DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy

Impact evaluation: From black boxes to more realistic alternatives

Johann Mouton

CREST, Stellenbosch

The storyline

The imperative to conduct (more) impact evaluation studies forces us to revisit some of the basic assumptions that underpin this design logic. There are a range of assumptions about the definition, logic and design of such studies that have in recent times been critiqued. We address these issues under the following themes:

- The first reflections will be on current definitions of "impact evaluation" .
- A second focus will be on the dominance of the RCT approach and how this has come about. We discuss the successionist notion of causation as well as the central role of the logic of counterfactual in the RCT approach and then show how these have kept the black box of interventions closed.
- We then discuss how alternative views of causation (configurational and generative approaches) have opened the "black box" paradigm.
- The implications of these shifts for alternative designs for impact evaluation are finally discussed.

DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy

Different definitions of impact evaluation

UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner

Definitions of impact evaluation

In a recent paper Michael Bamberger suggested that there are two quite different definitions of impact evaluation that are widely used. The first, that can be called the *technical* or *statistical definition,* defines IE as an evaluation that

> *… assesses changes in the well-being of individuals, households, communities or firms that can be attributed to a particular project, program, or policy. The central impact evaluation question is what would have happened to those receiving the intervention if they had not in fact received the program. Since we cannot observe this group both with and without the intervention, the key challenge is to develop a counterfactual—that is, a group which is as similar as possible (in observable and unobservable dimensions) to those receiving the intervention. This comparison allows for the establishment of definitive causality —attributing observed changes in welfare to the program, while removing confounding factors.* [**Source:** World Bank Poverty Net website]

The second, that can be called the *substantive long-term effects* definition is espoused by the Organization for Economic Co-operation and Development's Development Assistance Committee (OECD/DAC). This defines impact as:

> *positive and negative, primary and secondary, long-term effects produced by a development intervention, directly or indirectly, intended or unintended* [**Source:** OECD-DAC 2002, p. 24].

Methodological consequences of these two definitions

The OECD/DAC definition does not require a particular methodology for conducting an IE, but specifies that impact evaluations should only assess *long-term effects*.

*The World Bank de*finition requires a particular methodology (the use of a counterfactual, based on a pretest/posttest project/control group comparison) but does not specify a time horizon over which impacts should be measured, and does not specify the kinds of changes (outputs, outcomes or impacts) that can be assessed.

To some extent these definitions are inconsistent as the *technical* definition would permit an IE to be conducted at any stage of the project cycle as long as a counterfactual is used; while the *substantive* definition would only permit an IE to assess long-term effects but without specifying any particular methodology.

According to Bamberger these definitions have specific and far-reaching implications as they speak to the choice for a specific approach to impact evaluation.

The technical definition of IE and RCT's

Advocates of the *technical* definition of an IE often claim that randomized control trials (RCT's) and strong quasi-experimental designs are the "best" and "strongest" designs (some use the term the "gold standard"). However, it is important to understand that these designs should only be considered as the "strongest" in an important but narrow statistical sense as their strength lies in their ability to eliminate or control for selection bias within the framework of the definition of impact evaluation offered.

Critics have pointed out that these designs are not necessarily stronger than other designs with respect to other criteria (such as construct validity, the validity and reliability of indicators of outcomes and impacts, and the evaluators' ability to collect information on sensitive topics and to identify and interview difficult-to-reach groups). When used in isolation these "strong" designs, also have some fundamental weaknesses such as ignoring the process of project implementation and lack of attention to the local context in which each project is implemented.

SciSTIP

DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy

The RCT orthodoxy

UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner

Perpetuating the myths of the superiority of RCT's

- There is a long-standing and pervasive view that randomized (randomly) control trials or RCT's are the best and only scientifically defensible designs to assess the impact of interventions (specifically when defined in the World Bank definition).
- The RCT or *randomly controlled trial*, is an experimental design involving at least two groups of subjects, the control group and the experimental group (or study group, or treatment group), between which the subjects are distributed by a strictly random process (i.e., one with no exceptions), and which are not further identified or distinguished by any common factor besides the application of the experimental treatment to the experimental group.
- The dual design features of "randomization" and "control groups" are seen as providing sufficient grounds for making strong causal claims about the intervention and specifically the only design that allows the evaluator to make strong claims about causal attribution. Ultimately this is because – so it is claimed – it meets the conditions of the logic of the counterfactual. (showing that the intervention is the only factor that could have made the difference that was found).

