation (IE). The first is based on the logic of counterfactual analysis: 'the difference in the indicator of interest (Y) with the intervention (Y1) and without the intervention (Y2)'. The second, provided by the OECD-DAC (2002) glossary's definition of impact, is based on what we could call the evaluator's modus operandi: 'positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended'.

White believes that the two definitions denote two different things; he is only interested in the first one, which he rates to be able to fulfil the most urgent task, namely providing commissioners

Corresponding author:
Nicoletta Stame, Universita di Roma 'La Sapienza', Italy.
Email: nstame@aconet.it

with a clear-cut answer about 'what works and what doesn't and at what cost', based on 'evidence', so that they can take straightforward decisions about funding or discontinuing current programmes. As for the other definition, according to him it provides only vague answers about what works and is therefore of no use to commissioners. Therefore, White leaves it to readers to follow their preferences, while expecting that they should nonetheless follow him ('give ground') in his argumentation, i.e. as to the limitations that he ascribes to their methodological preferences. In this way, he closes the door to any possible dialogue because he does not grant the other definition any legitimacy to provide the answer to the crucial problem.

Undoubtedly, two different definitions of impact exist; but perhaps it has more to do with disciplinary background than with evaluation uses. I contend that the two definitions do not denote two different things, but two different approaches to the same thing: indeed, all evaluators are concerned with what works and what doesn't, and consequently with what causes 'working' or 'not working'. And all evaluators embark on this enterprise because they acknowledge the uncertainty of results.

Indeed, there is widespread awareness that:

- Goals may not be clear.
- There may be a black box between intervention and effects.
- Implementation may be very different from site to site.
- Anticipated effects can be both positive and negative, as can unanticipated effects.
- Commissioners may not use evaluations.

Different approaches have provided answers to the same problems from alternative perspectives, even if there have been encounters and 'contaminations' between the two camps. Perhaps revisiting some milestones along these parallel paths and unexpected encounters may clarify the scope for a debate that the mere opposition between the two definitions appears to rule out.

White is mainly worried by a methodological problem that besieges his 'definition': if there is a selection bias in the experimental group, the effect cannot be ascertained, and any attribution to the intervention may be false; hence it will not be possible to show whether a programme (or at least one having characteristics that allow for a quantitative impact evaluation) indeed works.

But to know whether a program 'works' or 'doesn't work' other aspects are implied beyond the ability to use valid tools to demonstrate the sequence of treatment effect on a given population. Suchman is credited among evaluators for having distinguished between 'implementation failure' and 'programme failure'. In fact, he was interested in identifying the role of intervening variables between the 'programme' activity (independent variable) and the desired effect (dependent variable). Hence Suchman (1969: 16) stated that a programme failure means 'the inability of the program to influence the "causal" variable', while a theory failure means 'the invalidity of the theory linking the "causal" variable to the desired objective'. Weiss (1972: 38) had originally utilized this distinction as a way of dealing with short-term vs long-term effects: 'stated another way, program failure is a failure to achieve proximate goals; theory failure occurs when the achievement of proximate goals does not lead to final desired outcomes'. Later on (1997: 45; 1998: 59), she built her theory-based evaluation approach on these concepts, although using slightly different terms: what Suchman called 'program' Weiss calls 'implementation'; what Suchman called 'theory' she calls 'program theory'. And it is the inter-relationship between implementation theory and programme theory that matters most in her approach.

It is however noteworthy that Suchman's and Weiss's argument had been utilized by Lipsey *et al.* (1985) who added a third failure beyond 'implementation failure' and 'theory failure': 'method

failure', by which he meant statistical power failure. And this *method* failure was to him more relevant than the other two,[1] because it implied that the evaluator might be responsible for *not* being able to show that a programme worked. This is the kind of failure addressed also by White.

Looking backward, I think that while some approaches to evaluation may be more concerned with the methodology of evaluation and others with the theory of the evaluandum, all three failures are – perhaps unconsciously– conceptualized and addressed by all approaches. Understanding how this has happened may provide a fruitful terrain for a dialogue between the two 'definitions' identified by White.

This article aims to achieve the following:

- Demonstrate that the two definitions refer to the same thing.
- Refer the two definitions to a distinction that is more relevant to the field of evaluation : 'goal-oriented' vs 'goal-free' approaches.
- Investigate the main issues in the three failures identified, and how they have been handled by the two approaches to impact, in order to see whether there could be a possible meeting point.

I leave it to another occasion to discuss whether policy-makers are as eager to be given exact answers on impact before deciding their policy, something that White takes for granted, contrary to much evidence.

Two Approaches

Seen in retrospect, and in the light of the issue of 'What works? Nothing works!', the main difference between evaluation approaches seems to be the attitude toward goals. Borrowing from Scriven (1981) we may distinguish between 'goal-oriented' and 'goal-free' approaches. The first includes all those who follow an idea of 'olympic rationality' or 'synoptic rationality', according to which a prerequisite for 'working' is that goals are clear, implementation follows precise protocols, and evaluation can show a positive impact (net effect) on the desired goals. The second includes those that are based on the respective opposite idea of 'bounded rationality' (Simon, 1947) or 'process rationality' (Lindblom, 1968), according to which programmes can work through mechanisms that are enacted by situated actors exploiting favourable opportunities, hence results are strictly linked to process; and evaluation can show how this happens.

Behind these two approaches lies a different conception of what is a policy: goal-oriented approaches – using a medical metaphor – consider policies as 'treatments' (single actions) that are administered to a 'sick' subject (or 'target') in order to achieve recovery; goal-free approaches consider policies as a set of actions in a system of relationships, in order to tackle a problem.[2]

The two main approaches can be seen as addressing the common problems in different ways:

- Regarding goals: from mapping programme objectives (goal-oriented), to redefining criteria and goals in the course of implementation (goal-free).
- Regarding implementation: from assessing implementation performance against protocols (goal-oriented), to inquiring into 'evolutionary implementation' which implies the renegotiation of goals (goal-free).
- Regarding impact: from sticking to anticipated effects only (goal-oriented), to openness towards any effects (goal-free).
- Regarding causation: from accepting black box (goal-oriented), to looking for mechanisms in contexts (goal-free).

