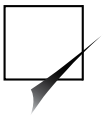

Article

Evaluation
2017, Vol. 23(1) 42–60

DOI: 10.1177/1356389016654584
journals.sagepub.com/home/evi

Process Tracing and Bayesian Updating for impact evaluation

Barbara Befani
University of Surrey and University of East Anglia, UK

Gavin Stedman-Bryce
Pamoja Evaluation Services Ltd, UK

Abstract
Commissioners of impact evaluation often place great emphasis on assessing the contribution made by a particular intervention in achieving one or more outcomes, commonly referred to as a 'contribution claim'. Current theory-based approaches fail to provide evaluators with guidance on how to collect data and assess how strongly or weakly such data support contribution claims. This article presents a rigorous quali-quantitative approach to establish the validity of contribution claims in impact evaluation, with explicit criteria to guide evaluators in data collection and in measuring confidence in their findings. Coined 'Contribution Tracing', the approach is inspired by the principles of Process Tracing and Bayesian Updating, and attempts to make these accessible, relevant and applicable by evaluators. The Contribution Tracing approach, aided by a symbolic 'contribution trial', adds value to impact evaluation theory-based approaches by: reducing confirmation bias; improving the conceptual clarity and precision of theories of change; providing more transparency and predictability to data-collection efforts; and ultimately increasing the internal validity and credibility of evaluation findings, namely of qualitative statements. The approach is demonstrated in the impact evaluation of the Universal Health Care campaign, an advocacy campaign aimed at influencing health policy in Ghana.

Keywords
Advocacy Evaluation, Bayesian Updating, Impact Evaluation, Mixed Methods, non-experimental approaches, Process Tracing

Introduction

Currently, much of the demand placed on evaluators by commissioners of impact evaluations revolves around the formulation and validation of a 'contribution claim' about the role an

Corresponding author:
Barbara Befani, University of Surrey and University of East Anglia, UK.
Email: befani@gmail.com

intervention, or specific aspects of it, played in the achievement of one or more outcomes. The most popular theory-based approaches that evaluators are currently offering, in response to such demand, are perhaps Contribution Analysis (Lemire et al., 2012; Mayne, 2001, 2008, 2012), various forms of Systems-Based Evaluation (Derwisch and Löwe, 2015; Grove, 2015; Williams, 2015; Williams and Hummelbrunner, 2010) and Realist Evaluation (Pawson and Tilley, 1997; Westhorp, 2014) (sometimes used in combination with Qualitative Comparative Analysis, Befani et al., 2007).

Contribution Analysis requires the creation of a 'causal chain' where each link represents an intermediate outcome, associated with risks that might prevent it from taking place and assumptions that need to hold if the intermediate outcome is to materialize. Systems-Based Evaluation allows for complex interrelations to emerge among programme components, resources, actors, and intermediate outcomes. Realist Evaluation, finally, requires the identification of one or more Context-Mechanism-Outcome (CMO) configurations that explain in-depth why certain links do or do not materialize in the Theory of Change, illustrating in detail the reasoning behind the decision-making of specific actors.

These three approaches are methodologically neutral: they do not provide clear guidance on how to collect data and assess the strength of such data towards (or against) a contribution claim. In other words, they do not provide a 'methodological bridge' between the contribution claim and the type of empirical data needed to assess it, nor the most appropriate data-collection techniques to use. In addition, they do not provide touchstones nor benchmarks to assess the quality of the evidence, to measure its power to change our confidence that a given contribution claim holds true, or not.

The existence of such a large methodological gap in evaluation can lead to unsatisfactory practices. A classic example is the evaluation of the often cited 'Final-Push' advocacy campaign, aimed at influencing a US Supreme Court decision (Patton, 2008). The approach used is Scriven's General Elimination Methodology (GEM) (Scriven, 2008), which shares similarities with Process Tracing in that an observable outcome is traced back to several possible causes, and alternative explanations are considered and assessed. There is also a clear contribution claim which the evaluation seeks to affirm (or deny): the Final-Push campaign contributed to the Supreme Court's decision; and the evaluation eventually concludes that the campaign contributed significantly to the Supreme Court's decision.

Empirical data collection is often mentioned ('a thorough review', 'careful analysis', 'well documented', '45 knowledgeable people interviewed', etc.), but carries an almost defensive, self-referential tone, which is understandable given that by reading the article it is not possible to link the observations made during the data-collection process with the contribution claim; or to understand why the contribution was considered 'significant' as opposed to 'fair' or 'small' or 'non-existent'.

This has important implications for the internal validity of the evaluation results: if another team had carried out the evaluation, would the findings have been the same? Another important implication of this methodological gap is that evaluators can easily build evidentiary cases towards (or against) a contribution claim on the basis of the quantity of evidence gathered (abundance or dearth), rather than on the probative value of such evidence. By probative value, we mean the power of specific items of evidence to increase or decrease our confidence in a specific claim. The latter is not necessarily low simply because we might be looking at just 'one' piece of evidence.

This methodological gap leads to a kind of inefficiency that is particularly undesirable when resources for impact evaluation are limited and evaluators can only gather a limited amount of

data. If they do not approach data collection with the idea that different pieces of evidence can have different probative values and that the latter can be assessed rigorously and transparently, under time and resource constraints, they will likely fail to observe the most probative evidence. Indeed, the lack of clear guidance to support impact evaluators in assessing contribution claims may dissuade evaluation commissioners from evaluating the impact of specific interventions on the grounds that it is 'technically impractical'.

This article argues that the principles of Process Tracing and Bayesian Updating can be combined with any of the above-mentioned theory-based approaches (in particular with Contribution Analysis) to offer clear guidance on what data to collect, when and how; together with standards to measure how much the evidence increases, or decreases, our confidence in a contribution claim.

Process Tracing has been previously discussed in connection with evaluation (Bjurulf et al., 2012; Byrne, 2013; Frey and Widmer, 2011; Mohr, 1999; Ton, 2012). The uses of Process Tracing proposed in these contributions do not identify the implications of this method for evaluation practice in terms of 1) clear identification of questions guiding data collection; 2) measuring the evidential strength or probative value specific observations provide for given contribution claims or mechanisms or theories of change under test; and 3) making the evaluation process transparent enough to be scrutinized by other interested parties. One recent contribution (Schmitt and Beach, 2015) explicitly refers informal Bayesian logic and classifies different pieces of evidence on the basis of their inferential power and types of Process Tracing test for which they can be used. However, it stops short of making the method accessible and ready for ' "hands-on" usage in complex evaluations', and identifies this work as 'an important task for the future'.