RCT versus quasi-experiments

- There is in fact an attempt in certain circles to elevate the RCT to the status of the only true experiment. (See overleaf) But it is important to keep in mind that the RCT is simply a special case of the family of experimental designs. This effort at persuasive redefinition is allied with an implicit denigration of the so-called "quasi-experimental" designs, which are in fact perfectly respectable experiments, only 'quasi' with respect to the one respect in which they have less control over one possible way of excluding one type of alternative explanation (Scriven).
- RCT's derive their dominant support from cases where interventions are relatively simple, where the target group is clearly defined and where temporal and contextual factors are relatively unambiguous. Perhaps not surprisingly, the best field of application of RCT's are in the field of clinical trials where new drugs and vaccines are assessed.

RCT's: Randomized control/clinical trials: An example

"The randomized clinical trial is an experimental study conducted on clinical patients (with their consent of course!). The investigator seeks to completely control the exposure (in terms of its type, amount, and duration), and (most importantly) who receives it through the process of randomization. RCT's are regarded as the most scientifically vigorous study design.

Because:

- *An unpredictable (i.e., concealed) random assignment eliminates (or at least greatly reduces) confounding from known and unknown prognostic factors (that is, it makes the groups equivalent in terms of their prognosis at baseline).*
- *Blinding eliminates biased measurement, so that outcomes are measured with the same degree of accuracy and complete ness in every participant.*

Because of these conditions, it is then possible to confidently attribute cause and effect – that is, because the only thing that differed between the groups of the trial was the presence or absence of the intervention, any effect can be ascribed to it (assuming a well conducted, unbiased study). The RCT is therefore described as having high internal validity – the experimental design ensures that, within reason, strong cause and effect conclusions can be drawn from the results.

Scriven's critique of RCT's

In a 2008 paper entitled "A Summative Evaluation of RCT Methodology: An Alternative Approach to Causal Research" lists five objections to the current uncritical adoption of RCT's.

1. *The RCT design is a theoretical construct of considerable interest, but it has essentially zero practical application to the field of human affairs.*
2. *Even the best double-blind drug studies do not have the unique explanatory power claimed by the proponents of RCTs for their zero-blind studies.*
3. *Generalizing from even a true RCT to real world use of the experimental treatment tested in the RCT is hazardous, and any true RCT study has to be supplemented with extensive high quality field reports on real world use.*
4. The other logical advantage of the RCT design that is claimed by its protagonists is that it supports what is said to be the key logical property of causes, the fact that they support counterfactuals (i.e., a cause is something without which the effect would not have occurred). *However, this counterfactual-supporting property is certainly not a logical property of causes, as stated, and even more certainly is not a logical property possessed by the quasi-RCTs being supported by RCT protagonists.*
5. *The threats of confounding variables to RCT designs are extremely serious and numerous, and costly to handle, and require continual highly skilled attention that is often not budgeted or staffed in RCT studies*

RCT's have little practical use in human affairs

The RCT design is a theoretical construct of considerable interest, but it has essentially zero practical application to the field of human affairs.

It is important to be clear that a true RCT study has to be (at least) double-blind, as are all sound pharmacological studies. But its applications in fields such as public health, education and social services are neither double-blind nor even single blind, but 'zero-blind.' In most, if not all cases, the intended beneficiaries are aware that they are part of an intervention and also – where there are control or comparison groups – in which group they fall! *Hence the common argument that the RCT designs being advocated in areas like education, public health, international aid, law enforcement, etc., have the (unique) advantage of "eliminating all spurious explanations" is completely invalid.*

According to Scriven the application of the RCT approach to the human sciences is not a matter of methodological neglect, "but of the almost complete impossibility, at least within the constraints of the usual protocols governing experimentation with human subjects, of arranging for even single blind conditions". His point is that it is simply impossible – in the real world of social interventions – to hide to the participants the conditions under which the intervention is implemented.

SciSTIP

DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy

The positivist origins of the RCT approach

UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner

Going back in history

How is it that the RCT-approach has become so pervasive and dominant OR at least the "belief" that it is both the best and most appropriate approach to impact evaluation?

It has two roots (which are themselves related) in the early history of programme evaluation and specifically the experimental tradition (Campbell, Stanley and others) of the 1960's.

- The first is the positivist rejection of context and value in all matters scientific.
- The second is the positivist appropriation of the successionist (Hume) interpretation of causation.