The alternatives implicit in these differences, as old as evaluation itself, have been recently challenged by the emergence of new issues that have dealt with all three failures, and that have been addressed inside both camps, as we will see in the next paragraphs. There is more to be debated than simply methodological failures: not all failures are to be understood as methodological failures; and methodological failures are not only a matter of selection bias.

Programme Theory Failure

What does it mean that a programme theory fails? That it is wrong? What is a programme theory about, and what should it include? And how can one distinguish right from wrong?

Programmes are designed to bring about a desired change. Programmes have underlying theories holding that a planned action will produce a (positive) result – a change in line with the goal – for a given population.[3] Theories are a way of understanding the world, of making sense of that which seems complex, disordered, chaotic, and the more so if what happens is not just the spontaneous unfolding of events, but the result of an effort at changing course. (For example, Pawson, 2006, emphasizes that evaluation is concerned with how regularities are altered.)

Programme theories have been mainly represented as logical frameworks, showing a linear *causality* from input to output. It is assumed that programmes are simple: action (a) produces outcome (b); or the programme (independent variable) produces a change in the target population (dependent variable). More important, it is assumed that all that is referred to as the programme (action, effects) can be reduced to something straightforward: one main variable, a single indicator on which the mean effect is calculated, one simple piece of evidence that can be offered to the decision-maker, a single cause that is responsible for a particular effect.

Complexity

This way of conceiving the programme has, however, also been seen as *a factor in programme theory failure*. There are aspects of programmes that cannot be reduced to something simple without not only losing sight of what they are intended for but also what they achieve. Such aspects need to be analysed through an alternative lens that acknowledges complexity instead of rejecting it. In a now classic contribution, Rogers (2008) has offered a valuable distinction – based on Glouberman and Zimmerman (2002) – between simple, complicated and complex aspects that may characterize an intervention. Simple interventions are discrete and standardized, implemented by a single organization (or 'actor'), pretty much the same everywhere. Complicated interventions have multiple components (e.g. they may be multidimensional, as integrated projects of welfare, or interventions for sustainable development), or work only in conjunction with other interventions. They are implemented by multiple identifiable organizations in predictable ways (they are multi-site), and they work differently in identifiably different situations (depending on the implicated population or on the characteristics of 'implementation environments'). Complex interventions are non-standardized and adaptive, 'emergent' in response to changing needs, opportunities and understandings of what is working; they are implemented by multiple organizations with emergent and unpredictable roles; generalizations rapidly decay, and results are sensitive to initial or starting conditions as well as to context.

This distinction has implications for causation (Rogers, 2010). For simple programmes, intervention is both necessary and sufficient to produce results. For complicated programmes, intervention may not be sufficient, or not be the only cause of impact. For complex programmes, there may be either a tipping point effect (a small change can produce a disproportionate effect) or recursive causality – vicious or virtuous circles.

White maintains that causation is demonstrated when *attribution* (of the result to the programme) is possible; and confronted with the issue of complexity he affirms that what is needed is showing the *contribution*[4] (in the sense of partial attribution) of the programme to the whole effect: if a net effect (i.e. that the result is only attributable to the programme) cannot be demonstrated, then what should be demonstrated is the proportion of the impact that is attributable to the programme. But how can one separate that proportion from the whole impact, if the programme works in a combined way, if something could not happen without something else? Using Rogers's distinction, one could say that attribution is only possible with simple programmes. If used in complicated programmes it may collapse things that ought to be distinguished (what works somewhere but not elsewhere). And as for complex programmes, the effect cannot be attributed to the programme but to sets of circumstances that cannot be anticipated in advance. All in all, given the kind of programmes that we face in most contemporary policy domains (international development, local development, public-health programmes, etc.) what would be needed is not the calculation of the net effect, but of what is a good mix! On the contrary, in these latter cases the search for a net effect may be counterproductive: ignoring unexpected positive results, the process of empowerment of stakeholders and the like.

Rival Explanations

The link between complexity and causation has been at the centre of evaluation theory ever since and has nurtured thinking about 'plausible rival hypotheses' (Campbell, 1969). Although it was originally treated as a methodological problem of validity, it has recently been revisited from the substantive perspective of programme theory. Commenting on Campbell's interest in 'reforms', that are by definition 'complex social change', Yin contrasts two strategies of Campbell: that of the experimental design and that of using rival explanations. He concludes that the second – as Campbell himself came to admit in Campbell (1994) – is better suited to complex interventions (that are changing and multifaceted), as it is with the complex case studies that have been Yin's turf for a long time[5] (Yin, 2000: 242). The use of rival explanations is common in other crafts (journalism, detective work, forensic science and astronomy), where 'the investigator defines the most compelling explanations, tests them by fairly collecting data that can support or refute them, and – given sufficiently consistent and clear evidence – concludes that one explanation but not the others is the most acceptable' (Yin, 2000: 243). These crafts are empirical: their advantage is that while a 'whole host of societal changes may be amenable to empirical investigation', especially those where stakes are currently the highest, they are 'freed from having to impose an experimental design' ('the broader and in fact more common use of rival explanations covers real-life, not craft, rivals', Yin, 2000: 248). Nonetheless, rival explanations are by no means alien to evaluation, as is shown by how Campbell himself has offered Pawson good arguments for criticizing the way systematic reviews are conducted (Pawson, 2006).

The problem that remains is how to identify rival explanations. From a methodological starting point, Yin says that 'evaluation literature offers virtually no guidance on how to identify and define real-life rivals'. He proposes a typology of real-life rivals, that can variously relate to targeted interventions, to implementation, to theory, to external conditions; and proposes examples of how to deal with them taken from such fields as decline in crime rates, support for industrial development, technological innovations, etc. However, Yin appears to overlook something that had indeed fascinated theory-based evaluation since its first appearance: the possible existence of different theories to explain the working of a programme, and the need to choose among them in order to test them.[6] And – as Patton (1989: 377) has advised – it should be noted that in this way it would

be possible to engage stakeholders in conceptualizing their own programme's theories of action. Nonetheless Yin's contribution in its explicitness and methodological 'correctness' is an important step forward.