This article aspires to be a step towards achieving the 'important task' of making Process Tracing principles and tests ready for application in real-life evaluations. Building on a previous attempt to make Process Tracing more relevant for evaluators, in particular those interested in using Contribution Analysis (Befani and Mayne, 2014), it aims to provide simple and straightforward guidance on what questions to ask, how to assess the strength of different pieces of evidence, and how to expose this work to the scrutiny of a 'jury', with the aim of increasing the robustness and transparency of the findings. In the conclusions we acknowledge that further guidance needs to be offered if we are to provide evaluators with a fully loaded 'off the shelf' framework, and sketch the contours of what we believe to be the task of future research.

Basic principles of Process Tracing and Bayesian Updating

Process Tracing has been referred to as a method (Beach and Pedersen, 2013; Collier, 2011) but also as a tool (Bennett, 2010; Collier, 2011) and a technique (Bennett and Checkel, 2014a) for data collection and analysis, reflecting its focus on theory development as much as on the search and assessment of evidence for a causal explanation (also reflected in the distinction between the two 'deductive' and 'inductive' variants (Beach and Pedersen, 2011; Bennett and Checkel, 2014a). Its purpose is to draw causal inferences from 'historical cases', broadly intended as explanations of past events. It is based on a mechanistic understanding of causality in social realities, and starts from the reconstruction of a causal process intervening between an independent variable and an outcome, which could for example be a Theory of Change, a complex mechanism or a CMO configuration.

The method makes a clear distinction between: a) the process described in the Theory of Change, considered a possible 'reality', or an ontological entity which might or might not exist or have materialized; which is usually unobservable; b) the evaluator's hypothesis on the existence of that reality (which is an idea in 'our head' [Bennett and Checkel, 2014a] rather than a reality 'out there'); and c) the observable and therefore testable implications of the existence of such reality. This tripartite conceptual framework stems from the awareness that mechanisms in the social sciences are usually not directly observable; we can never attain perfect certainty of their existence but nevertheless we formulate hypotheses about their existence and look for evidence in an attempt to increase or decrease our confidence in such hypotheses. Put differently, the aspiration of Process Tracing is to minimize the inferential error we risk making when producing statements about an ontological causal reality.

The backward perspective takes advantage of the fact that, at the time of the investigation, the mechanism has presumably had enough time to leave traces which provide a strong indication of its existence. Process Tracing recognizes that not all of these traces are equally informative, and as a consequence focuses on assessing the quality, strength, power, or probative value that select pieces of evidence hold in support of (or against) the causal mechanism.

One of its advantages is that it allows a clear distinction between 'absence of evidence', which has little inferential power and does not add much value to what the researcher already knows, and 'evidence of absence', which on the contrary can strongly challenge a hypothesis, if it contradicts observable implications stemming from such a hypothesis.

In Process Tracing, four well-known metaphors are often used to describe the different ways evidence affects our confidence about a certain mechanism or Theory of Change: the Hoop test, the Smoking Gun test,[1] the Straw-in-the-Wind test and the Doubly-Decisive test (Bennett, 2010; Van Evera, 1997). See Box 1 for the properties of these tests.

Box 1. Differences and similarities among the four Process Tracing tests.

Smoking Gun *(confirmatory)*: If the evidence is observed, the hypothesis is confirmed. If the evidence is not observed, the hypothesis is not confirmed; but it is not rejected, either.
Hoop Test *(disconfirmatory)*: If the evidence is not observed, the hypothesis is rejected. If the evidence is observed, the hypothesis is not rejected (it 'goes through the hoop', passes the test); but it is not confirmed, either.
Doubly Decisive *(both confirmatory and disconfirmatory)*: If the evidence is observed, the hypothesis is confirmed. If the evidence is not observed, the hypothesis is rejected.
Straw-in-the-Wind *(neither confirmatory nor disconfirmatory)*: If the evidence is observed, this is not sufficient to confirm the hypothesis. If the evidence is not observed, this is not sufficient to reject the hypothesis.

One possibility offered by Process Tracing, attempted in social science (Fairfield and Charman, 2015) and law (Edwards, 1986; Friedman, 1986; Kaye, 1986) but so far mostly untapped in evaluation, is its combination with a rigorous mathematical formalization. While the concepts of Process Tracing can be modelled with different mathematical concepts and tools[2] one branch of mathematics we find very useful in connection with the method is Bayesian Updating (see also Beach and Pedersen, 2013; Befani and Mayne, 2014; Bennett, 2008, 2014) which we would like to refer to, specifically, as 'Bayesian Confidence Updating'. In this formalization of Process Tracing, the inferential power or probative value of a piece of evidence E for a theory T can be measured in a number of ways (Bennett, 2014; Friedman,

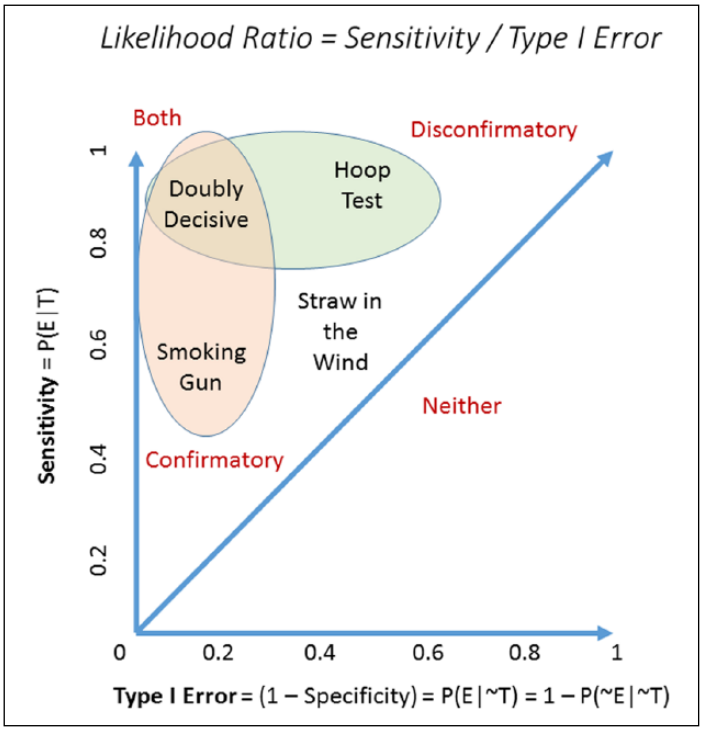

Figure 1. Adapted from Humphreys and Jacobs (2015).

1986; Kaye, 1986), all related to the difference between the true positives rate or 'sensitivity' (the probability that the evidence confirms that the theory holds when this is in fact the case) and the false positives rate or 'Type I error' (the probability that the evidence confirms that the theory holds when this is actually not the case). The larger the difference between the true positives rate and the false positives rate, the higher the probative value of evidence E for theory T. This is shown in Figure 1 (above): the tests with the highest probative value are furthest from the diagonal 'Sensitivity = Type I error' line.