The positivist choice for methodology over theory

According to Nicoletta Stame (2004: *Theory-based Evaluation and Types of Complexity)* the 'original sin' of mainstream evaluation (with a positivist imprint) lies in choosing to play a low-key role. Because of the positivist abhorrence to anything remotely normative or value-laden (hence their commitment to value-free science) the early evaluators concentrated their efforts on methodological challenges: - and specifically a methodology for verifying the internal validity (causality) and external validity (generalization) of programmes. This resulted in a commitment to the experimental tradition that soon became associated with "black-box evaluations".

The black box is the space between the actual input and the expected output of a programme. Moved by the need to tackle serious social problems, programme designers often gloss over what is expected to happen, the how and why, when an input is put in place; and evaluations do the same concentrating on measuring outputs, whilst attributing the observed difference to the input. All this is hardly informative for a policy design wishing to build upon previous experience.

Consequences of the "low-key" approach to programme evaluation (Stame)

- Not discussing programme theories amounted to warranting programmes with 'absolute rationality' (assuming that needs are known, decision makers are informed about the risks and opportunities implied in each option, decisions are taken with the aim of maximizing the gains from existing resources) at a time when most policy analysis had accepted bounded rationality and incrementalism as its new paradigm.
- If programmes could be regarded as 'rational actions', then politics was seen as a disturbance or interference and the political context itself never became an object of inquiry.
- Concentrating on verifying the validity of programmes whose theoretical implications were not questioned, led evaluators to believe that the outcome of evaluation would be 'instrumental' use: saying that something worked or did not work. This provided the commissioner with a clear answer about 'go/no go' and the decision maker would then follow suit.

(Source: N Stame, 2004)

SciSTIP

DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy

Successionist interpretation of causation

UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner

The successionist theory of causation

The positivists of the first part of the 20th century (because of their anti-metaphysical stance) were happy to align with the account of causality given by David Hume in the 18th century. On his account causation is evidenced through the observation of regularities: if potential cause C and effect E are always found together, then either C causes E, or E causes C. This is what Hume meant when he said that "causation" is nothing but the "constant conjunction of events". The assumption is that a true cause does not work by accident, but operates constantly and regularly, producing the same effect over time and in different settings (hence its characterization as "lawlike").

What matters in regularity is the simultaneous observation of two separate entities, while the description of the causal connection (the nature of the "arrow"), or the process leading from C to E remains unknown; what happens in-between cause and effect, what the cause does in order to produce the effect, is kept closed inside what much literature has called the "black box". In Hume's words: "we can never penetrate so far into the essence and construction of bodies as to perceive the principle, on which their mutual influence depends".