Weiss responded to Yin's provocative stance. In an article entitled 'What to do until the random assigners come' she locates Yin's contribution as the next step beyond Campbell's ideas about plausible rival hypotheses: 'where Campbell focused primarily on rival explanations stemming from methodological artefacts, Yin proposes to identify *substantive* rival explanations' (Weiss, 2002: 217). She describes the process by which an evaluator 'looks around and collects whatever information and qualitative data are relevant to the issue at hand' (2002: 219), in order to see 'whether any [other factor, such as other programs or policies, environmental conditions, social and cultural conditions] could have brought about the kind of outcomes that the target program was trying to affect', thus setting up systematic inquiries into the situation. Weiss concludes that alternative means to random assignment in order to solve the causality dilemma can be 'a combination of Theory-based Evaluation and Ruling Out' (the rival explanation).

Implementation Failure

What does 'implementation failure' mean? Is it possible to identify right vs wrong implementation?[7] The implementation phase is where a programme is put into effect by real people, concerned stakeholders, in specific contexts: it is therefore where the real world creeps in, mobilizing diversity and latent energies.

Goal-oriented approaches conceive implementation as the moment in which the 'treatment' is administered to the target group according to protocols, that should be matched to the objects of the programme, and as clearly as possible so as to produce uniform behaviour that can be measured and correlated with results. Goal-free approaches rest on the observation that during implementation, programmes are redefined by differently situated actors who follow different courses of action, according to different contexts, cultural or environmental: 'implementation is a complex, multistage process of institutional and individual learning' (McLaughlin, 1984: 100). Implementation is when a programme becomes what it is.

Two main issues have been recently addressed as problems of implementation failure: they refer to the key questions of the implementors' behaviour and to context.

Guidelines and Protocols

The first one is centred around the guidance offered. Goal-oriented approaches maintain that in order to have an accurate evaluation of implementation there should be precise guidelines and implementation protocols, against which it would be possible to assess 'disciplined performance', and 'implementation fidelity' (Donkoh et al., 2006: 15). Knowing that implementers will use different means (they are after all typical examples of 'street-level bureaucrats': Lipsky, 1980), the problem becomes that of identifying a model that can become a means of assessing distance or proximity. In the same way as in economics 'perfect competition' is almost never found, because of monopolies, oligopolies, and the like, but still works as a heuristic device to understand the workings of the market, so in the administrative environment where programmes are implemented it is believed that 'perfect administration' (Hood, 1976)[8]– characterized by clear objectives, division of tasks, good choice of leaders, enforced sanctions for deviants – is a useful device for understanding what happens in what is in reality a far from perfect world.

The idea of a precise norm to which implementers have to adhere, contrasts strongly with some features of implementation that we know to be crucial to programmes and to evaluation, and may by themselves be a cause of malfunctioning if not taken into account:

- implementation is the process through which it is possible to improve a flawed programme;
- learning is a matter of people understanding what is working better;
- things change during implementation, and no protocol, however flexible, can forecast all possible situations;
- implementation is an emerging process.

Goal-free approaches, as applied for example to street-level evaluation (Brodkin, 2003), have emphasized the idea of learning and reflexivity. According to this view, guidelines are seen as broad suggestions that are not compulsory and should be interpreted by responsible implementers able to mobilize hidden resources and to understand how to master difficult situations. This debate is particularly relevant in the European context, where the diversity of environments where EU programmes are implemented was originally tackled with strict bureaucratic rules that have failed to achieve harmonization, but instead suppressed local capacities. More recent evaluation guidelines (for instance, European Commission, 2006) have been issued that allow for more room to manoeuvre, and potentially allow for improved evaluation of the implementation of programmes.

Context

Real-life implementers do not respond only to their own attitudes and beliefs, but live in specific contexts. Context plays a varying role in goal-free approaches, and in none is it so crucial as in Pawson and Tilley's (1997: 70) realist evaluation. Context is not seen as simply 'the spatial or geographical or institutional location into which programs are embedded', but as 'the prior set of social rules, norms, values and interrelationships gathered in these places which sets limits on the efficacy of program mechanisms'. This embedded stratified situation interacts with the mechanism at work in the programme to produce outcomes that vary according to place, time and circumstance: any evaluation of an intervention will be confronted with results whose variations are explained by different 'medium range theories' (see the versatile application of the reference group theory by Pawson, 2010). And although realist evaluation is more interested in programme theory than in implementation theory – to use Weiss's distinction – it has attributed to context a role that no one will be able to dismiss, because – as Rogers notes – it has attributed a role to beneficiaries who can choose among implementation alternatives, thus challenging mechanistic ways of looking at programmes that deny discretion to both beneficiaries and implementers (Rogers, 1999: 382).

But context is a *bête noire* for followers of the counterfactual, who consider it as a confounding element in the relationship between treatment and impact that has to be kept under control. They need to consider it just as an intervening variable, and have jumped at the distinction between mediator and moderator variables (Baron and Kenny, 1986). Moderator variables are those that affect the strength and direction of treatment, and are therefore responsible for a varying impact, without affecting the logic of treatment/impact. The results may vary according to the values attributed to such moderator variables as related to people (women vs men, young vs old, resident vs immigrant), places (urban vs rural; technologically advanced vs backward) and even delivery itself (having received the treatment regularly vs irregularly). This is how White treats this topic, when he says that 'context is one aspect of impact heterogeneity', and admits that

> A study which presents a single impact estimate (the average treatment effect) is likely to be of less use to policy-makers than one examining in which context interventions are more effective, which target groups benefit most, and what environmental settings are useful or detrimental to achieving impact.'