Intuitively, this means that if an observed piece of evidence has a higher chance of being observed if theory T holds true (sensitivity), than if theory T does not hold true (Type I error), this constitutes a confirmation of the theory. If the opposite is true, and the evidence has a higher chance of being observed if the theory does not hold, compared with if the theory holds, observation of that evidence weakens the theory. Finally, if the evidence has a similar chance of being observed whether the theory holds or not (sensitivity is roughly the same as Type I error), observing it will not significantly alter our confidence in the theory.

In Bayesian Confidence Updating, different pieces of evidence have different values of sensitivity and specificity, hence different likelihood ratios, and thus different abilities to alter the evaluator's initial confidence in the contribution claim. The evaluator is thus forced to be transparent about their assumptions and confidence on the existence of the claim, and to 'declare' its observable implications ('if the claim holds true – or does not – what should I expect to observe? With what probability?'). Making these assumptions (which, by other methods, are mostly left out or at best are left implicit) transparent means making them open

to challenge; if no major objections are offered, this will increase their legitimacy and credibility. Just like in a judicial trial where evidence is produced in favour or against a defendant and the jury is left to assess the probative value of that evidence, if the prosecution cannot produce any significant evidence of guilt or if the defence finds proof that the suspect is innocent, then the suspect is considered innocent by the jury.[3]

The building blocks of the 'Contribution Tracing' approach

The approach we illustrate starts with the formulation of a 'contribution claim' about the role an intervention (or parts of it) might have had in the achievement of an outcome. This stage is followed by the symbolic launch of a 'contribution trial', where evidence is gathered in support or against the contribution claim.

When a case is brought before a court of law, the burden of proof principle dictates that the evidence presented must prove, beyond reasonable doubt, the guilt of the accused. The evaluator, thankfully, is not determining someone's fate in a court of law, rather the likelihood that a particular contribution claim holds true. There are two types of evidence the evaluator is particularly interested in: evidence they would 'expect-to-see', under the hypothesis that the contribution claim holds, and evidence they would 'love-to-see', if they want to prove that it does.

'Expect-to-see' evidence are the observations we expect to make under the assumption that the contribution claim holds true; our confidence in the claim changes significantly only if – after having looked carefully – we fail to observe it.[4] This is the same logic followed during a murder trial when evaluating the evidence about a specific suspect A, who is assumed guilty of murdering victim B. Our claim in this example becomes 'suspect A killed victim B'; this premise triggering our search for 'expect-to-see' evidence. If A really did kill B, we expect, for example, the suspect not to have a solid alibi which places them far from the physical vicinity of the victim near the time of the crime. If suspect A does not have a robust alibi, this might strengthen our confidence in the contribution claim, but not to the point where we declare them guilty and close the case. On the other hand, if evidence is found that suspect A was on another continent during the exact time of the murder, the assumption that they are guilty is ruled out and our claim rejected.

Put differently, expect-to-see evidence has disconfirmatory but not confirmatory power. It can be identified by answering two related questions. First, 'what evidence do we expect to find if the contribution claim holds?' Second, 'what would prove, beyond reasonable doubt that the contribution claim does *not* hold?' In the above-mentioned case, in response to the first of our questions, we would expect suspect A not to have a robust alibi. Finding such evidence strengthens our confidence in the contribution claim but the absence of an alibi does not in itself infer guilt. Turning to the second of our questions, finding evidence that the suspect does have a water-tight alibi, immediately rules out the contribution claim and our focus would shift to other plausible causes, or in our example, other suspects.

In Process Tracing, observing 'expect-to-see' evidence is known as 'passing the Hoop test': the contribution claim needs to 'pass through the hoop' if it is to be retained as a possibility. In this sense 'expect-to-see' evidence is necessary, but not sufficient to prove the contribution claim.[5]

'Love-to-see' evidence, as the title suggests, is the Holy Grail for evaluators. It is evidence that has the power to confirm the contribution claim beyond reasonable doubt; but it is unfortunately harder to find, because most pieces of evidence are compatible with different, sometimes opposite, claims and explanations. In most cases we are not so lucky to observe 'love-to-see' evidence, but when we do, our confidence in the contribution claim is greatly strengthened.

The two, inter-related questions we ask in order to identify 'love-to-see' evidence are: 'what evidence is not compatible with any other explanation or causal claim'? and 'what would prove, beyond reasonable doubt that the contribution claim holds?' In Process Tracing, observing 'love-to-see' evidence is known as 'passing the Smoking Gun test'; it is considered akin to catching the murder suspect with a smoking gun in their hand, in the vicinity of the victim. In such a scenario, it would be reasonable to link this observation with the murder, because the sight of a person with a smoking gun – especially standing over a dead body – is extremely difficult to explain unless that person has killed the victim. While alternative explanations are possible, such as the person holding the gun claiming they had only picked it up and fired it to chase away the real killer; this scenario is usually far less credible, unless very special (rare) evidence is found to substantiate the suspect's remarks.

Observing 'love-to-see' evidence provides strong confirmation for our contribution claim; but failing to observe it does not weaken the claim. If we fail to observe the suspect with a smoking gun in their hand in the vicinity of the victim, it does not mean that they are not guilty; it might simply mean that they have been quick to escape from the crime scene. In this sense 'love-to-see' evidence is sufficient, but not necessary, to prove the contribution claim.[6]

The reason 'love-to-see' evidence strongly increases our confidence in the contribution claim is that it is hard to imagine that specific evidence being observed, unless the contribution claim holds (unless the intervention has actually made that specific contribution in reality). In other words, 'love-to-see' (or Smoking Gun) evidence is very specific to the assumption that we are trying to prove. This is why it is usually rare: because there are not many reasons that can account for it. In Bayesian terms, the likelihood of that evidence being observed if the theory holds is much higher than its likelihood if the theory does not hold.

Imagine the scenario of a top-ranking politician stating on record that their dramatic policy shift was due to a particular NGO-led public campaign. In some contexts, it might be hard to conceive of any other plausible scenario that would lead to this pronouncement. Unless evidence of special motives is found (such as privileged treatment of that NGO by the politician, secret agreements, willingness to hide an uncomfortable real reason, etc.) the hypothesis that the shift was due to the campaign appears the most likely explanation of that observation. Put differently, this kind of statement is rare and there are not many reasons that can account for it; and if no evidence is found of alternative reasons, this can be a good example of 'love-to-see' evidence, strongly confirming the contribution claim.

For every trial, a specific type of evidence usually exists which, if found, is enough to condemn the accused. For example, a non-coerced admission of guilt is a relatively rare event, and it is usually difficult to think of other reasons the suspect would confess, other than lifting a burden off their guilty conscience. But even if the suspect does not confess, it is usually possible to conceive of decisive, framing evidence. Not rarely such evidence, if available, is easily recognizable and the subject of manipulation, concealment or destruction (also known as 'withholding evidence'). The corollary in Contribution Tracing is knowing what evidence to look for and understanding its importance in relation to the claim under assessment.