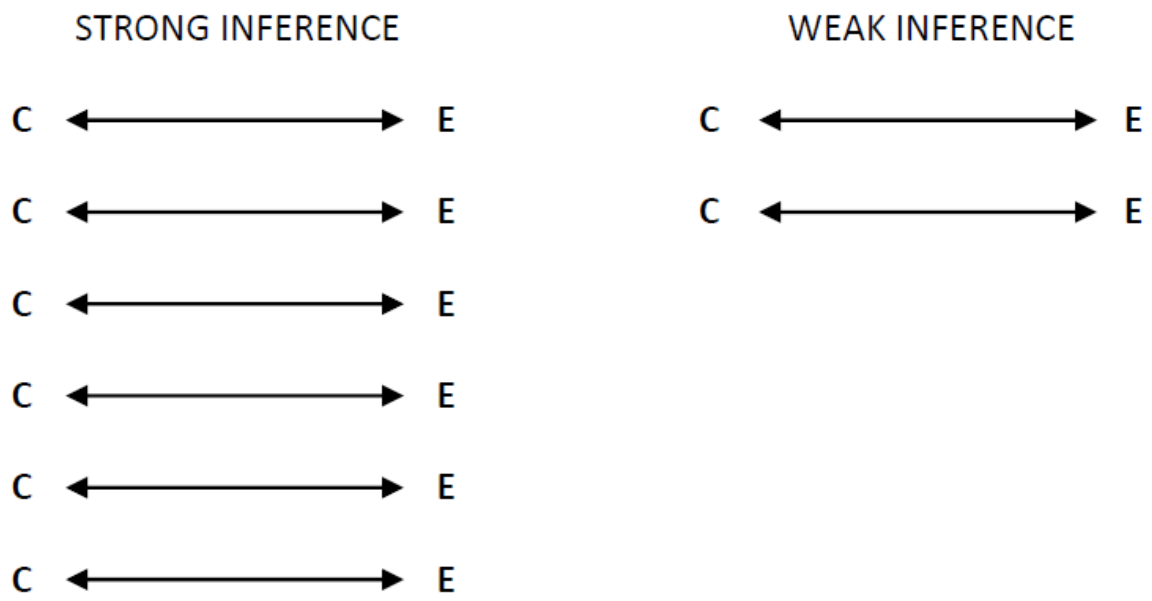
Strength of causal association

In the successionist view, the cause is both necessary and sufficient for the effect: sufficient because all events where the cause is observed also present the effect; "we may define a cause to be an object, followed by another [the effect], and where all objects similar to the first are followed by objects similar to the second" . But also necessary in that "if the [cause] had not been, the [effect] never had existed".

The main problem with this view lies in checking for the difference of "all possible elements": in practice this requires comparing a high number of events taking place in a wide range of settings. Because we cannot reach absolute certainty of inference, we have developed ways to evaluate the "quality" (likeliness) of the causal link: as in statistical modeling, within the more general regularity framework, the frequency of association between cause and effect strengthens the causal assumption; and theoretical relationships / associations / models that apply to a high number of cases are "more reliable" in their "explanatory power" than theories applying to a small number of cases or only one case.

The strength of causal association

In other words, **the strength of the causal association** increases with the number of cases where conjunction between cause and effect is observed; and finding cases where the cause is not simultaneously observed with the effect weakens the inference. In regularity causation, the closer we are to a "law" the better.



Regularity and impact evaluation

The successionist or regularity view of causation has some value in impact evaluation studies. It is useful, for example, if one wants to know what the numbers of beneficiaries are that demonstrate certain characteristics after having received an intervention. But this approach does not provide answers on why the beneficiaries have these characteristics, nor on how they were developed following the intervention. It simply shows that the more frequent the number of "administrations"of the intervention, the more likely one would expect to see certain changes or effects.

Thus regularity does not allow the evaluator to trace the process leading from cause to effect, and the attribution of the impact, while shown to hold in many cases, lacks "depth" on how the causal association happens. The black box remains unopened.

SCISTIP

DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy

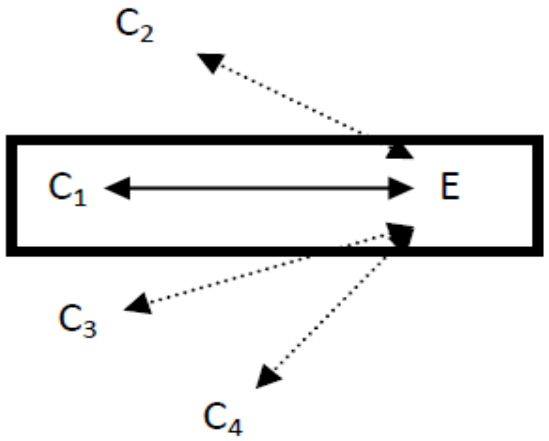
Counterfactuals as the solution?

UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner

On counterfactuals

The central impact evaluation question is what would have happened to those receiving the intervention if they had not in fact received the program (World Bank)

On the face of it counterfactuals have the advantage of needing only two events (as opposed to infinite) to infer causation. In evaluation terminology these two events are exemplified in the treatment and control groups ("events"). However, this advantage is only apparent because those two "events" need to be identical on an infinite number of elements except cause and effect. The cause is isolated through the careful choice of the two events to be compared:



How the logic of the counterfactual works

The logic of the counterfactual is traced back to the work of the British philosopher John Stuart Mill and his Method of Difference. According to Mill, the causality of C with regard to effect E is claimed in the following way:

f g h i j k | f g h i j k C E => C is the cause of E (or E the cause of C)

The above two events are compared and C is the only "new entry" in the second event: all the other elements f g h i j k l m are present in both. When other factors are present in only one event along with C and E, we cannot infer causation; for example in:

f g h i j k | f g h i j k L C E => either L or C could be causes of E.

While f, g, h, i, j and k are rejected on the grounds that they are present in both events, L cannot yet be rejected. In order to reject L, too, we need to find the case f g h i j k L.