Followers of the counterfactual are, however, less interested in mediator variables that act as an intermediate effect and would introduce an element of complexity that they are trying to roll back. Not by chance Donaldson (2007: 29) notes that 'most programs are considered to be multiple component interventions which are accurately conceptualized to contain multiple mediator variables', and that using mediator variables allows one to 'expose, often implicitly, theoretical program mechanisms'.[9] In a similar vein, Mark suggests utilizing a 'mediation model', by which a cause/effect relationship could be established between intervention and mediator, and between mediator and effect (as in the sequence 'intervention → attitude change → behaviour change').

Methodology Failure

Methodology is the terrain where White has chosen to fight his battle. Having admitted that 'a program theory provides a framework for an evaluation' he signals that 'it still needs an analytical approach to determine if outcomes have changed as a result of the intervention', and the missing approach is counterfactual analysis.

On the other hand, it is clear that both goal-oriented and goal-free approaches are concerned with methodology failure, although they have a different understanding of what such failure is about. For the goal-oriented it is confined to 'statistical power failure', whereas for goal-free approaches it implies problems in a wider range of methods, indeed, the same idea of mixed methods (Greene et al., 1989), so dear to goal-free approaches, comes from considering that each method has its limitations, hence the necessity to combine them.

Lipsey, as we have seen, is concerned with statistical significance that is obtained by correlating 'difference on sample data before and after an experiment' with 'difference between treatment and control group means' (2000: 103). The two-variables matrix offers two correct and two false conclusions. The two correct conclusions are: one, null hypothesis false, and rejected, hence net effect demonstrated, the programme works; two, null hypothesis true and accepted, hence no effect demonstrated, the programme does not work. The two false conclusions are called respectively 'Type I error' (the null hypothesis is rejected when it is likely to be true: false positive), and 'Type II error' (the null hypothesis is falsely accepted when it is likely to be untrue: false negative).[10] Type I error means that a failing programme continues to be funded, Type II error means that a good programme fails to be continued. It will be noted that Lipsey's original concern was for the false negative, i.e. not being able to demonstrate that a programme *worked*, whereas White and most contributions from the international development field seem to be more worried about the false positive, i.e. that programmes that *do not work* continue be funded.

Internal and External Validity

The main concern here is the validity of data: how far a measure actually captures the characteristics it is supposed to be measuring; and in our specific case, the validity of data showing that an intervention works (internal validity) and the findings are generalizable (external validity). Internal validity refers to the ability to demonstrate that in a given experiment the effect depends on the treatment; external validity refers to 'the extent to which the effects can be generalized to other

populations, settings, and treatment and measurement variables' (Campbell and Russo, 1999: 74). The distinction between these two forms of validity was originally raised in an article by Campbell and Stanley (1966) and rehearsed in Campbell's seminal article on 'Reforms as experiments' (1969). As we might expect from Campbell, he seeks to deal with threats to both types of validity. Among threats to internal validity Campbell included not only selection biases (what White is mainly interested in), but also history, maturation and instability. He focuses on designing an experiment, but also on what happens during the experiment (the programme implementation). Among threats to external validity Campbell included many instances of complexity, such as 'multiple treatment interference', and 'irrelevant replicability of treatments' (treatments are complex). These threats refer to the conditions in which an external factor may affect the laboratory situation of an experiment. But they also show Campbell's awareness of the problem of complexity, which he treated by the method of plausible rival hypotheses: in fact, the threats to validity come from 'plausible rival hypotheses to account for the data' (Campbell and Stanley, 1966: 36; see also Mark, 1986: 48).

This issue has relevant policy implications, as is shown by a debate on external validity between Campbell and Cronbach that took place at the beginning of the 1980s.[11] The core of this debate is the relative importance attributed to internal rather than external validity. Campbell seemed to most commentators to be more interested in internal validity, although he denied it, saying that after all he was responsible for coining the concept of external validity, that he later on relabelled as 'proximal similarity' (Campbell, 1986). Cronbach (1982), who on the contrary was more interested in external validity, distinguished between:

- causal generalization among *similar populations*;
- generalizing from the samples to populations that are *manifestly different*, that is – as Cook (2000: 7) puts it – invoking causal explanation as the means for creating *transferable knowledge.*

Mark argues that the difference between Campbell's and Cronbach's approaches was about criteria they used, Campbell's emphasis being on scientific inquiry and Cronbach's on more immediate applied policy concerns. 'Campbell seems to assume that having confident inferences at a low level of generalization (internal validity) ultimately increases the confidence about higher level inferences'; 'Cronbach argues that evaluation should maximize not internal validity but relevance, that is the ability to draw inferences about the UTOS that is of interest in policy making' (Mark, 1986: 54).[12] Greene (2004: 174) notes:

> Cronbach underscored the importance of going from situation to situation and refining our understanding as we go, as well as extrapolating what is learned in one setting to others. … Cronbach's primary emphasis … was thus on the quality and defensibility of inferences to other contexts, on external validity rather than internal validity. … to situating evaluation as serving to enhance our understanding of the character of enduring social problems and how we can best address them in this context and the next and the next, rather than to strong inferences about the causal effect of a particular intervention in a given set of sites.

Although it is not possible to deny that Campbell's scientific attitude was strongly linked to his commitment to social improvement (as all his elaborations on 'reforms' and on the 'experimenting society' demonstrate), one can see in this debate the roots of current concerns about how to derive evidence on what works and what doesn't. It could be said that Campbell was worried lest external validity could *not* be shown (hence, the programme would not work), whereas Cronbach was looking

for *new areas* where programmes could work (or not work). Both these preoccupations persist even in current debates.

In a sense, only diehard followers of the counterfactual can claim to be mainly interested in internal validity, without realizing that this runs against their desire to offer the policy-maker a strong argument for taking far-reaching decisions. This paradox is even more obvious in the field of development evaluation, where goal-oriented approaches have focused on threats to internal validity (of single interventions). This is at a time when the thrust of globalization, cooperation among international agencies and new paths to development (as shown by the entry of 'new' powers such as China, Brazil and India onto the world stage) would argue that the greatest attention should be paid to problems of generalizability.