Finally, some pieces of evidence (even rarer than 'love-to-see') can be considered, at the same time, necessary and sufficient to prove the contribution claim: they can both confirm or increase our confidence in the claim (if observed); and disconfirm or decrease our confidence in it (if not observed). Watching the murder taking place in a piece of CCTV footage, where the killer can be clearly recognized, can both condemn suspect A, if we see them in the footage, and exonerate suspect A, if we see someone else in there. Similarly, a DNA test can both

condemn and exonerate a specific suspect, depending on whether the DNA sample matches their DNA or not. In Process Tracing language, these are known as 'Doubly-Decisive' tests (see Box 1).

Application: Evaluating the Universal Health Care Campaign in Ghana

Let us now leave the court room and consider a real case based on an evaluation which applied Process Tracing principles, conducted on behalf of Oxfam GB (Stedman-Bryce, 2013).

The Universal Health Care Campaign in Ghana was a civil society-led movement that aimed to use the upcoming 2012 presidential election as a window of opportunity to promote their health-related policy priorities. The campaign had significant reach comprising over 500 health-focused civil society organizations, who when taken together, were operational in all regions of Ghana. The ultimate goal of the campaign was to have the Government of Ghana legislate for free universal health care for all. While the ultimate goal had not been achieved at the time of the evaluation (and remains to be achieved at the time of writing), the campaign did claim a significant victory which is the subject of this example.

In the course of the campaign a research report was produced and widely disseminated that highlighted a number of alleged shortcomings in the country's National Health Insurance Scheme (NHIS). The scheme, administered by the National Health Insurance Authority (NHIA) – a Government department – is for all Ghanaians, who, with some exceptions, are eligible to receive health care assuming they pay an annual membership fee. The main contention between the campaign and the NHIA related to the number of people registered on the scheme: the NHIA claimed 67 per cent of all Ghanaians were registered, and hence eligible for free health care, while the campaign claimed only 18 per cent were registered.

At the crux of the debate was the methodology used by the NHIA to calculate membership of the scheme. The campaign claimed the methodology was flawed, while the NHIA countered that the campaign had inaccurate information and stood by their methodology and calculations. Remarkably, several months after the publication of the campaign's report, and in a context of vociferous debate between the campaign and the NHIA, each defending their position, the NHIA announced it would be changing its methodology for calculating membership in the scheme, citing methodological inaccuracies. This led the NHIA to reduce its coverage figures from 67 per cent to 34 per cent. The campaign claimed that their report and lobbying efforts contributed to this turnaround by the NHIA – to change its methodology and revise its coverage figures. The evaluation thus sought to assess the following contribution claim: 'the campaign affected the NHIA's decision to revise the methodology for calculating the membership of the scheme'.

Testing the contribution claim

Let us now show how the contribution claim – 'The advocacy campaign affected the Government's decision to revise the methodology for calculating membership of the National Health Insurance Scheme (NHIS)' – was tested. The first step was to formulate both our expectations and 'dreams': or in other words, what evidence we expect-to-see under the hypothesis that the claim holds true; and what evidence we would love-to-see, because these would either strongly confirm or disconfirm the contribution claim.

Table 1. Description of evidence expected to be observed under the hypothesis that the contribution claim holds.

Expectation One	At least partial congruence between the revised methodology and the suggestions made by the campaign.
Expectation Two	The revision to the methodology to happen sometime AFTER the campaign published its report.
Expectation Three	The campaign and its report to have sufficient reach or to be targeted in a way that the Government could have, at least potentially, access to the report.
Expectation Four	The majority of the stakeholders responsible for the campaign (who have an incentive to say it has been successful) believe in the contribution claim.

Table 1 describes the expectations held by the evaluators in terms of evidence to be observed if the contribution claim holds true. For each expectation, we now discuss why observing that specific evidence, while strengthening our confidence in the contribution claim, is not enough to prove influence, while failing to observe it strongly weakens the claim (that the campaign had an influence).

- Partial congruence between the revised methodology and the suggestions made by the campaign might be a coincidence: the changes might have been inspired by knowledge unrelated to the campaign. This is why observing evidence described in Expectation One does not confirm that influence has taken place. However, failing to observe at least partial congruence between the suggestions and the implemented changes makes it much more difficult to argue that influence has taken place. If this is the case, we would observe some congruence.
- In Expectation Two: causal transformations take place in a temporal sequence, the cause happening before the effect. If the changes took place before the campaign even published its report, it is impossible to argue that the NHIA were influenced by something that had not yet been released! Unless the Government had access to draft reports, which in the specific situation appeared very unlikely. Clearly, the mere observation that the changes take place after the campaign is not enough to prove influence: many other events took place after the campaign that are obviously unrelated to it.
- As for Expectation Three: if the campaign were not strongly picked up by the media and did not have wide resonance, and no evidence was found that the Government had access to the report through other channels, an argument could be made that the Government was oblivious to it and could not have been influenced in any way by it. This expectation was largely met when the Government explicitly reacted against the campaign and the findings of its report, which proved that they had had access to its content. However, just because it appears uncontroversial that the NHIA had access to the report, it does not necessarily mean that it was influenced by it. In theory, having access to the report might have made the Government more aware of potential dangers and encouraged it to avoid what was suggested, doing exactly the opposite! So mere exposure to the campaign is not sufficient to prove influence.
- Finally, Expectation Four stems from the fact that stakeholders responsible for the campaign have strong incentives to prove that they are being successful at what they do; so

> if the campaign is actually successful, we would expect them, if not to brag about it, at least to share their positive view of it. Of course, just because stakeholders responsible for the campaign declare it a success, it does not prove that this is actually the case (precisely because they have career-related incentives to say so which are unrelated to whether the campaign was actually successful or not. If we had access to an independent view, say of someone whose career is unrelated to the success of the campaign, we would have different expectations).

Notice that, if observing E is a Hoop test for a contribution claim, not observing E becomes a Smoking Gun test for the negation of the claim (Humphreys and Jacobs, 2015). In this case, if we fail to observe expectation four and most of the above stakeholders admit that the campaign failed, this becomes strong evidence that the campaign did fail; indeed, it might be difficult for the evaluator to think of reasons different from the campaign actually failing to explain these accounts. It would be difficult to imagine a situation where the campaign has been clearly successful and yet individuals, who have an incentive to show that this is the case, deny it. Why would they say that? Failing to observe Hoop test evidence for an influence claim thus constitutes Smoking Gun evidence that the intervention did not have an influence.

In other words, the above would constitute Smoking Gun evidence that the contribution claim does *not* hold. On the other hand, among the evidence we would 'love-to-see', because it greatly increases our confidence that the contribution claim *does* hold, being rare under alternative circumstances, we can include:

1. Admission of influence on behalf of the NHIA in a public statement; and
2. The NHIA using exactly the same formula suggested in the report to revise its methodology.

Both of the above were observed during the evaluation, and both would be very unlikely unless the NHIA was influenced by the report. In theory, it is possible to find alternative explanations. For example, for number one, we could construe that the NHIA considered the organizations responsible for the report as allies and wanted to support them, even though the report had not influenced the reform substantially. This assumption, however, was strongly weakened by evidence of the obvious tension and strong language the NHIA used against the campaign organizations to refute their report's findings and in directly attacking the credibility of those organizations within the campaign. Similarly, in number two above, the NHIA might have been pursuing independent research in parallel to the campaign, leading exactly to the same recommendation. This would have been unlikely because the recommendation was very specific (a mathematical formula); in any case, no evidence of previous research was found whatsoever.

Absence of evidence is not equivalent to evidence of absence; we might not find evidence simply because we do not look for it! However, if we show that we have looked hard and have not found anything, we are leaning more towards evidence of absence than absence of evidence. In the example we are illustrating, the specific attitude of the NHIA during the campaign makes it hard to believe that it was undertaking parallel research to decide how to improve the methodology. If the NHIA wanted to downplay the influence of the campaign, they could have taken the opportunity to claim during the campaign that they were working on something similar. If they had evidence that the campaign did not add any value to their

process, they could have produced it. Instead, they attacked the campaign's report as a whole, only to do some of what the report suggested a few months later.

The strong language used by the NHIA in attacking the campaign also suggests that they were uncomfortable being highlighted and influenced in this way by civil society organizations, least of all Oxfam GB who had a role in supporting the campaign (these attitudes are to an extent captured in the proverbs 'the lady doth protest too much' and 'he that blames would buy'). In short, the NHIA was unable to prove that they were not influenced by the campaign, even though it appeared that they would have liked to, judging from the tension during the campaign. Then during a WHO/World Bank Ministerial meeting, the official Ghana delegation, on public record, admitted the influence of the campaign and how the intervention had actually been useful.

While observing either of the two 'love to see' items listed above strongly increases our confidence that the campaign influenced the NHIA's revision of the methodology, failing to observe either does not weaken the contribution claim. The campaign could still have had an influence, even if the NHIA did not admit it publicly. Similarly, the report might have influenced the new methodology only partially, which would have resulted in the new methodology not being identical to the one eventually adopted. Finally, in Bayesian terms we can claim that, in general, the probability of the Government admitting to be influenced by an NGO if they actually are is much higher than if they are not; and that the probability of the formula used by the Government being identical to the one used in the report 'by chance', or for reasons other than the campaign exercising its policy influence, is in general quite low; and specifically much lower than the probability of the authority using the same formula in case of influence.

The above strategies help prove a specific contribution claim; but events (or outcomes) rarely have one single cause. For example, the upcoming elections might have provided additional pressure to be accountable to citizens and civil society, which might have provided the final push for the NHIA to adopt the new methodology. In addition, the NHIA might have been finally convinced by the methodological soundness of the technical recommendation, realizing it was superior, which might not have happened if the suggestion were not as clear cut or more 'political' than 'technical'.

A number of modified, more complex contribution claims, taking account of other influencing factors besides the intervention, can be developed and tested using the same method described above, gradually incorporating and consolidating evidence found for the single factors (Befani and Mayne, 2014; Befani et al., 2016).

Measuring the strength of the evidence with Bayesian probability

Throughout this article we have used terms such as 'unlikely', 'confidence', 'chances', and so on because the goal of this approach is not to assess impact directly but to assess our 'confidence' that the intervention had an impact (Befani et al., 2016). Other authors (Beach and Pedersen, 2013; Bennett, 2008, 2014; Schmitt and Beach, 2015) have observed how the principles of Process Tracing can be fruitfully combined with Bayesian probability to quantify the probative value of specific pieces of evidence, or their power to change our pre-observation confidence that a specific contribution claim holds.

In Bayesian terms, before starting data collection, we can quantify our 'prior' confidence about the contribution claim (that we could, for example, do in Steps 1 to 4 of a Contribution Analysis (see Befani and Mayne, 2014) into the probability that the claim is true and

represent it as P(CC) (probability of the contribution claim being true). If we have no prior information about the claim and essentially no reason to believe that it is more valid than not or less valid than not, our prior confidence will be 0.5, which is known as the 'no information' situation in Bayesian statistics (Fairfield and Charman, 2015; Piccinato, 2009).

The role of data collection is then to 'update' our confidence in the contribution claim, increasing it or decreasing it compared to the prior, pre-observation situation. Following the Bayes formula, our 'post-observation' or 'posterior' confidence in the claim (represented as the conditional probability of the claim given that evidence E has been observed) is obtained as follows:

$$P(CC|E) = P(CC) * P(E|CC) / P(E)$$

While the above is the most popular representation of the Bayes formula, an alternative sees the probability of evidence E, denoted as P(E), being represented in a way that explicitly shows the importance of the sensitivity P(E|CC) and Type I error P(E|~CC)[7]:

$$P(CC|E) = P(CC) * P(E|CC) / \left[P(CC)*P(E|CC) + P(\sim CC)*P(E|\sim CC)\right]$$

The right side of the formula shows that the power of the evidence to change our prior confidence P(CC) depends on the 'sensitivity' or true positives rate: in particular, the evidence has higher power to change our prior confidence if the sensitivity is high. This can be associated with the Process Tracing Hoop Test: if the sensitivity (P(E|CC)), or our expectation of observing a specific piece of evidence under the contribution claim, is high, it means that the Hoop test is strong, and failing to observe E drastically reduces our confidence in the claim.

The same side of the formula also shows that the probative value of evidence E is high when its Type I error (the false positives rate), denoted as P(E|~CC), is low. The link with Process Tracing is that Smoking Gun, confirmatory evidence is highly specific to the claim (has high specificity or high true negatives rate), because it is highly unlikely under alternatives to the claim. Another advantage of this formalization is that it shows the link between probative value and the likelihood ratio 'sensitivity/Type I error', formalizing the statement we made early on about how the probative value increases as the difference between sensitivity and Type I error increases. Figure 1 illustrates the links between sensitivity, specificity, Type I error, the likelihood ratio and Process Tracing tests. Notice how the probative value/conclusiveness of tests increases as we move away from the diagonal, 'sensitivity = Type I error' line.

While not everyone is an enthusiast about quantifying confidence and probabilities in qualitative research (Fairfield and Charman, 2015), approaching Process Tracing in this way is useful to overcome a dilemma which is often encountered in practice: the uncertainty around classifying specific pieces of evidence as smoking gun, straw-in-the-wind, or hoop tests. Figure 1 should help clarify how in straw-in-the-wind evidence, sensitivity is not very high and Type I error is not very low; while in more conclusive evidence we either have a sensitivity which is close to one or a Type I error which is close to zero. The Doubly-Decisive tests describe the fortunate situations where we have both. Many practitioners are now attempting to classify evidence under these tests without formally or informally referencing such probabilities, while we argue that such a classification is only defensible after having made statements, either qualitative or quantitative, about sensitivity and Type I error.

In order to illustrate how the Bayes formula can be applied, consider the example of our favourite piece of evidence from above: the admission of influence on behalf of the

Government. Let's assume that we have no prior information on how likely the intervention is to have influenced the NHIA's decision to revise the formula in that particular way: this will set the prior probability of the contribution claim at 0.5.

We now need to assess the power of the Government's admission to change our initial confidence. In order to do this, we need to estimate the sensitivity and Type I error of this piece of evidence for our claim. Since precise estimation of these probabilities is not always straightforward, we use three different scenarios to help us navigate intervals of possibilities: the standard, conservative and super-conservative scenarios.

In our standard scenario, the Type I error is set at 0.01: a Government admitting to be influenced by an NGO while this is actually not the case can be considered an extremely rare event, happening in only 1 per cent of situations where Governments take decisions without being influenced by NGOs. At the same time, the probability that the Government admits influence when having been influenced (the sensitivity) is higher, although the event is not very frequent either because Government in general tend to be resistant to admitting influence of non-governmental entities. Our standard estimate for this context is 0.2: which means that there is a 20 per cent chance that Government admit influence when they have actually been influenced.

In this scenario, observing the Government admit influence increases our confidence in the contribution claim from 0.5 to 0.95[8] (Table 2).

One typical critique of this approach is that the post-observation confidence in the contribution claim is over-dependent on the estimates of the probability of observing that evidence under different hypotheses, which might not be considered reliable if there is not enough quality data available. The good news is that the robustness of the approach (or the sensitivity of the findings to these probabilities) can be tested *while applying* the approach. Table 2 presents alternative estimates of the post-observation confidence under different scenarios and different values of sensitivity and Type I error.

Under a more conservative estimate, where we think Governments would tend to praise NGOs more frequently, even without being influenced by them, the Type I error P(E|~CC) is higher and set at 0.05; in the super-conservative estimate this value reaches 0.1, which means that Governments for some reason praise NGOs for one out of 10 policy decisions they make, even if they are not influenced by these organizations.

Similarly, if we are to make more cautious and uncertain assumptions on the sensitivity of the evidence, we can assume that it's more unlikely for Governments to admit influence even when they are influenced, and set the sensitivity P(E|CC) at a lower 0.15 in the conservative and at 0.1 in the super-conservative scenarios.

Table 2 reports the values of the post-observation confidence under the three scenarios. In the conservative estimate, the evidence still increases our confidence in the contribution claim (from 0.5 to 0.75), but less than in our standard scenario. In the extreme, super-conservative scenario, our prior confidence is left completely unchanged by the Government's admission. Even if we do not trust the values we used to reach these conclusions, the formula allows us to elicit the implicit assumptions we make when making such statements on the weakness or inconclusiveness of the evidence. Saying that the evidence is useless requires the strong assumption that the Government is equally likely to make such statements, whether they have been influenced or not. But if we believe that this assumption does not make any sense and that Governments will tend to admit influence more easily if such influence is real than if it is not, then the admission does increase our confidence in the contribution claim. A comparison can be made, again, with admissions of guilt on behalf of suspects: this is usually taken as

Table 2. Probabilities needed to estimate the updated, post-observation confidence in our hypothesis of influence.

	Standard	Conservative	Super-conservative
SENSITIVITY: Probability of the Government admitting to be influenced by an NGOs under the hypothesis of influence taking place – P(E|CC)	0.2	0.15	0.1
TYPE I ERROR: Probability of the Government admitting to be influenced by an NGOs under the hypothesis of no influence taking place – P(E|~CC)	0.01	0.05	0.1
Prior Confidence in the hypothesis of influence – P(CC)	0.5	0.5	0.5
Post-observation Confidence in hypothesis of influence: P(CC|E)	0.95	0.75	0.5

Table 3. Qualitative rubrics describing different quantitative levels of confidence.

Practical Certainty	0.99+
Reasonable Certainty	0.95–0.99
High Confidence	0.85–0.95
Cautious Confidence	0.70–0.85
More Confident than not	0.50–0.70
No information	0.50

evidence of guilt when the assumption that suspects have the same tendency to confess whether they are guilty or not is not considered credible.

The point of this exercise is not to estimate quantities with a high degree of precision, but to illustrate how the formalization of sensitivity and Type I error values allows a reasonably accurate assessment of the strength of given pieces of evidence for a claim; or at least allows us to uncover the extreme assumptions we would need to make in order to argue that the evidence is completely uninformative.

If evaluators or stakeholders struggle to use the language of probability, we can also use qualitative rubrics to evaluate our confidence as illustrated in Table 3. Using qualitative rubrics, we can express our assumptions above as follows:

1. We have no prior information about the chances that the intervention has had an influence – P(CC) = 0.5;
2. We are cautiously confident (0.80, standard), or highly confident (0.90, super-conservative) that the Government would not admit to be influenced by an NGO, if this were the case; and
3. We are practically certain (0.99, standard), or reasonably certain (0.95, conservative), or highly confident (0.90, super-conservative), that the Government would not declare to have been influenced by an NGO, if this were not actually the case.

Finally, our conclusions would read:

- The evidence does not alter our confidence in the contribution claim in any way (super-conservative scenario, P(CC|E) = 0.5)

- The evidence makes us cautiously confident in the validity of the contribution claim (conservative scenario, P(CC|E) = 0.75)
- The evidence makes us reasonably certain of the validity of the contribution claim (standard scenario, P(CC|E) = 0.95+)

Even just a small difference between the expectations of observing that specific evidence under the two hypotheses (influence/no influence) can increase our confidence about the truth of the contribution claim. Using the Bayes formula to elicit our implicit assumptions and make our expectations transparent is not just a pedagogical or sophisticated formalization: it allows other stakeholders to check the robustness of our reasoning, to make alternative proposals, to see the implications of considering the evidence irrelevant or uninformative; and ultimately increases the rigour, robustness and credibility of the evaluation.

How to apply the approach: Practical considerations and the 'contribution trial'

The theories of change we encounter in our daily practice as evaluators are often not amenable to the derivation of testable implications. One of the benefits of the approach is to refine theories of change in such a way that their statements are much more limited and precise, which is often the only way to connect the theories with testable implications. For example, the initial contribution claim of the Health Care Campaign was a much more generic 'the campaign has had significant policy influence'; it was through the examination of evidence that the claim was made more specific and connected to the revision of the formula to calculate health system coverage. The evaluator discovered that many other interesting claims were simply not supported by the available evidence, and decided to focus on what could be proven. Using a more forgiving approach that would have tolerated looser linkages between evidence and contribution claims might have failed to refine the claim in such a precise way. In other words, contribution tracing has a low tolerance for vaguely expressed, unfalsifiable theories of change and untestable claims, and its application increases the conceptual precision, clarity and scientific quality of theories of change.

We have made use of the criminal case metaphor given many people's exposure to legal processes and procedures via televised crime dramas and the prevalence and popularity of crime fiction generally. This helps root the application of Process Tracing tests in examples which are familiar to most.

Building on this, we introduce the concept of a 'contribution trial' whereby evaluators can assess their evidence against the principles of what they 'expect-to-see' and what they would 'love-to-see'. We propose that – in order to minimize confirmation bias – this should be done in collaboration with key stakeholders and include 'critical friends' to the intervention under investigation; as well as actors representing other plausible influencing sources.

The 'contribution trial' is a constructive conversation focused around the assessment of the probative value of given pieces of evidence for given claims. The purpose of the 'contribution trial' is to reach agreement, through dialogue, about the most compelling evidence the evaluation should look for and what bearing, if found, such evidence would have on our confidence about the contribution claim.

This approach presumes that the 'contribution trial' is used as a collaborative process to aid primary data-collection design, and to inform appropriate selection of data-collection tools.

It raises the likelihood that the evaluator asks the 'right' questions of the 'right' people and looks in the 'right' places with the most appropriate tools.

The judgment on whether the evidence is compelling or not, and on how compelling or strong a given piece of evidence is for a claim, is something that can be measured transparently and potentially agreed upon by a 'jury of judges', ideally representing 'stakes' around different contribution claims (in order to reduce confirmation bias). In order to facilitate agreement, it is important to formalize three probability assessments, representing the evaluator's expectation of (or confidence in):

1. The contribution claim being true (pre-observation of a specific piece of evidence, or simply 'prior');
2. The given piece of evidence being observed if the contribution claim holds (the 'sensitivity'); and
3. The given piece of evidence being observed if the contribution claim does not hold (the 'Type I error').

The 'jury' can expose their reasons during the course of a symbolic trial, and either broadly agree on the assessments of the three levels of confidence (for example following the above proposed rubrics) or propose different, alternative assessments. The judgments can be used to create extreme, opposite scenarios that would formalize the extent of disagreement. The ultimate outcome of these trials, the one directly useful for the evaluation, would be lower and upper limits of post-data collection confidence in the contribution claim.

If the evaluator formulates a series of different contribution claims and collects a number of pieces of evidence, different matrices can be created that link each observation (e.g. reported in the columns) with each claim (e.g. reported in the rows) on the basis of their sensitivity, Type I error, likelihood ratios, prior and posterior confidence, probative value, and so on. There should be a different matrix for each of these measures. In order to minimize bias and conflicts, the trial should not be conducted on the matrices indicating the confidence or the probative value directly, but only on those illustrating sensitivity and Type I error.[9] When the 'jury' has validated the latter two measurements, the evaluator can compute the posterior and other measures, thus spotting at a glance which pieces of evidence have the highest probative value for each claim; and more generally which claims are most strongly supported.

Concluding remarks

Qualitative methods for impact evaluation have traditionally been considered second-best and inferior to quantitative methods because of their lower internal validity and replicability. We believe that Contribution Tracing, an approach focused on the replicable and transparent testing of mechanisms and qualitative statements, has the potential to contribute to qualitative evaluation methods being taken more seriously for two main reasons.

The first is that the questions guiding data collection directly address the core reason we conduct data collection, which is to increase (or decrease) our confidence in a hypothesis. Every piece of evidence considered is assessed in terms of its ability to alter this confidence, and different Process Tracing tests can be used for this purpose, allowing the evaluator to fruitfully and transparently navigate a high number of hypotheses and a high number of pieces of evidence at the same time. This will allow evaluators to gradually connect claims with pieces

of evidence and obtain a shortlist of a few, plausible claims which are strongly supported (as with inductive Process Tracing: see Beach and Pedersen, 2013; Bennett and Checkel, 2014b). This process can be made fully transparent, illustrating why the evidence for the eliminated hypotheses was weak, and why it was stronger for the hypotheses that have been retained.

The second reason is that, similarly to what happens in traditional quantitative and statistical methods, confidence can be measured, both with probability and qualitative rubrics. The measurements and qualitative assessments of confidence can be fully shared with a 'jury' of experts, stakeholders, or fellow evaluators, who can either validate or refine the assessments. The process is replicable and will produce confidence intervals describing the power of given pieces of evidence to change our confidence in the contribution claim (e.g. from 0.75 to 0.95 as in the Table 2 above), increasing the robustness and ultimately the 'objectivity' of qualitative evaluation findings.[10]

As we anticipated at the beginning of this article, we are still not providing a fully-fledged textbook guidance on the approach, which would both articulate its epistemological premises, and offer step-by-step, application-relevant indications. A more robust defence of the approach would need to build more explicitly on the application of Bayesian Updating in other fields (law, medical diagnosis, geology, archaeology, social science), and answer important practical questions on how to estimate the required probabilities, including for example those relating to 'packages' of pieces of evidence.

Nonetheless, we hope that – as we continue to test the approach on our real-life evaluations – the guidance provided so far is sufficient to influence the practice of at least some other evaluators, contributing to making data collection more transparent, systematic and ultimately efficient. We hope to encourage and enable evaluators to act as 'evaluation knowledge translators', supporting others with responsibility for the design and planning of development intervention; for example, by assisting in the downstream design of development interventions' evaluation M&E frameworks based on a comprehensive understanding of what evidence will be sought upstream as impact evaluation gets underway. Greater understanding of the probative value of items of evidence in relation to specific contribution claims can focus M&E frameworks and plans towards attaching greater importance to collecting data with higher probative value over the lifetime of the intervention, in the context of an impact evaluation upstream.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. Notice that the Hoop test for the existence of a mechanism is the Smoking Gun test for its non-existence (Bennett, 2014; Humphreys and Jacobs, 2015). If failing to observe E rules out theory T (hence T fails the Hoop test), observing 'absence of E' confirms that theory T does not exist or that 'non T' exists. In other words, 'absence of E' is a Smoking Gun test for the non-existence of T. Similarly, if observing evidence F strongly confirms (is a Smoking Gun test for) theory S, not observing F means that the alternatives to S cannot be ruled out, or in other words the alternatives to S pass the Hoop Test.
2. For example, Set Theory and Directed Acyclic Graphs (DAGs or multitrees).

3. The application of Bayesian analysis to law court cases has been questioned (Brilmayer, 1986); however, the matter is far from being settled and the vast majority of the participants of the symposium maintained that the advantages outweigh the disadvantages.
4. We need to look carefully if we want to avoid confusing evidence of absence with absence of evidence!
5. The language of necessity and sufficiency is used for illustrative purposes and is not to be intended as if we are embracing a deterministic perspective: set theory incorporates the possibility of 'fuzziness' which is compatible with the realm of probability and 'degrees of confidence'.
6. See caveats above on fuzzy set theory.
7. $P(E) = P(E \cap CC) + P(E \cap \sim CC) = P(CC)*P(E|CC) + P(\sim CC)*P(E|\sim CC)$
8. $P(CC|E) = 0.5 * 0.2 / (0.5 * 0.2 + 0.5 * 0.01) = 0.1 / (0.1 + 0.005) = 0.1 / 0.105 = 0.952381$
9. And prior confidence, if different values of it are used.
10. The term 'objectivity' is used because the claims would have been agreed upon and validated by a jury

References

Beach D and Pedersen R (2011) *What is Process-Tracing Actually Tracing? The Three Variants of Process Tracing Methods and Their Uses and Limitations*. APSA 2011 Annual Meeting Paper. Available at: http://ssrn.com/abstract=1902082

Beach D and Pedersen R (2013) *Process-Tracing Methods: Foundations and Guidelines*. Ann Arbor, MI: University of Michigan Press.

Befani B and Mayne J (2014) Process Tracing and Contribution Analysis: A combined approach to generative causal inference for impact evaluation. *IDS Bulletin* 45(6): 17–36.

Befani B, D'Errico S, Booker F and Giuliani A (2016) *Clearing the Fog: New Tools for Improving the Credibility of Impact Claims*. London: International Institute for Environment and Development.

Befani B, Ledermann S and Sager F (2007) Realistic Evaluation and QCA: Conceptual parallels and an empirical application. *Evaluation* 13(2): 171–92.

Bennett A (2008) Process Tracing: A Bayesian perspective. In: Box-Steffensmeier J, Brady H and Collier D. *The Oxford Handbook of Political Methodology*. Oxford University Press.

Bennett A (2010) Process Tracing and Causal Inference. In: Brady H and Collier D. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. New York: Rowman and Littlefield.

Bennett A (2014) Appendix: Disciplining our conjectures. Systematizing process tracing with Bayesian analysis. In: *Process Tracing: From Metaphor to Analytic Tool*. Cambridge: Cambridge University Press.

Bennett A and Checkel J (2014a) Introduction: Process tracing: From philosophical roots to best practices. In: *Process Tracing: From Metaphor to Analytic Tool*. Cambridge: Cambridge University Press.

Bennett A and Checkel J (eds) (2014b) *Process Tracing: From Metaphor to Analytic Tool*. Cambridge: Cambridge University Press.

Bjurulf S, Vedung E and Larsson C (2012) A triangulation approach to impact evaluation. *Evaluation* 19(1): 56–73.

Brilmayer L (1986) Second-order evidence and Bayesian logic. *Boston University Law Review* 66: 673–91.

Byrne D (2013) Evaluating complex social interventions in a complex world. *Evaluation* 19(3): 217–28.

Collier D (2011) Understanding Process Tracing. *Political Science and Politics* 44(4): 823–30.

Derwisch S and Löwe P (2015) Systems dynamics modelling in industrial development evaluation. *IDS Bulletin* 46(1): 44–57.

Edwards W (1986) Summing up: The Society of Bayesian Trial Lawyers. *Boston University Law Review* 66: 937–41.

Fairfield T and Charman A (2015) Applying formal Bayesian analysis to qualitative case research: An empirical example, implications, and caveats. *SSRN*. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2647184.

Frey K and Widmer T (2011) Revising Swiss policies: The influence of efficiency analyses. *American Journal of Evaluation* 32(4): 494–517.

Friedman R (1986) A close look at probative value. *Boston University Law Review* 66: 733–59.

Grove J (2015) Aiming for utility in 'Systems-based Evaluation': A research-based framework for practitioners. *IDS Bulletin* 46(1): 58–70.

Humphreys M and Jacobs A (2015) Mixing methods: A bayesian approach. *American Political Science Review* 109(4): 653–673.

Kaye D (1986) Quantifying probative value. *Boston University Law Review* 66: 761–6.

Lemire S, Bohni Nielsen S and Dybdal L (2012) Making contribution analysis work: A practical framework for handling influencing factors and alternative explanations. *Evaluation* 18(3): 294–309.

Mayne J (2001) Addressing attribution through contribution analysis: Using performance measures sensibly. *The Canadian Journal of Program Evaluation* 16(1): 1–24.

Mayne J (2008) *Contribution Analysis: An Approach to Exploring Cause and Effect*. Rome: Institutional Learning and Change (ILAC) Initiative (CGIAR).

Mayne J (2012) Contribution analysis: Coming of age? *Evaluation* 18(3): 270–80.

Mohr L (1999) The qualitative method of impact analysis. *American Journal of Evaluation* 20(1): 69–84.

Patton MQ (2008) Advocacy impact evaluation. *Journal of Multidisciplinary Evaluation* 5(9): 1–10.

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

Piccinato L (2009) *Metodi per le decisioni statistiche*. Milan: Springer-Verlag.

Schmitt J and Beach D (2015) The contribution of process tracing to theory-based evaluations of complex aid instruments. *Evaluation* 21(4): 429–47.

Scriven M (2008) A summative evaluation of RCT methodology & an alternative approach to causal research. *Journal of MultiDisciplinary Evaluation* 5(9): 11–24.

Stedman-Bryce G (2013) *Health For All: Towards Universal Health Care in Ghana. End of Campaign Evaluation Report*. Oxford: Oxfam Great Britain.

Ton G (2012) The mixing of methods: A three-step process for improving rigour in impact evaluations. *Evaluation* 18(1): 5–25.

Van Evera S (1997) *Guide to Methods for Students of Political Science*. Ithaca, NY: Cornell University Press.

Westhorp G (2014) *Realist Evaluation: An Introduction*, London: Overseas Development Institute (ODI).

Williams B (2015) Prosaic or profound? The adoption of systems ideas by impact evaluation. *IDS Bulletin* 46(1): 7–16.

Williams B and Hummelbrunner R (2010) *Systems Concepts in Action: A Practicioner's Toolkit*. Stanford, CA: Stanford University Press.

Barbara Befani, Research Fellow at the University of Surrey and Research Associate at the University of East Anglia, UK. She has spent 10 years adapting and tweaking methods for evaluation, in the last 5 focusing on impact evaluation, causal inference, QCA and methodological appropriateness.

Gavin Stedman-Bryce, Director of Pamoja Evaluation Services Ltd, UK evaluation professional with background in public health; specializing in advocacy and policy influencing impact evaluation with a particular focus on applying innovative approaches, including process tracing.