How the logic of the counterfactual works

As with the regularity approaches, the **strength of inference** through counterfactuals increases as the number of alternative causes we are able to reject increases; the higher the number of elements that can be shown to be equal in the two events, the better.

f g h | f g h C E => **WEAK INFERENCE** (i, j, k, l, m and n haven't been rejected yet)

f g h i j k l m n | f g h i j k l m n C E => **STRONG INFERENCE** (many more causes have been eliminated)

The logic of the counterfactual embedded in the RCT design

The most common method used to find a specific event presenting a number of specific factors without C and E is the experiment in controlled settings and specifically the RCT. The RCT randomizes who receives an intervention (or service, or pill) – the treatment group - and who does not – the control. It then compares the outcomes between those two groups; this comparison gives us the impact of the programme. In this way, the control mimics the counterfactual. The counterfactual is defined as what would have happened to the same individuals at the same time had the programme not been implemented. It is, by definition, impossible to observe – it's an alternative universe! There are two basic ways to approximate the counterfactual: (i) using the outcome observed for non-beneficiaries (the control group); or (ii) using the outcome observed for beneficiaries before they are exposed to the intervention (the baseline).

However, even when the known threats to experiments are controlled for and the causal association covers a high number of cases, a knowledge gap remains on the characteristics of the association between a given factor and a given effect. Like the regularity approach discussed above, counterfactuals do not provide answers as to what happened "between" the alleged cause and the effect; e.g. on what the "true" cause was at the micro, in-depth level.

SciSTIP

DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy

Peeking into the black box: Causal fields/ configurational causation

UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner

INUS causality and causal fields

- In an attempt to remedy the inadequacy of the successionist view in accounting for the effects arising from interaction of causal agents, John Mackie introduced in 1965 the notion of "causal field", exemplifying the idea that the link to be studied is not between the effect and a single cause, but between the effect and a causal package: a "block" of single causes that might not have an independent influence on the effect. In 1974 the same author theorizes a special type of cause called the INUS: an insufficient (I) but necessary (N) part of a causal package, which is in itself unnecessary (U) but sufficient (S).
- Fire cannot only be caused by gas fumes and cigarette lighting, although they are jointly sufficient for it. It can also be caused by, for example, striking a match on a red phosphorous surface. Each of these four causes is an INUS: none of them is sufficient for fire; but each of them is necessary for a combination to be sufficient for fire. The match in itself does not light a fire, but neither so does the red surface: none of them alone are sufficient, and both of them need the other to produce fire. In other words they are jointly sufficient in that they are part of a sufficient combination. This combination, however, is not necessary for fire to happen: fire also happens when cigarette lighting devices being activated meet gas fumes.

Configurational causation – a peek into the black box

The analysis of necessity and sufficiency thus reveals that many causes come in "packages": pairs or vectors of causes, and are unable to produce the effect unless they are combined with other causes; and that the same effect could be produced by several different combinations of different single causes. In what is called the configurational view of causation, a cause is thus identified with a configuration / constellation of conditions, or more precisely with a combination of single causes producing the same effect. Different combinations may lead to the same outcome; and similar combinations may lead to different outcomes, because individual conditions can affect the outcome in opposite ways, depending on what other factors they are combined with.

Configurational causation solves some problems connected with the successionist view but not all. Causality is no longer essentially single and independent, and is now seen as properly multiple and conjunctural; however the problem of establishing the direction of causation is not yet satisfactorily addressed. The black box has become partially transparent and some characteristics of the causal process begin to emerge (we know more than just beginning and end, we do have some in-between elements), but the understanding is not yet fine-grained and combinations need to be interpreted. We have started to "peek in" but still haven't opened the black box.

SciSTIP

DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy

Opening the black box: causal mechanisms (generative causation)

UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner

The black box and a successionist theory of causality

According to the realists - Pawson and Tilley - the problem of the 'black box', and of the failure of experimental designs of evaluation, lies in the 'successionist' theory of causality on which experiments are based. According to this view, we cannot know why something changes, but only that something has changed from status 'a' (without stimulus, without programme) to status 'b' (with stimulus, with programme) in a given case. And that is why it is so difficult to say whether the change can be attributed to the programme.

The realist approach is based on a 'generative' theory of causality: it is not programmes that make things change, it is people, embedded in their context who, when exposed to programmes, do something to activate given mechanisms, and change. **So the mystery of the black box is unveiled: *people inhabit it*.** This makes for a completely different design of evaluation. The evaluator elaborates how the mechanism could work in a given context and asks the people who could know about it to provide evidence.

Pawson and Tilley on realistic evaluation

The characteristic of this approach is to stress what the components of a good programme theory should be: context (C) and mechanism (M), which account for outcome (O). Evaluation is based on the CMO configuration. Programmes are seen as the opportunities that an agent, situated inside structures and organizations, can choose to take, and the outcomes will depend on how the mechanism that is supposed to be at work will be enacted in a given context. Mechanisms are not infinite in number, and programmes can be grouped in relation to the mechanisms around which they are built (naming and shaming, incentives, etc.). Nor are contexts limitless, but vary according to certain characteristics (density, marginality, etc.). Contexts and mechanisms are part of middle-range theories.