As Perrin (2000: 275) noted while commenting on Campbell's intellectual legacy to the *art* of evaluation, 'evaluation findings that cannot be generalized if only to the same program with identical characteristics at a future time, are of little or no use. Without being able to identify what factors are responsible for impact, findings about impact have little or no practical value.' And since in the reality of programme delivery little remains stationary, responsive programmes should be changing and adapting. Hence, in order to generalize findings, 'rather than attempting to eliminate as many extraneous factors as possible, we should strive instead toward differentness rather than sameness in program elements and contexts'.

Syntheses

There have been two main ways of continuing the search for external validity, which – not surprisingly – are represented by the different approaches to syntheses, as a way of generalizing. The first one is well represented in Lipsey's reading of Campbell's late elaboration on external validity which has found its way into meta-analysis. Meta-analyses[13] provide – Lipsey (2000: 25) declares – 'large and heterogeneous samples of persons, settings, manipulations and outcome measures' that are lacking in single studies but are 'crucial for confident causal generalization'. According to him, in his later years Campbell 'welcomed the many heterogeneous replications that characterized most meta-analyses and saw that, through methods based on such replication, external validity could be placed on a firmer footing' (2000: 39).

It is interesting to contrast this view with a piece written by Pawson (2004) entitled 'Would Campbell be a member of the Campbell Collaboration?', which he answered by a loud 'No'. Pawson criticizes the way in which 'replication' takes place in meta-analyses, based on 'procedural uniformity' and on 'hierarchy of evidence' (alias 'the gold standard') that Campbell's methodology of 'evolutionary epistemology' would not have approved of, hinging as it did around the 'crucial question of balance between variation and retention' (Campbell and Russo, 1999: 134).

This leads me to the second way of elaborating on external validity, represented by the need to learn from good practice, to transfer successful programmes from one field to another, in ways that could be reminiscent of Cronbach's attitude. Contrary to the many optimistic renderings of 'best practices',[14] that would imply imposing uniformly to other places something that has been seen to work elsewhere, the real problem in this way of transferring knowledge is the search for the conditions that make (or do not make) it transferable. In realist syntheses, where it is not programme effects on populations that are compared but mechanisms in contexts, what is at stake is not whether the same 'treatment' produces the same effects on different populations (or different effects depending on the moderators), but whether theories of programmes that can be found 'across policy, disciplinary and organizational boundaries' (Pawson, 2006: 178) can explain 'what works for whom, in what circumstances and in what respect'. Because 'social

interventions … are never implemented the same way twice' (2006: 170) and interventions are interpreted and reinterpreted by their participants, it is not possible to demarcate programmes-that-work from programmes-that-do-not: 'understanding how a particular intervention works requires a study of the fate of each of its many, many intervention theories' (2006: 171). Hence, instead of synthesizing studies that replicate the same intervention, it is suggested to review programme theories as they are found in different situations. This is a way of openly addressing complexity and heterogeneity. Pawson notes (2006: 173) – that while 'heterogeneity is normally considered the curse of systematic reviews – … from a theory development perspective much can be learned, for example, about the utility of league tables by comparing their application in schools and hospitals'. As the realist synthesis had shown (Pawson, 2006: ch. 7), public disclosure initiatives had varying effects according to the 'susceptibility and status of the named party', the kind of sanctions following disclosure, the ability to control the information agenda, the power and independence of the responsible body, which corresponds to as many theories as could emerge from the synthesis.

Thus, Pawson concludes with a rebuttal of the main tenets of meta-analysis (as preached by the Campbell Collaboration), that is that generalizations (external validity) can be obtained through reproducibility of the research that is being synthesized. In the first place, decisions taken during a synthesis (for example, when to consider a search terminated) are mainly based on tacit knowledge of the researchers, which defies any idea of codified transparency. In the second place, Pawson maintains that a model of validity based on refutation (of theories) is preferable to one based on replication, and calls to witness Campbell's method of eliminating rival explanations, according to which 'organized distrust (among researchers) produces trustworthy reports' (Campbell, 1984: 38; Pawson, 2006: 182).

Gold Standard?

This problem leads us to the 'mother of all debates', that on gold standards, that is echoed by White. The point is whether there is a method that could be considered as the gold standard in evaluation, and therefore the best, in a hypothetical pyramid that sees at the top random control trials (RCT), then going down step by step: quasi-experiments, quantitative research (surveys), qualitative research and on down to ethnographic research.

Here is how White puts it: there is not a hierarchy of methods, rather it is imperative to use the best available method. For the evaluation question about impact, where it is a matter of attributing changes to a specific intervention, the best available method depends on the nature of the intervention being evaluated. When the unit of assignment is a large *n* experimental approaches should be used (the unit of assignment drives the power of calculation); however, if not RCTs, then some other quantitative (quasi-experimental) method will be the best available method. Instead, when there is a small- or medium-sized *n*, then qualitative approaches may be the best available methodology, but sometimes quantitative approaches can also be the most appropriate. In other words, counterfactual analysis is the best available method in particular cases, but there are limits to its use, although – according to White – at the moment there are many more cases where it should be used and isn't than the other way round.

This position has been contested on two main grounds:

- counterfactual analysis is not the only method for ascertaining causality;
- if there are limitations to the use of counterfactual analysis, how do we establish where to draw the line?

In a recent debate (Donaldson et al., 2009) Scriven has criticized the current myth of 'causation and evidence based on counterfactuals', that maintains that causal connections can only be inferred (statistically) and not observed. He has listed alternative designs for understanding causation used by natural science disciplines such as astronomy, epidemiology, as well as criminal law among others. According to Scriven, the only gold standard that exists is the 'General elimination method' (that is eliminating rival explanations). On the one hand, he maintains that causation is 'directly and reliably – indeed, trivially and universally – observable' (Scriven, 2009: 138), in the sense of 'critical observation', that is 'subject to the usual checks for the usual sources of error, including reflections and the likelihood of those' (2009: 140). On the other hand, according to Scriven the RCT approach is rarely able to substantiate the 'degree of certainty' that it claims, since RCTs are also 'entirely situation-dependent[15] in the normal context of social and educational inquiry' (2009: 142).

Scriven warns against the danger of overgeneralization about science itself, as when 'thinking that excellent designs for demonstrating causation and evidence in their own sphere are definitive for the whole of science' (2009: 150). In so doing, he comes close to the point of departure of this article, when he criticizes 'the attempted annexation of the concepts of *significance* by *statistically significant*' (2009: 151; original emphasis). If improving 'rigor in the applied social sciences' is a commendable motivation, the 'way to do that is by increased care in picking up the right tool for each job and using it properly, not by an oversimplification of the task' (2009: 151).

The second ground for criticism of RCTs stems from the paradox of admitting limitations to RCTs. Followers of the counterfactual advocate that RCTs or quasi-experimental designs are the best way of conducting an impact evaluation, and at the same time are eager to warn that these can be applied only in a limited subset of cases. This point was at the centre of topical debates at the Cairo Conference on Impact Evaluation in development policies (see Chambers et al., 2009).

Indeed, as the European Evaluation Society (2007) statement on the importance of a methodologically diverse approach to impact evaluation noted, the instances when RCTs are suitable are limited to simple interventions (linear causality), where it is possible to control for context, when experimental and control conditions remain fixed and when it is ethically acceptable to engage in randomization; they are not appropriate in complex situations, or when there are emergent and unanticipated outcomes.

Other limitations that goal-oriented approaches should take into account are of a different kind. It is now fashionable to propose prospective evaluation,[16] that is to say to prepare a collection of data on a variable of interest for an experimental group and a control from the inception of a new programme. But – as Patton (2008: 6) reminds us – RCTs 'are not appropriate at the start up of new projects and new initiatives, which need time to work out inevitable implementation problems and get the intervention stabilized and standardized prior to implementing an impact evaluation'.

All these limitations mean that if only the scant cases thus defined can be evaluated, the majority of policies – and even the most important ones – will not be evaluated. Put another way: one evaluates only what can be evaluated with RCTs, programmes are evaluated according to their evaluability rather than according to their importance or the needs of stakeholders. This is not a good service to the impact evaluation-hungry policy-maker depicted by White, not to speak of stakeholders and beneficiaries.[17]

Situational Responsiveness

The variety of answers to both queries (alternative designs for causal explanation and limits to counterfactuals) has brought us to a different position on the goal-free side: the quality of impact evaluation does not depend on a single gold standard method, but on the appropriateness of the methods

utilized to the situation.[18] Patton (2010) calls this situational responsiveness. Utilizing this approach Rogers (2010) has provided examples of how to choose the appropriate methodology in cases of simple, complicated or complex programmes. This is provided by answering the following questions:

- What is the nature of the intervention? Simple, complicated or complex?
- What is the nature of the impact? Produced directly or indirectly? Short term impacts or long term impacts? Transformational (unlikely to be reversed) or fragile? Acting alone or in conjunction with favourable circumstances?
- Why is the impact evaluation being done? Who are the intended users? Whose values will be used? Will the focus be on the average effect, or the effect on the most disadvantaged? Is it being done to retrospectively justify expenditure, in which case credible estimates of net benefit will be sufficient? Or is it being done to inform possible scaling up of a pilot (in which case good information will be needed on how it works)?

Appropriate methods will have to be found for each of these different situations, and, apart from cases in which the limits of counterfactuals are manifest, it will be a matter of combining different methodologies – that might include any method, RCTs included – in a mix suitable to the specific situation.

Conclusion

The need for better understanding by policy-makers of policy impact and effectiveness cannot be denied. Nor can the wish of evaluators to support sound policy-making be dismissed or belittled. However, concern about what doesn't work should engage evaluators in an effort to uncover failures that cannot be blamed on methodology alone, but also relate to programme theory and implementation. In reviewing ways in which these failures have been addressed by the two approaches identified as goal-oriented and goal-free, this article has envisioned a terrain in which encounters between them are possible.

No doubt, there are points on which the two approaches show a fundamental difference of perspective, as with the concept of what an intervention is like, whose stakes are to be considered central and what is the 'best available method' in evaluation.

There are, however, other points on which the boundaries between the different approaches are more blurred, or where both approaches invoke similar arguments to tackle questions that have common roots. It is interesting to note that such points have been raised across different types of failure, and that it has often been done in the name of Campbell, from opposing camps: indeed, Campbell's legacy has been claimed by a host of evaluators from different persuasions, including experimentalists like Lipsey, qualitative researchers like Yin,[19] critical realists like Pawson (2004).

Let me recall these 'boundary' topics. First, the method of ruling out plausible *rival explanations*. It has been invoked as a way of addressing *programme theory failure*. Yin used it when looking for causal explanations in complex programmes; he was supported by Weiss's elaboration on multiple programme theories. But Campbell had used it as a way of addressing *methodology failure*, in order to eliminate threats to validity. In his turn, Pawson has used it in realistic syntheses, where evidence emerges from theory elaboration, as an alternative to replication and reproducibility advocated by meta-analyses.

Second, the debate on *internal versus external validity*. Although it was seen as a problem of methodology failure related to the quality of data, it was shown to have greater policy implications, insofar as external validity is linked to the transferability of knowledge (Cronbach), and is therefore

also a means of theory improvement. Moreover, external validity is linked to the possibility of learning from 'good practices', itself as much an instance of 'what works' as anything else.

Third, the concept of *situational responsiveness* (Patton) that is mirrored by *situation-specific knowledge* as invoked by Campbell. This is an antidote to the 'gold standard' (i.e. as the best method to be used everywhere) in the name of a search for the method that is more appropriate to any situation/evaluandum. True, it will be difficult to establish a consensus on what exactly the evaluandum is like (simple? complicated? complex?), but it is something that cannot be overlooked, not even by followers of the counterfactual who admit to limitations in the use of their preferred method.

I suggest that these blurred boundaries could indicate possible sites for future debates, and new possibilities to make progress in better handling recurrent problems in evaluation.

Notes

1. Lipsey (2000: 39) is skeptical about 'using program theories as the central focus in program evaluation' because he fears (with Campbell) that as is usual in 'causal modelling' one is 'tempted to explain effects that (are) themselves not well documented'.
2. There are, of course, other differences between the two conceptions: e.g. in goal-free approaches policy-makers and implementers are seen to have greater discretionary than in goal-oriented.
3. See Weiss (1998: 55) for various definitions of theory.
4. This use of contribution is different from Mayne's (2010), who uses contribution analysis within the frame of theory-based evaluation:

   > Contribution analysis builds on the idea of using a program's theory of change to infer causation. … The result of a contribution analysis is not definitive proof, but rather evidence and argumentation from which it is reasonable to conclude the program has made an important contribution and why, within some level of confidence.

5. The story of this unpredictable encounter between Campbell and Yin is told in Yin (2000), and refers to Campbell's introduction to Yin's book on case study research (Yin, 1994).
6. See e.g. Weiss (1972) on the programme based on visits by teachers to students' families, in order to improve students' performance.
7. McLaughlin (1984: 99) speaks of a further failure: 'label fallacy', when the same name of a programme may mean different things in different institutions. This, however, does not mean that one is right and the other wrong, but they may be different.
8. For an elaboration on this point, see Martini and Sisti (2009).
9. One will also remember that Weiss (1997) had noted the need to make a better use of mediator variables.
10. See also Yin (2000: 246).
11. Greene (2004: 174) reminds us of the 'superb exchange' between Cronbach and Campbell on the 'relative merits of external vs. internal validity'.
12. UTOS is an acronym for Units (population, sites), Treatments, Observations (data collected), Settings (for the design and use of evaluation studies).
13. Lipsey (2000: 25): 'meta-analysis was originally developed to provide quantitative reviews of the very type of descriptive causal connection whose generalization is under discussion here. The technique has since been extended to cover non causal descriptive questions (e.g. 'do boys and girls differ in science achievement or persuasibility?') and to identify factors moderating a causal connection'.
14. For a full account of all the fallacies of 'best practices', and on wise ways of dealing with 'good practices' see Perrin (2006).

15. This includes both the well-known threats to validity and the ability of evaluators to conduct RCTs.
16. Even in EU regional policy: see Barca (2009).
17. One could also reflect on the policy implications of the goal-oriented approach. If the benchmark is that of evaluability by RCTs, many projects and programmes, such as large humanitarian projects, risk not being funded because they are considered not amenable to adequate evidence of efficacy, should the basis for that judgement not be supported by counterfactually derived scientific knowledge and research methodology. But this would be the stuff of another article.
18. See also GAO (2009).
19. See Bickman (2000) for such different appraisals of Campbell legacy as those provided by Lipsey, Yin and Perrin.

References

Barca F (2009) *An Agenda for a Reformed Cohesion Policy*, independent report prepared at the request of the Commissioner for Regional Policy, EU.

Baron RM, Kenny DA (1986) The mediator-moderator variable distinction in social psychological research: conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology* 51(6): 1173–82.

Bickman L (ed.) (2000) *alidity and Social Experimentation*, vol. 1 of *Donald Campbell's Legacy*. Thousand Oaks, CA: SAGE.

Brodkin E (2003) Street-level research: policy at the front lines. In: Corbett t, Lennon MC (eds) *Policy into Action: Implementation Research and Welfare Reform*. Washington, DC: Urban Institute Press.

Campbell D (1969) Reforms as experiments. *American Psychologist* 24: 409–29. Now in Campbell D, Russo J (1999) *Social Experimentation*. Thousand Oaks, CA: SAGE.

Campbell D (1984) Can we be scientific in applied social sciences? In: Connor RF, Altman DG, and Jackson C (eds) *Evaluation Studies Review Annual*, 26–48. Newbury Park, CA: SAGE.

Campbell D (1986) Relabeling internal and external validity for applied social scientists. In: Trochim WMK (ed.) *Advances in Quasi-Experimental Design and Analysis*, New Developments in Program Evaluation, 31: 67–78.

Campbell D (1994) Foreword. In: Yin R, *Case Study Research*, ix–xi. Thousand Oaks, CA: SAGE.

Campbell D, Russo MJ (1999) *Social Experimentation*. Thousand Oaks, CA: SAGE.

Campbell D, Stanley JC (1966) *Experimental and Quasi-Experimental Design for Research*. Chicago: Rand-Macmillan.

Chambers R, Karlan D, Ravallion M, and Rogers P (2009) *'Designing Impact Evaluations: Different Perspectives.* 3IE Working Paper, 4. Washington.

Cook TD (2000) Toward a practical theory of external validity. In: Bickman L (ed.)*Validity and Social Experimentation*, vol. 1 of *Donald Campbell's Legacy*, 3–43. Thousand Oaks, CA: SAGE.

Cronbach LJ (1982) *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass.

Donaldson S (2007) *Program Theory-Driven Evaluation Science*. New York: Lawrence Erlbaum Associates.

Donaldson S, Christie C, and Mark MM (eds) (2009) *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* Los Angeles: SAGE.

Donkoh C, Undershill K, and Montgomery P (2006) Independent living programs for improving outcomes for young people leaving the care system. *Campbell Systematic Reviews*, Campbell Collaboration. URL:www.cochrane.org/reviews/en/ab005558.html

EES (European Evaluation Society) (2007) *The Importance of a Methodologically Diverse Approach to Impact Evaluation – Specifically with Respect to Development Aid and Development Interventions*. URL: europeanevaluation.org.

European Commission, DG Regional Affairs and DG Employment (2006) *Indicative Guidelines on Evaluation Methods: Evaluation during the Programming Period*. URL: http://www.3kps.gr/ex-ante2007-2013/html/arxeia/WD5_Ongoing_Evaluation_EN_0906.pdf.

GAO (2009) *A Variety of Rigorous Methods Can Help Identify Effective Interventions*. Washington, DC: GAO.

Glouberman S, Zimmerman B (2002), *Complicated and Complex Systems: What Would Successful Reform of Medicare Look Like?* Commission on the Future of Health Care in Canada, Discussion Paper, 8. URL: http://www.hc-sc.gc.ca/english/pdf/romanow/pdfs/8_Glouberman_E.pdf.

Greene JC (2004) The educative evaluator: an interpretation of Lee J. Cronbach's vision of evaluation. In: Alkin MC (ed.) *Evaluation Roots: Tracing Theorists' Views and Influences*. Thousand Oaks, CA: SAGE.

Greene JC, Caracelli V, and Graham WF (1989) Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis* 11(3): 255–274.

Hood C (1976) *The Limits of Administrations*. London: John Wiley.

Lindblom CE (1968) *The Policy-Making Process*. Englewood Cliffs, NJ: Prentice Hall.

Lipsey M (2000) Statistical conclusion validity for intervention research: a significant (p <.05) problem. In: Bickman L (ed.) *Validity and Social Experimentation*, vol. 1 of *Donald Campbell's Legacy*. Thousand Oaks, CA: SAGE.

Lipsey M, Crosse S, Dunkle J, Pollard J, and Stobart G (1985) Evaluation : the state of the art and the sorry state of the science. In: Cordray DS (ed.) *Utilizing Prior Research in Evaluation Planning*, New Directions for Program Evaluation, 27. San Francisco: Jossey-Bass, 7–28.

Lipsky M (1980) *Street-Level Bureaucracy*. New York: Russel Sage Foundation.

McLaughlin MW (1984) Implementation realities and evaluation design. In: Shotland RL, Mark MM (eds) *Social Science and Social Policy*. Beverly Hills, CA: SAGE.

Mark M (1986) Validity typologies and the logic and practice of quasi-experimentation. In: Trochim WMK (ed.) *Advances in Quasi-Experimental Design and Analysis*, New Developments in Program Evaluation, 31: 47–66. San Francisco: Jossey-Bass.

Martini A, Sisti M (2009) *Valutare il successo delle politiche pubbliche*. Bologna: Il Mulino.

Mayne J (2010) Addressing cause and effect in simple and complex settings through contribution analysis. In: Schwartz R, Forss K, and Marra M (eds) *Evaluating the Complex*. New Brunswick, NJ: Transactions Publishers .

OECD, DAC (2002) *Glossary of Key Terms in Evaluation and Results Based Management.* URL: www.oecd.org/dac/evaluation.

Patton MQ (1989) A context and boundaries for a theory-driven approach to validity. *Evaluation and Program Planning* 12: 375–7.

Patton MQ (2008), *State of the Art in Measuring Development Assistance*, Mimeo.

Patton MQ (2010) *Developmental Evaluation*. New York: Guildford Press.

Pawson R (2004) Would Campbell be a member of the Campbell Collaboration? *The Evaluator* (Winter), 13–15.

Pawson R (2006) *Evidence Based Policy: A Realist Perspective*. London: SAGE.

Pawson R (2010) Middle range theory and program theory evaluation: from provenance to practice. In: Vaessen J, Leeuw FL (eds) *Mind the Gap: Perspectives on Policy Evaluation and the Social Sciences*, 171–202. New Brunswick, NJ: Transactions Publishers.

Pawson R, Tilley N (1997) *Realistic Evaluation*. London: SAGE.

Perrin B (2000) Donald T. Campbell and the art of practical 'in-the-trenches' program evaluation. In: Bickman L (ed.) *Validity and Social Experimentation*, vol. 1 of *Donald Campbell's Legacy*, 262–282. Thousand Oaks, CA: SAGE.

Perrin, B. (2006) How evaluation can help make knowledge management real. In: Rist R, Stame N (eds) *From Studies to Streams*. New Brunswick, NJ: Transactions Publishers.

Rogers P (1999) Review of 'Realistic evaluation'. *American Journal of Evaluation* 20(3): 381–3.
Rogers P (2008) Using program theories to evaluate complicated and complex aspects of interventions. *Evaluation* 14(1): 29–48.
Rogers P (2010) Implications of complicated and complex characteristics for key tasks in evaluation. In: Schwartz R, Forss K, and Marra M (eds) *Evaluating the Complex*. New Brunswick, NJ: Transactions Publishers.
Scriven M (1981) *Evaluation Thesaurus*, 3rd edn. Inverness, CA: Edgepress.
Scriven M (2009) Demythologizing causation and evidence. In Donaldson S, Christie C, and Mark MM (eds) *What Counts as Credible Evidence in Applied Research and Evaluation Practice?*, 134–152 Los Angeles: SAGE.
Simon HA (1947) *Administrative Behaviour*. New York: Macmillan.
Suchman EA (1969) Evaluating educational programs. *Urban Review* 3(4): 15–17.
Weiss C (1972) *Evaluation Research*. Inglewood Cliffs, NJ: Prentice Hall.
Weiss C (1997) Theory-based evaluation: past, present and future. In: Rog DJ (ed.) *Progress and Future Directions in Evaluation,* New Directions for Evaluation, 76. San Francisco: Jossey-Bass.
Weiss C (1998) *Evaluation*, 2nd edn. Upper Saddle River, NJ: Prentice Hall.
Weiss C (2002) What to do until the random assigner comes. In: Mosteller F, Boruch R, *Evidence Matters: Randomized Trials in Education Research*, 198–224. Washington, DC: Brookings Institution Press.
White H (2010) A contribution to current debates in impact evaluation. *Evaluation*16(2).
Yin R (1994) *Case Study Research*. Thousand Oaks, CA: SAGE.
Yin R (2000) Rival explanations as an alternative to reforms as 'experiments'. In: Bickman L (ed.) *Validity and Social Experimentation*, vol. 1 of *Donald Campbell's Legacy*, 239–266. Thousand Oaks, CA: SAGE.

Nicoletta Stame works at the Universita di Roma 'La Sapienza', Italy. [Email: nstame@aconet.it]