> *. . . the basic idea of middle-range theory is that these propositions do not have to be developed de novo on the basis of local wisdom in each investigation. Rather they are likely to have a common thread running through them traceable to [a] more abstract analytic frameworks [. . .]*. (Pawson and Tilley, 1997: 123–4)11

An everyday analogy

While everyone might know the recipe to a meal, just having some ingredients on the table does not make the meal. Someone must actually put together the ingredients in a certain way to obtain the final effect – the meal. While the configurational view of causation sheds some light on necessity and sufficiency (the recipe), we now focus on a notion of causality that informs on how the ingredients must put together; following what order and techniques. In other words, we will address the specific processes of mixing, mashing, flavouring, cooking, etc. that can make different meals out of the same ingredients depending on how someone mixes them together.

This leads us nicely to the notion of "causal mechanisms" as a description of the causal process taking place between cause and effect.

Chen and Rossi were among the first evaluators to introduce the term "mechanism" and point out its importance for theory-driven evaluation. In an early article, they argue that "the theory-driven approach avoids the pitfalls of black-box evaluation and provides better understanding of the causal mechanisms underlying the relationship between treatment and effects" (Chen & Rossi, 1987,)

Mechanism analysis and impact evaluation

There is considerable support for this view of mechanisms in much of the contemporary literature on mechanism-based analysis. According to this perspective, mechanisms are underlying entities, processes, or structures which operate in particular contexts to generate outcomes of interest. There are three essential clues located in a "realist" reading of mechanisms. These are that: (a) Mechanisms are usually hidden; (b) Mechanisms are sensitive to variations in context; and (c) Mechanisms generate outcomes.

This approach is important depending on the use we want to make of impact evaluation. If we need it to justify public spending, or for accountability purposes, then a simple link (or an unopened black box) might suffice. But if we need it to improve our interventions, or to extend it to other areas, then we need to know the details of the actual working gears – the hidden "mechanisms": what are these mechanisms exactly and how can we improve them / make it work in other areas? What mechanisms are actually important and most relevant?

SciSTIP

DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy

Concluding musings

UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner

Impact evaluation using case study designs

Some years ago I proposed that we consider the *evaluative case study design* as an appropriate design when assessing the impact of interventions. Three reasons were proposed:

- The necessity to take into account the specific context of an intervention under different conditions of implementation. Case studies adhere to the "logic of contextualisation" and the need to understand causal mechanisms within the context in which they are triggered and operate.
- In contradistinction to the typical experimental logic (and the focus on controlling for confounding variables), it embodies the logic of "causal narration" and hence is compatible with theory-based evaluation approaches.
- Counterfactual thinking is still crucial. But not in the sense of the logic of the counterfactual that dictates the use of randomisation or control groups – neither condition which is found in the vast majority of (complex) social interventions. But, within the logic of case studies, in the sense of assessing rival explanations or theories for programme outcomes.

Some concurrence

John Gerring in 2004 writes:

One can readily see why the investigation of causal mechanisms (including both process-tracing and pattern-matching) is commonly associated with the case study research design. The in-depth analysis of a single unit is useful in elucidating causal mechanisms because its characteristic style of evidence-gathering— over-time and within-unit variation—is likely to pro-vide clues into what connects a purported X to a particular Y.

And Scriven as well:

….since it is often suffused with causal claims based on observation, is immediately reinstated as a live candidate for respectable demonstration of causation.

In conclusion then….

- The uncritical belief in the superiority of RCT's in impact evaluation needs to be suspended. RCT's are strong designs and have their place in certain fields (most notably pharmacological interventions).

- Other quasi-experimental designs (and the use of the term "quasi" remains unfortunate) and especially time series designs are sufficiently powerful and much more relevant to large scale social interventions.

- The point of this paper was to focus on a more realistic alternative - the evaluative case study design. Although it often involves more time and resources than RCT's and quasi-experimental studies, it is firmly embedded in a realistic paradigm (with the emphasis on causal mechanisms as explanatory devices), it is sensitive to context and the different conditions under which interventions succeed or fail and it takes the construction of theories of change (as causal narratives) seriously.

SCISTIP

DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy

Thank you

UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner