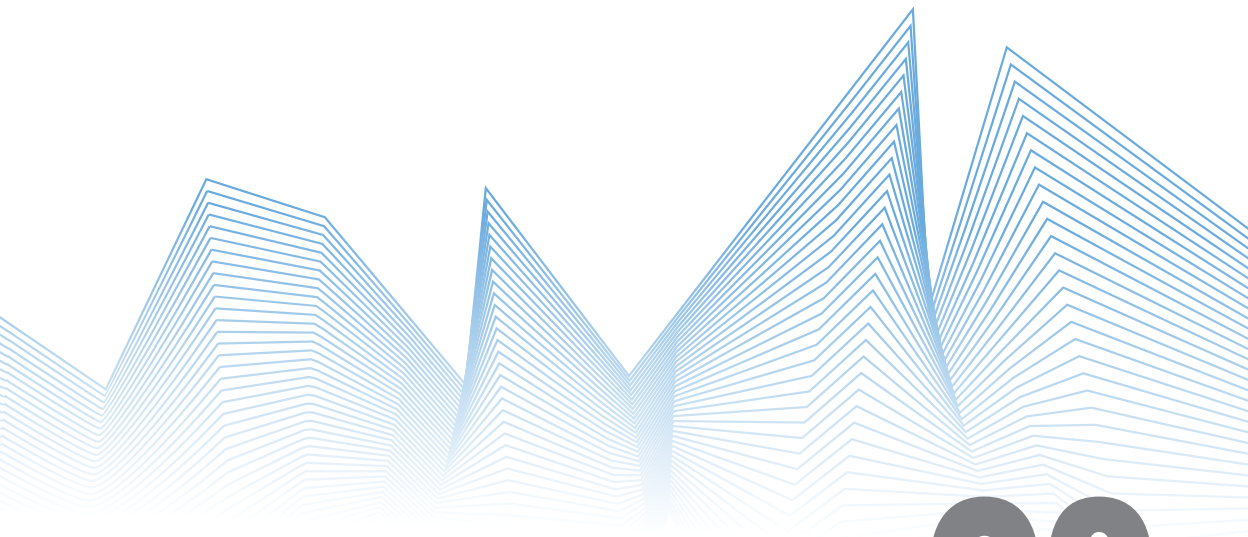
EBA



03
2021

CREDIBLE EXPLANATIONS OF DEVELOPMENT OUTCOMES: IMPROVING QUALITY AND RIGOUR WITH BAYESIAN THEORY-BASED EVALUATION

Barbara Befani

Credible Explanations of Development Outcomes: Improving Quality and Rigour with Bayesian Theory-Based Evaluation

Barbara Befani

Report 2021:03

to

The Expert Group for Aid Studies (EBA)

Dr. **Barbara Befani** specialises in developing and adapting innovative methodologies for evaluation. Currently Visiting Fellow at the University of Surrey and working mostly as an independent researcher/consultant, she's a former Research Fellow of the Institute of Development Studies, the University of Surrey, and a former Research Associate of the University of East Anglia. Barbara has fifteen years of experience in evaluation methodology research, training, advisory work, and community building. Her main fields of expertise are hybrid, quali-quanti methods like QCA and (Bayesian) Process Tracing; the theory and practice of appropriate methodological choice; and unified quality assessment systems (that equally apply to qualitative, quantitative, and mixed-methods evaluations).

Please refer to the present report as: Befani, B (2021), *Credible Explanations of Development Outcomes: Improving Quality and Rigour with Bayesian Theory- Based Evaluation*, EBA Report 2021:03, The Expert Group for Aid Studies (EBA), Sweden.

This report can be downloaded free of charge at www.eba.se

ISBN 978-91-88143-67-9
Printed by Elanders Sverige AB
Stockholm 2021

Cover design by Julia Demchenko

Acknowledgements

My Bayesian journey in evaluation began in 2013 when I was reading Daniel Kahneman's work on cognitive biases. I knew the Bayes formula from my university days and the lectures of Ludovico Piccinato (who had been a pupil of Bruno De Finetti); but it was after reading Kahneman that I could intuitively see the Bayes theorem's potential in evaluation. My ideas were very confused at the time but I hung a sketch of the Confusion Matrix on the wall at my office at the Institute for Development Studies because I knew that was the conceptual "kernel" on which the methodological developments needed to be grounded on. An incredible journey started that left me indebted to many illustrious colleagues. First, John Mayne who recognised the commonalities with Contribution Analysis. Then, Gavin Stedman-Bryce who allowed me to rework what, to my knowledge, was the first application of Process Tracing to development evaluation (pioneered by the brilliant Claire Hutchings at Oxfam) and imagine what the Bayesian version of PT could look like in that context. Shortly after, Stefano D'Errico of the International Institute for Environment and Development decided to invest in Bayesian innovation and with the precious help of Francesca Booker we applied the method to an evaluation of policy influence. The basic tenets of our method were already developed at that point but more needed to be developed regarding qualitative descriptors of confidence associated to ranges and facilitating confidence updating under various circumstances. The Centre for the Evaluation of Complexity Across the Nexus (CECAN) allowed me to develop the excel tool and Chris Rees of Risk Solutions introduced me to the literature on the elicitation of expert judgement. Laura Camden first gave me the opportunity to teach the method and see that its learning curve was different from other methods; as well as that workshop participants tended to

get confused between hoop tests and smoking guns which further convinced me that the Bayesian formalisation was needed. At the time, various actors tried to apply the method but not everyone persevered long enough to overcome its initial challenges; Kirsty Wilson of LTS International was one such person and she allowed me to discover the risks and sensitivities of seeking information that would normally be confidential. Mary Anderson and Charles Michaelis were two other such colleagues, and together we discovered and developed another piece of the methodological puzzle. Corinna Elsenbroich and Jennifer Badham helped develop and crystallise another complex idea which was initially just an intuition: using the Bayes theorem to assess evidence in support of given parameters of "theories" expressed as computer-based simulation models; while Frederico Brandao has contributed his immense knowledge of forestry policy to prove that Bayes had something to offer to that sector, too. Over the years, I've had enlightening and constructive discussions with Zoe Sutherland of Itad; Johannes Schmitt of the German Institute of Development Evaluation; and again, with Claire Hutchings of OPM who is playing an enormous role in communicating the method's potential and possibilities to commissioners.

Last but not least, I am immensely grateful to EBA for the opportunity to write this report; and in particular to Markus Burman who was equally patient and unwavering in his support for the project through a series of setbacks and delays mostly due to my personal circumstances. The reference group "dream team" he put together was the best I could ask for. I was incredibly pleased to finally work with Gustav Petersson after a long wait; and to once again benefit from Rick Davies's unique insights, after many years of brilliant and wide-ranging methodological discussions. I am humbled that Derek Beach, whose textbook made me discover Process Tracing in 2013, and Nancy Cartwright,

who has revolutionised development evaluation thinking in the last few years, have accepted to review my work; their insights have been precious and improved the report considerably. Finally, I am grateful to Naïma El Hawary for her help in proofreading the report. I take full responsibility for all the errors, omissions, typos and other shortcomings.

In loving memory of my mum, Rosella Santorelli, who enjoyed teaching maths to children.

Table of Contents

Foreword by EBA 1

Sammanfattning 3

Summary 5

1. Introduction: why read this report? 8

1.1 Quality in research and evaluation 15

1.2 TBE: potential and typical weaknesses 18

2. What is Diagnostic/Bayesian TBE? 22

2.1 Theory 24

2.2. Empirical observations 25

2.3 Confidence in the truth of the theory 25

2.4 Quantitative Confidence and the Bayes formula 28

2.5 Expressing confidence qualitatively 45

2.6 Metaphors and typologies: the Process Tracing tests 51

3. How to apply Bayesian TBE in practice 55

3.1 Step One: Developing a testable theory 56

3.2 Step Two: Identifying diagnostic tests and designing data collection 68

3.3 Step Three: Estimating the Bayes formula values and updating confidence 99

4. Concluding remarks 128

References 133

Appendix: What a commissioner should know 140

Previous EBA reports 143

Foreword by EBA

Establishing whether development programmes lead to development results is one of the most important tasks in evaluation. At the same time, it is arguably the most difficult task. In addition to find whether an observed effect was caused or not by an intervention, it is also of importance to learn for whom, under what circumstances, how and why the intervention made a difference, not to mention the study of unforeseen positive or negative effects caused by the intervention.

In this report, Barbara Befani presents a method that combines advantages from both qualitative and quantitative impact evaluation methods. Bayesian Theory-Based Evaluation can draw from and explain outcomes on single case studies, while using traceable, repeatable operations and explicit confidence levels, which increase the reliability of the findings.

The report contains an accessible introduction to the method, an in-depth theoretical discussion of the method's epistemological tenets, and a comprehensive section with practical applications for those who want to apply the method.

Evaluation methodology is a rapidly developing field, generating numerous new methods for impact evaluation presented as articles in scientific journals and used by the international evaluation community. With this report, EBA aims to contribute to increased use of such methods for impact evaluation in the Swedish community, a scarcity evidenced e.g. in EBA report 2021:02.

I hope this report will find an audience among commissioners of evaluations, evaluators, researchers, and persons with an interest in methodology and impact assessments.

The study has been conducted with support from a reference group chaired by Janet Vähämäki, member of the Expert Group. The author is solely responsible for the content of the report.

Gothenburg, september 2021

Helena Lindholm

Sammanfattning

Denna rapport introducerar en ny innovativ metod för teoribaserad utvärdering av insatser inom bistånd och andra politikområden. Teoribaserade utvärderingar syftar oftast till att förklara resultat snarare än att som experimentella metoder mäta nettoeffekter. De är inriktade på lärande snarare än ansvarsutkrävande eller ja/nej-beslut om insatser ska fortsätta, avbrytas eller skalas upp. Detta fokus på lärande fångas av utvärderingsfrågor som ”hur och varför har interventionen fungerat/inte fungerat”, ”kommer den fungera i andra sammanhang eller i framtiden”, ”för vem fungerar insatsen, för vem inte”. De förändringsteorier som analyseras med metoden återfinns på mikro- och mesonivå, vilket innebär att de främst relaterar till enskilda fall.

Metoden (som också är känd som Bayesian Process Tracing, Contribution Tracing eller Process Tracing with Bayesian Updating och Diagnostic Evaluation) har uppmärksammats i den internationella utvärderingsdiskussionen då den kombinerar styrkor från etablerade kvalitativa och kvantitativa metoder. Som de förra kan den förklara resultat och arbetar med enskilda fall. Som de senare kan metoden anses vara ”rigorös” i betydelsen att dess processer är spår- och reproducerbara vilket oftast ökar tillförlitligheten i slutsatser. Att slutsatser kopplas till uttryckliga konfidensnivåer (som vid statistisk analys) bidrar till att ökad trovärdighet i utvärderingsresultat.

Metoden har utvecklats under ett antal år och viss litteratur finns tillgänglig. Det finns dock ett behov att anpassa litteraturen till olika behov och att sprida lärdomar från en hittills ganska begränsad användning.

Rapporten innehåller en fördjupad teoretisk diskussion om metodens kunskapsteoretiska principer och hur dessa förhåller sig till utvärderingspraktiken, något jag tror tilltalar metodspecialister. Rapporten har också ett utförligt avsnitt med praktiska tillämpningar

som jag tror kommer tilltala utvärderare, forskare och konsulter som vill tillämpa metoden i praktiken. Detta avsnitt bygger på sex fall (inom områdena policypåverkan, skogspolitik, energi och hälsa) och som används för att illustrera och exemplifiera metodens olika steg. Dessa beskrivs som en linjär process men i praktiken handlar det om iteration fram och tillbaka mellan tre steg där teorin succesivt förfinas och ytterligare belägg införlivas i analysen.

De tre primära tillämpningsstegen för metoden är: 1) utveckla en testbar teori ("testbar" betyder här detaljerad och nära kopplad till empiriska observationer), 2) utforma datainsamlingen runt tester som övertygande försvagar eller stärker teorin, 3) bedöm bevisvärde eller "styrka" utifrån dessa och gör konfidensuppdatering. Tidigare teoribaserade utvärderingsmetoder har inriktats på det första steget. De bedömer inte explicit förtroendet för teorin eller styrkan i de belägg som anförs. Inte heller andra förändringsteoretiska utvärderingsmetoder försöker identifiera denna typ av "ideala" belägg för teorin.

Att använda Bayes sats i utvärdering kräver skattning av sannolikheter. Arbetet underlättas dock genom användande av kvalitativa konfidensnivåer utifrån sannolikhetsintervall. Användaren behöver endast bedöma konfidensnivån som exempelvis "mycket säker", "praktiskt taget säker", "mer säker än inte", och teknisk expertis i utvärderingsteamet kan sedan genomföra själva uppdateringen. I rapporten diskuteras flera sätt att åstadkomma dessa skattningar och olika sätt att hantera aggregering när man har olika empiriska observationer att ta hänsyn till i analysen.

Bilaga 1 i rapporten har tagits fram för beställare av utvärderingar. Den varnar för att metoden både är relativt avancerad och att få utvärderare idag har erfarenhet av den. Det poängteras dock att metoden inte kräver mer resurser än traditionella fallstudier och att det snarare handlar om att utvärderare gör samma saker något annorlunda än att göra helt nya och andra saker. Det handlar mer om ett förändrat tankesätt vid utformning av datainsamling och analys än om i grunden förändrade moment.

Summary

This report presents an innovative methodology to conduct theory-based evaluations (TBEs) of development interventions as well as programmes relevant to other policy sectors. TBEs usually aim at explaining development outcomes rather than measuring net effects like experimental methods do; they are oriented towards learning and improvement rather than accountability and yes/no decisions on whether to abolish, suspend, continue, or scale up programmes. The learning element is captured by evaluation questions to the effect of "how and why has the intervention worked or not"; "will the intervention work in other contexts or in the future", or "who is the intervention working and not working for". The theories covered by this method are medium-low level, which means they relate to single case studies.

The method (which is also known as Bayesian Process Tracing, Contribution Tracing, or Process Tracing with Bayesian Updating, and Diagnostic Evaluation) is receiving increasing attention from the international (development) evaluation community because it retains several advantages of both qualitative and quantitative methods: like the former, it can explain outcomes and work on single case studies; like the latter, it is "rigorous" or in other words, its operations are traceable, repeatable and usually increase the robustness/ reliability of the findings. In addition to this, the fact that statements in the findings are associated with explicit levels of confidence (as in statistics) contributes to greater credibility of evaluation results.

While the method has been in development for a few years and there is a limited body of literature available, there is a large need of adapting this literature for evaluation purposes and of sharing the lessons learned from the small number of applications that have so far been carried out in real life evaluations. This guide aims to accomplish both and – after an in-depth theoretical discussion of the epistemological tenets and how they relate to evaluation, which is most likely to appeal to methods specialists – the report includes a

section on the practical applications in policy evaluations that is likely to appeal to evaluators, researchers, and consultants who want to apply the method in practice. This section builds on six applications (policy influence in development, forestry policy, energy policy, and public health) that are used to illustrate and exemplify the practical steps that need to be taken. The sequence is described as a linear process but in practice there will be iteration and back-and-forth between the three steps, as the theory is refined on the basis of evidence and as additional evidence is incorporated.

The three application steps are the following: a) developing a testable theory (where "testable" means detailed and closely connected to empirical observations); b) designing data collection around tests that conclusively weaken and / or strengthen the theory; c) formally assessing the probative value or "strength" of the tests and carry out formal confidence updating. Typically, TBE methodologies will focus on the first step but fail to carry out the second and third; in other words, they don't formally measure confidence in the theory nor assess the strength of the evidence in support or against it. Doing so would require the design of strong tests, for example imagining observing (or not observing) particular content in documents or interviews that would substantially weaken or strengthen the theory. Traditional TBE methods do not explicitly identify this kind of "ideal" evidence and thus cannot directly seek it. Elsewhere this evidence has been described from the perspective of someone who wishes to confirm the theory as "love-to-see" (the "smoking gun" kind that would confirm the theory upon observation) and "hate-not-to-see" (the "hoop test" kind that would rule out the theory if not observed).

Using the Bayes formula requires the estimation of probabilities but this can be circumvented by working with qualitative levels of confidence associated to probability ranges; the user would just need to assess their confidence level in terms of, for example, "highly confident", "practically certain", "barely more confident than not", etc. and the technical expert would input the assessments into the

updating procedure. We discuss several ways to produce these estimates as well as several ways to deal with the aggregation of multiple observations.

The annex written for commissioners warns about the possible challenges such as the relatively steep learning curve and the current (at the time of writing) relative rarity of personnel who possess knowledge and experience of the method; but mostly aims at reassuring commissioners that the method will not require substantially larger resources than traditional case studies and it does not require stakeholders to do different things, but rather to do the same things differently. It aims at reframing the way evaluators think about designing data collection and analysis for TBEs, not at substantially changing these activities. But since it's a relatively new way of thinking and thrives on detail and documentation, it will occasionally require that special attention is devoted to building or strengthening trust between evaluators who collect and analyse data, and stakeholders who are supposed to provide information and documentation.

1. Introduction: why read this report?

CHAPTER SUMMARY: In this chapter we explain why diagnostic (Bayesian) theory-based evaluation, the method presented in the report, is a useful addition to the evaluator's toolbox. We frame the argument in terms of quality improvement and bias reduction; we offer that it helps make sense of the evidence for a given theory when it's most difficult to understand what data or other empirical observations reveal about a theory (for example when most of the evidence seems to be weak or when evidence is scarce or when the topic is sensitive and the findings are at high risk of being biased). More generally, applying Bayesian Updating to TBE helps "shield" TBE from the accusations of low reliability or robustness that are sometimes directed to it by evaluators who prefer quantitative methods, because of improvements gained on transparency of the analytical process, reliability (or robustness / rigour), and ultimately credibility of the evaluation findings. The chapter outlines the structure of the report with the aim of making it more readable and discusses the audiences the report has been written for.

Development interventions are subject to increasing scrutiny from both an accountability and a learning perspective, both in Sweden[1] and within the international community. Accountability is about enquiring whether an intervention reached the expected targets; while learning covers a broader range of knowledge gaps, which are not always known upfront. The "learning" evaluation questions are why an intervention worked or didn't, if it's likely to work under different circumstances and in the future, under what circumstances does it work better, for whom, why, and so on.[2] According to the

[1] https://www.sida.se/English/how-we-work/evaluation/

[2] This is consistent with SIDA's evaluation approach, according to which evaluation can contribute to "Learning about what works for whom, under what circumstances and how" and "Accountability by providing transparency in

learning perspective (Hummelbrunner, 2015), the main goals of the evaluation are to understand how and why an intervention had the consequences it had; with a view to improving it and possibly replicating it in different contexts and sectors (Stern, et al., 2012).

The movement towards quality and rigour in evaluation (Savedoff, Levine, & Birdsall, 2006) often tends to favour quantitative approaches because of their relatively high performance on some quality standards, like replicability, robustness, and internal validity/credibility (Gertler, Martinez, Premand, Rawlings, & Vermeerch, 2011). However, quantitative methods present a double limitation: 1) they usually present strict/multiple requirements and can only be applied under limited circumstances; and 2) they are not always satisfactory in terms of which specific evaluation questions they are able to answer, focusing on quantitative values or net effects, rather than explanatory or contextualised mechanisms that deepen understanding (Befani, 2016; Stern, 2015; Befani, 2020).

Theory-based (impact) evaluation is now a widely accepted group of approaches for development interventions (AA.VV., 2017; White, 2009; Schmitt, 2020), providing the capacity to answer how and why questions, explain outcomes, and clarify the role played by the intervention (and which parts thereof) in achieving outcomes. But, at the same time, TBE mostly draws on qualitative methods, and is hence generally subject to validity and reliability challenges to a larger extent that quantitative methods are. An EBA commissioned report (Befani, 2016) discussed a qualitative method, Qualitative Comparative Analysis or QCA, that attempts to improve on these quality dimensions (as well as transparency and construct validity). However, unlike the method presented in this report, QCA is not applicable to single case studies, nor to situations in which only a handful of cases are available.

Swedish development cooperation.". https://www.sida.se/English/how-we-work/evaluation/.

We steer clear of establishing or implying any methodological hierarchy and support methodological equality (or at least the idea that all methods should be treated equally, because they present different analytical strengths and weaknesses and they can all be used to low or high quality standards (Clarke, Gillies, Illari, Russo, & Williamson, 2014; Befani, 2020). However, we are concerned that TBE evaluations are often criticised on the transparency, credibility, and reliability of their findings (Befani, 2020; Befani & D'Errico, 2020) and our aim with this paper is to propose potential solutions that can improve their quality; with the proposed solutions being applicable, this time, to single case studies. While important quality dimensions (Tsang, 2014), in this report we do not directly address transferability or generalisation because we do not believe this method necessarily brings improvements on this front compared to more traditional case studies; these are the dimensions that QCA improves upon, but Bayesian/diagnostic evaluation has other comparative advantages. Indeed, QCA and Bayesian logic are often seen as complementary (Beach, 2018; Schneider & Rohlfing, 2013). Notice that – while Bayesian TBE does not improve generalisability or transferability compared to more traditional TBE, it does so compared to methods geared towards measuring net effects without delving into the reasons behind intervention success or lack thereof (Cartwright & Hardie, 2012).

Since the method does not require a fundamentally different set of resources from traditional case studies (quoting from the executive summary: it's about doing things differently, not doing different things), but presents advantages in terms of transparency, credibility, and reliability, we suggest that it is used whenever the evaluation context requires a higher level of the latter three quality dimensions: broadly speaking, the same arguments that favour RCTs and quasi-experiments in the eyes of commissioners and decision makers (credibility and bias reduction) should favour diagnostic or Bayesian theory-based evaluation when the questions of interest (how and why did the intervention work or not) are different from those experimental evaluation is able to answer (how much and to what

extent did the intervention work or make a difference). Sometimes these requirements arise when the programme to be evaluated is controversial or high-stake, but in general, particularly in development, interventions are increasingly subject to greater accountability requiring more rigorous scientific standards as we discussed at the beginning.

We struggled to choose a specific audience for this report because we thought it was important to convey messages to both commissioners and evaluators; to both specialists and non-specialists; and to evaluators versed in qualitative methods as well as quantitative methods, which the method presented is supposed to bridge. It has not been easy to calibrate the language and some audiences might struggle with certain parts of the report; but hopefully, on the bright side, every type of audience will find something that speaks to them. Below we provide suggestions on which sections might be more relevant for which audience.

The executive summary and the annex for commissioners are aimed at a broad audience of non-specialists, while the introduction has been written with evaluators in mind who have been active in the community at least for the last 10 years. The second chapter is aimed at evaluation "philosophers" and those who want to dig deep into the theoretical and epistemological foundations of methodology; but evaluators who ultimately want to apply the method in practice should also find it useful that some conceptual aspects are clarified. The third chapter is the practical/applied chapter and is aimed at those evaluators who want to "get their hands dirty" with the method and do the whole thing in their evaluation practice. For this reason, at times the report becomes highly technical, because some practical actions require at least limited knowledge of probability theory and its mathematical underpinnings. But we've been careful to always consider the perspective of the average evaluator and have included alternative solutions to most of the technical procedures (see the translator rubrics in Section 2 and the strategy for assembling of multiple pieces of evidence in the absence of Bayesian estimates in Section 3.3.4).

We start by arguing that, in typical qualitative evaluation practice, the connection between empirical data and mechanisms or theories is often weak, reprising and enriching some of the introductory arguments of the other report (Befani, 2016); and secondly, we present an approach, suitable for single case studies, that makes this connection clearer, more explicit, and in some cases even "automatic", with, we believe, substantial quality improvements on transparency, credibility/internal validity, and robustness. We do not believe applying this method is the only way to improve on these criteria, we just argue that it does bring improvements on these fronts compared to traditional case studies and the way theory-based evaluations are commonly conducted.

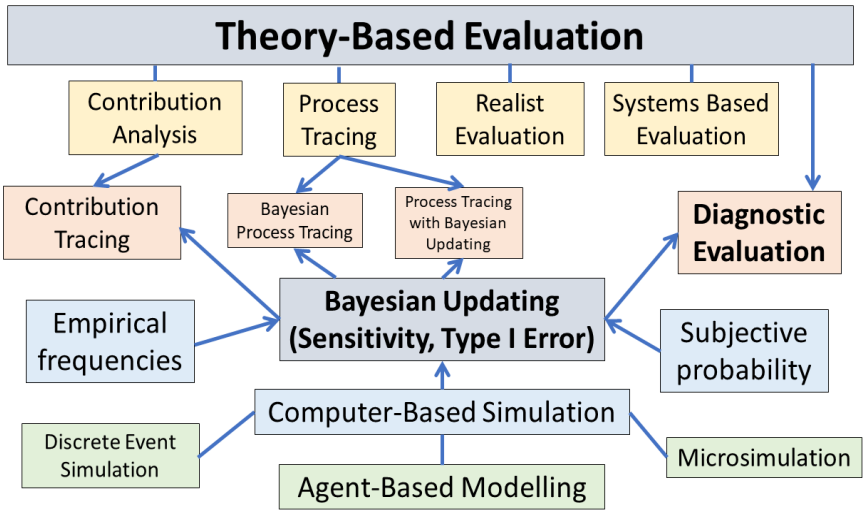
The improvements are discussed in detail for two types of common bias: confirmation bias and conservatism. We do not argue that applying this method eliminates all bias; the only kind of bias it eliminates completely is conservatism (which is directly linked to the lack of use of the Bayes formula (Kahneman, 2012) and then we suggest that it contributes to a potentially great reduction of confirmation bias. But the main point we want to convey is that bias must be acknowledged and the conditions must be created for bias to emerge; and we believe this method, besides virtually eliminating conservatism, creates favourable conditions for confirmation bias to be detected; and even suggests how specific "red flags" can be spotted that should alert the evaluator to the presence of bias (section 2.4.1.1).

The report does not address why evaluators should use theory-based evaluation, and mostly takes TBE's relevance and usefulness for granted. We are also not overly prescriptive in terms of which theories should be used, only recommending the criterion of "testability" (see introduction of section 3.1), which tackles one very specific aspect of theory-based evaluations and theories of change: their connection with empirical observations. The diagram below (Figure 1) should help clarify the intent and content of the report.

The upper grey box represents the vast world of Theory-Based Evaluation, which comprises the above-mentioned variants in yellow (Contribution Analysis, Process Tracing, Realist Evaluation, Systems-Based Evaluation). The grey BU box represents formal Bayesian Updating, with its focus on the estimation of Sensitivity and Type I error for given observations and theories (and theory components). Various labels have been used to connect forms of TBE with the logic and tools of Bayesian Updating (boxes in orange): *Contribution Tracing* meant to bridge Contribution Analysis and Bayesian Updating (Befani & Stedman-Bryce, 2017; Befani & Mayne, 2014); *Bayesian Process Tracing* (Fairfield & Charman, 2017) connects Process Tracing and Bayesian Updating, exactly like *Process Tracing with Bayesian Updating*. But the ambition of Diagnostic or Bayesian Theory-Based Evaluation (Befani, 2020) is to connect all forms of TBE with Bayesian Updating, by arguing that the Bayes formula does not discriminate on how the *event* the probability of which we are trying to estimate is represented; and also to move beyond formal Bayesianism by providing qualitative tools and strategies that build upon Bayesianism but do not formally include it (Befani, 2020) (see section 3.4).

So, for example, *Diagnostic Evaluation* is different from *Bayesian Process Tracing* in that the theory does not necessarily take the form of a causal process mechanism; but it's similar because they both use the likelihood ratio and some values of the Bayes formula. The bottom part of the chart indicates that all forms of Bayesian Updating require the estimation of probabilities, in particular of Sensitivity and Type I error; and that there are three main strategies to satisfy this requirement: empirical frequencies, elicitation of expert judgement (or subjective probability), and computer-based simulation. As it turns out, not all forms of CBS can be used for this purpose, but only a handful like discrete event simulation, microsimulation and Agent-Based Modelling (Befani, Elsenbroich, & Badham, 2021).

Figure 1: Connections among various approaches and methods mentioned in the report



The fundamental link with probability or confidence estimation implies that a transparent process behind the estimation of probabilities or the corresponding qualitative confidence levels is required. Possibly, this is the area where the method's learning curve is steeper and the practice that is the most unusual for evaluators. While acknowledging the challenges of both elicitation and synthesis, the report aims at providing defensible solutions for the practicalities of both, endorsing in particular an adaptation of the SHELF method for the elicitation of expert judgements.[3]

Our overall ambition is that this report becomes a tool for both an in-depth understanding of diagnostic approaches and Bayesian Updating, and a How-To-Guide that provides detailed indications to evaluators who want to apply it. It also covers commissioner's concerns such as conditions required for a high-quality application, mentioning typical challenges and how to overcome them (see Annex). Its target audience are evaluators wishing to apply Bayesian Updating to theory-based evaluations like for example those using

[3] http://www.tonyohagan.co.uk/shelf/

Process Tracing or Contribution Analysis; and commissioners wishing to gain a better understanding of what this process entails so that they can improve procurement and follow up.

The report is structured as follows. The introductory chapter covers quality and explains which quality dimensions are relevant for this method. It also tackles the idea of theory-based evaluation and its several variants, along with some of the main weaknesses and critiques. The second chapter is theoretical and covers the basic notions of diagnosis in evaluation – what it is useful for and how the Bayes formula relates to it; the relation between qualitative and quantitative confidence, the confusion matrix, how to measure probative value and how to represent and categorise empirical observations, including visually on a chart. There is also a reference to Van Evera's metaphors that have become famous as Process Tracing tests (Smoking Gun, Hoop test, etc). The third chapter discusses the application of the method in practice, using several examples from evaluations of policy influence, forestry policy, energy policy, and public health. It is structured in three parts, one for each application step: developing a testable theory, designing data collection and assessing evidence strength, and updating confidence estimating probabilities. The concluding remarks are followed by an annex written for commissioners presenting the method's benefits from a commissioner's perspective as well as requirements and typical challenges.

1.1 Quality in research and evaluation

Systems of quality assessment in research and evaluation[4] focus on a number of key quality dimensions (Andrew & Halcomb, 2009; Bryman, 2012; DFID, 2014; Lincoln & Guba, 1985; Sale & Brazil, 2004), out of which we can single out at least seven (Befani, 2020):

[4] This paragraph draws on unpublished work conducted by Maren Duvendack.

1. Conceptual Framing
2. Transparency
3. Appropriateness
4. Construct Validity
5. Credibility (Internal Validity)
6. Transferability (External Validity)
7. Reliability (Dependability, Consistency, Robustness)

Qualitative methods tend to be considered comparatively stronger on some criteria, like Conceptual Framing and Construct Validity; and weaker on others, like transparency, credibility, and reliability. Transparency (or repeatability/replicability) refers to the quality of information provided about designs, methods, data collection protocols and techniques; locations, geography and contexts, including possible conflicts of interest; and the criteria used to interpret data and resolve uncertainties on the road to findings. This information is assumed to be critical for the study to be replicated and in a qualitative study it's usually comparatively more difficult to disclose the entire sequence of steps and critical factors (critical in the sense that they affect the findings) leading from data collection to findings; in particular, the leap from empirical observations to theoretical findings is relatively self-referential and difficult to trace. Reliability (or consistency, robustness, dependability) refers to how consistent the findings stay when data collection and analysis are repeated; and if a group of methods struggles on transparency and repeatability, it's hard to consider its findings "stable" because it's simply not possible or very difficult to repeat the process that would inform such judgement. Credibility[5], a.k.a. Internal Validity, or "Truth Value", refers to the extent to which the findings can be trusted. One can argue that it is relatively more difficult to trust the

[5] Credibility of statements / findings / theories also depends on the quality / credibility of the data used to make such statements, but that's equally true for qualitative and quantitative methods so this factor does not penalise one group of methods more than the other.

findings of qualitative research because the researchers do not assign a formal level of confidence to them: unlike statistical inference that uses p-values and 90%, 95%, and 99% confidence levels (linked to sample quality and size and estimated through automatic and repeatable procedures), qualitative findings are presumed to be correct on the basis of loosely structured linkages with data, in a process where confidence is essentially never formally estimated. In addition, for causal statements, credibility is enhanced because confounding factors can be controlled with established protocols (Farrington, Gottfredson, Sherman, & Welsh, 2002; Gertler, Martinez, Premand, Rawlings, & Vermeerch, 2011; Treasury, 2011).

Software for the analysis of qualitative data is commonly used to develop or confirm theory but it still falls short of assigning a formal value to empirical evidence for a given theory or more generally estimating formal confidence levels. This makes it difficult to compare different theories, particularly in those common, ambiguous situations where theories are not mutually exclusive. Qualitative researchers and evaluators use a variety of arguments to support their findings, but there is no recognised standard of good practice linking theoretical statements with empirical data. Evaluation literature on qualitative methods tends to be focused either on theory alone (how theory should be represented, what ontology should be used as a framing device) or data collection and analysis alone (techniques for surveys, focus groups, how to analyse transcripts, textual analysis, etc.). When reading an evaluation report, it is often difficult to find systematic linkages between results from data collection or analysis and level of confidence in the findings. Information that can help enlighten on possible evaluator bias is usually included, and can help identify motivational biases, but it is less likely to help with confirmation bias and certainly doesn't help with conservatism (see section 1.2.1).

1.2 TBE: potential and typical weaknesses

Theory-based evaluation has been a cornerstone of evaluation methods for several decades (Chen, 1990; Pawson & Tilley, 1997; Weiss C. , 1974; Weiss C. H., 1997; Weiss & Connell, 1995). There are many variants of TBE, but all share a focus on the explanation of outcomes: on understanding how and why outcomes are realized. Either how they are supposed to be realised, if the theory is constructed ex ante or before data collection, or how they have actually realised, if the theory is constructed or refined ex post or after data collection.

In Realist Evaluation the central object of analysis is the CMO configuration (Pawson & Tilley, 1997), where mechanisms are interpreted within a realist ontology (Bhaskar, 2009) as dependent on a context. In Contribution Analysis (Mayne, 1999; Mayne, 2008) the "contribution story" explaining how an outcome was achieved is represented as a chain of intermediate outcomes, each coming with assumptions that need to hold and risks that need to be avoided for the next step to be achieved. Systems-based evaluation (Williams, 2015; Williams & Hummelbrunner, 2010; Befani, Ramalingam, & Stern, 2015), while recommending that attention be directed to the interrelationships among parts of complex causal mechanisms and the causal loops within them, do remind us that we need to take perspectives of stakeholders into account and keep an eye on who gets to define system boundaries; but the indications on how to design data collection and how to use data to strengthen or weaken theories are vague.

Indeed, as currently known and practiced, TBE guidelines strongly focus on the content of the mechanism or theory of change, and only to a lesser extent on how to verify or reject it. With the sole exception of Process Tracing, no form of TBE focuses on formally assessing the value of empirical observations for the theory under investigation, in terms of whether evidence strengthens or weakens the theory or neither; nor, if so, by how much. Process Tracing,

however, is relatively simplistic in the way it assesses probative value: as we argue later in the report (Section 2.3) the analyses conducted under this method have three possible outcomes: confirming the theory, rejecting the theory, or neither. Reality is more nuanced than this, which is why we need a more fine-grained diagnostic approach, providing us with the opportunity to *measure* the probative value of empirical observations for given theories. Another reason we need the precision and transparency offered by the Bayes formula is that it protects us against cognitive biases.

1.2.1 How confirmation bias and conservatism affect TBE and qualitative methods

Failing to assess the value of the evidence or assessing it in a way that isn't transparent or convincing, as often happens with TBE and to some extent also with qualitative Process Tracing, particularly in complex circumstances, leads to systematic bias in the interpretation of the evidence and thus fallacious judgement when it comes to revising our beliefs in the light of the evidence. We have mentioned above typical weaknesses in terms of lack of transparency, credibility, and reliability; in this section we focus on cognitive biases, two in particular: confirmation bias and conservatism.

Confirmation bias ("CB" from now on) is defined as the "tendency to search for, interpret, favour, and recall information in a way that confirms one's pre-existing beliefs or hypotheses, while giving disproportionately less consideration to alternative possibilities" (Plous, 1993). It's a cognitive bias related to cognitive dissonance (Elster, 1998); a systematic error of inductive reasoning, and it's stronger for "emotionally charged issues and for deeply entrenched beliefs". In evaluation, we fall victim to confirmation bias when we disproportionally focus our data collection and analysis activities on confirming a given theory we believe or would prefer to be true,

rather than on looking for plausible alternatives or on seeking evidence that might weaken the theory. CB manifests itself in three distinct phases of the thinking (and research) process:

1. Search: gathering or "gravitating towards" information that confirms one's pre-existing or preferred beliefs (or theories).
2. Analysis/interpretation: ambiguous information is interpreted as favouring or supporting one's pre-existing or preferred beliefs (or theories).
3. Memory: selective use of memory biased towards removing uncomfortable recollections (or recollections that might weaken the preferred beliefs or theories).

In evaluation, CB emerges when the theory preferred by the commissioner and/or evaluator receives preferential treatment compared to its alternatives, which are either not considered for testing, or are tested but not as systematically and thoroughly as the former (search CB). When alternatives are tested, only weakening evidence is sought for the latter, while only strengthening evidence is sought for the main or preferred theory.

CB also emerges when ambiguous information that doesn't particularly support the preferred theory more than alternatives is interpreted as favouring the preferred theory (analysis CB). Or when already available knowledge that is supposed to be considered but isn't supporting the preferred theory is omitted or not taken into account (memory CB). In general, CB is undesirable as it can lead to overconfidence and wishful thinking; and can be exacerbated in situations where limited resources are available for an evaluation.

Conservatism or conservatism bias is another systematic bias in the way humans process information. It manifests itself when, revising our beliefs in the light of (new) empirical evidence, we systematically underestimate the strength or probative value of this evidence; compared to the value that would be returned by the Bayes formula (Kahneman, 2012). In other words, we do change our opinions in a way that is proportional to the Bayes formula output; but we do it

insufficiently. Conservatism has been demonstrated in several experiments (Edwards, 1982)[6], and can be considered an extension of the anchoring bias (Tversky & Kahneman , 1974).

In evaluation, we see conservatism when we do not adequately take the evidence into account, and we don't understand its value in support or against our preferred theories. In Bayesian terms (see below), conservatism would imply that the posteriors are never too different from the priors: but explicitly using the Bayes formula ensures that – if this happens – it is for a good, empirically justified reason, and not because the human mind naturally tends to undervalue empirical observations. Applying the Bayes formula, which directly protects against cognitive conservatism, can surprise us at times as to how powerful evidence actually can be and how strongly it's supposed to alter our pre-existing beliefs.

[6] In one famous experiment, participants have been presented with two bags, the first one with 700 red and 300 blue balls, and the second one with 300 red and 700 blue ones; and invited to choose one without knowing which bag it was. They have been invited to sample from the bag randomly, with replacement after each ball. Those who, after 12 samples, got 8 reds and 4 blues (or 66% reds), were asked to estimate the probability that they were sampling from the predominantly red bag, and their average answer was 0.7. The correct value, or what their confidence should have been in that case after observing the evidence, is 0.97 (prior is 0.5, Sensitivity 0.231, T1E 0.008). In other words, they are much more uncertain than they should be and remain "anchored" to the initial / prior probability of having sampled from the predominantly red bag, comparatively ignoring the empirical evidence.

2. What is Diagnostic/Bayesian TBE?

CHAPTER SUMMARY: In this chapter, written mainly for evaluators who want to have a good theoretical grasp of the method, we lay out the bare bones or the conceptual foundations of the method; we only use a simplified example and later flesh out the details and content of theories in chapter three. We explain how theory and findings can be rigorously connected with evidence and empirical observations, using the "confusion matrix" which analyses the four possible states obtained by crossing the two states "theory true/theory not true" with the two states "evidence observed/not observed". All situations can fit into one of the four cells and we encourage evaluators to wonder, for each theory and each (set of) observation(s), which situation we are in. We then argue that it is useful to estimate the probabilities of being into these cells for various pieces of evidence and theories, because that will tell us how strong the evidence is for the theory; and whether the evidence strengthens or weakens the theory. Note that it is possible to describe probabilities qualitatively, and we explain how an evaluator who isn't numerically inclined can express qualitative judgements on the confidence of observing given pieces of evidence under various theoretical assumptions, which can then be "translated" into probability ranges (and vice versa). Finally, we link the four cells to the more famous Process Tracing metaphors for categorizing evidence, which are supposed to communicate how the evidence put into each box is supposed to change (or not change) our mind about the theory: can it strengthen the theory? Can it weaken the theory? Can it do both? Sometimes it can do neither.

In a nutshell, Diagnostic Evaluation is Bayesian Updating applied to Theory-Based Evaluation; and can also be called "Diagnostic Theory-Based Evaluation" or "Bayesian Theory-Based Evaluation" (see Figure 1). Currently, there are applications of Bayesian Updating to Contribution Analysis (Contribution Tracing); to Process Tracing

(Bayesian Process Tracing or Process Tracing with Bayesian Updating); but the message we want to convey here is that Bayesian Updating can be applied to virtually all forms of Theory-Based Evaluation (see Figure 1). We first focus on the diagnostic features of the Bayesian approach and clarify its links with the Confusion Matrix; we won't discuss features of theory-based evaluation which aren't strictly linked to this and to empirical observations. Once the conceptual underpinnings of diagnosis are clear, we show how it can be applied to Theory-Based Evaluation.

Diagnostic Evaluation could equally be named "Bayesian Theory-Based Evaluation", but the term Bayesian evokes numerical reasoning and technicalities that are not necessarily required if one uses qualitative confidence descriptors which are then translated into numerical intervals by a technical expert or facilitator. Another advantage of the term "diagnostic" is that it evokes an unobservable entity and that should resonate with evaluators who often deal with unobservable processes, or at least processes that are unobservable at the time the evaluation takes place.

We start by discussing the three conceptual building blocks of the method (Table 1), inspired by (Bennett & Checkel, 2014): Theory, Observations, and Confidence. We then move on to the Bayes formula details, its features and the opportunities it offers, learning to express confidence quantitatively and visualising our assessments of evidence strength. We draw several parallels with the confusion matrix, which we believe can – despite its unfortunate name – greatly clarify the conceptual underpinnings of the method. We then move on to the qualitative expression of confidence levels and qualitative confidence updating. We conclude the chapter by drawing parallels with the famous Process Tracing metaphors and stressing the differences between diagnostic evaluation and traditional qualitative process tracing.

2.1 Theory[7]

The first core element of the method is a proposition or *statement* about the existence of something. It could be about the impact a programme has had, the role it has played, an illness or unobservable condition, or about anything else that is stated to *exist*. A "Theory" is an ontological object: it might exist or not, it might be true or not. In evaluation theories are sometimes expressed in the form of a contribution claim; or a "mechanism"[8]; or the explanation of an outcome, describing the inner workings that produce it. For example, "the new regulation created a deterrent for farmers and made it more costly for them to deforest". Another form the theory takes is the description of a process or of a system, with varying degrees of complexity: for example, a model depicting the effects of different kinds of vaccination and other behavioural changes on overall infection rates. Anything that is potentially (but, usually, not obviously[9]) *real* belongs in the "theory" category. The problem we face is that such mechanisms, processes, or models are often largely unobservable and there is a certain level of disagreement on what is the exact mechanism, process, or system that has acted to produce the outcome; and over which statements and propositions are true or not. Similarly, when a physician tries to diagnose a patient, they initially don't know which condition they have fallen ill with.

---

[7] Notice that our working definition of "theory" is not limited to causal process mechanism (as one would expect in Bayesian Process Tracing) and expands to configurational theories, behavioural "inner" mechanisms constituting parsimonious explanations of snapshot events happening at a given point in time; complex mechanisms representing adaptive systems, network behaviours, etc. Bayesian Updating is based on probability theory, which handles something as generic as *events*, not something as specific as causal process mechanisms. For a list of "traditional" applications of Process Tracing, see the Process Tracing section in (Vaessen, Lemire, & Befani, 2020)

[8] For a recent overview of causal mechanisms in evaluation, see (Schmitt, 2020) and the whole NDE Special Issue.

[9] If it were obvious that the mechanism or theory existed, we wouldn't need to seek empirical traces or manifestations in order to increase our confidence that it does.

2.2. Empirical observations

The second core element of the method are tangible, observable phenomena: content of documentation, statistical data; recordings, transcripts and minutes offering accounts of what was said during interviews, meetings, and other conversations; databases, media products like photos and videos, drawings, etc. It is usually easier to agree on what is being empirically observed, at least at a literal level, than to agree on which theory, system, process or mechanism is true and lies behind those empirical manifestations; or what is the unobservable ontological entity that is leaving the observed traces. In medical diagnosis, empirical observations would be symptoms and test results: doctors easily agree on symptoms being presented by the same patient; blood samples are analysed by machines and imaging tests are also performed by machines. What is usually more controversial is how empirical observations relate to the underlying condition: which one do they support and how strongly? Conclusive tests are needed to achieve definitive diagnoses. So, we need a third core element that measures the "strength" or probative value of observations for certain theories.

2.3 Confidence in the truth of the theory

The third element of diagnostic evaluation is key to the added value of the approach because it is missing[10] from TBE as currently practiced: formal confidence in the truth of the theory. It's a belief about whether the proposition or statement is true or not, but far from being a yes/no kind of judgement, it's a formalised, declared degree of confidence in the existence of the mechanism or process.

[10] In Process Tracing there is a declared level of confidence when a test is identified as a "smoking gun", "hoop test", or "doubly decisive" vs. a "straw-in-the-wind". However, it's a binary, strong/weak kind of judgement (see also Section 2.3) while the formal Bayesian approach allows a more fine-grained assessment.

Neither unobservable nor tangible (see Table 1), confidence is a "thought in your head" (Bennett & Checkel, 2014), that can be stated and/or calculated; can be expressed with a probability (a number between 0 and 1) or with a qualitative scale (high, cautious, etc. see Table 3) and – if it's not the result of a Bayes formula calculation – is subject to the same cognitive biases that all judgements under uncertainty are subject to (Kahneman, 2012). We mentioned conservatism in particular because it can be eliminated by the Bayes formula, and confirmation bias because its risk seems to be particularly high in the field of policy evaluation. Imagine a commissioner wanting to believe that their intervention had a positive influence (confirmation bias); while struggling to acknowledge the strength of the evidence weakening their preferred claim (conservatism).

Table 1: The three building blocks of diagnostic approaches

	Theory	Confidence	Empirical Observations
Nature	Ontological entity	Human belief	Empirical object
Observability	Unobservable	It is held at the cognitive level ("a thought in your head"; can be formally declared	Observable, tangible
It manifests itself as:	Mechanism, Process, Theory of Change	Can be expressed along a scale of qualitative descriptors; or quantitatively as a probability (0-1)	Datasets, content of documentation, accounts as recorded in meeting minutes, transcripts, media
In medical diagnosis:	The underlying disease	Confidence that the diagnosis is correct	Symptoms, test results
In evaluation:	A proposition or statement e.g., about impact or	Level of confidence that	Timelines, data patterns, content of documentation,

	Theory	Confidence	Empirical Observations
	contribution "Intervention X influenced Policy Y"	the theory is true (or false)	content of interviews

Adapted from Befani (2020)

Why the term "diagnostic"? According to the Oxford dictionary, "diagnosis" is defined as "the identification of the nature of an illness or other problem by examination of the symptoms". This definition includes all the three core elements of the method we are presenting: first, the identification of the "nature" of the illness or problem is the ontological dimension: what the patient *really* has. Second, the "symptoms": that is, the empirical and perceivable/observable manifestations of the condition. Third, the "examination of the symptoms", which is equivalent to the assessment of the strength or probative value of the evidence in order to achieve a conclusive belief on the (ontological) nature of the illness/problem. In other words, the problem or disease leaves perceivable and observable traces through symptoms that the physician is supposed to interpret to make a diagnosis, discovering the true nature of the mysterious object.

If every diagnostic process entails examining empirical observations to identify the ontological nature of the object that is causing or producing the observed phenomena, we can call diagnostic a kind of Theory-Based evaluation that assesses the value of empirical observations to understand which process or mechanism produced the data we collect and analyse.

In the next sections we will cover a key feature of the approach: how confidence is expressed, measured, and altered on the basis of evidence. Namely, we address both quantitative and qualitative approaches to confidence level description and measurement: quantitative in section 2.4 and qualitative in section 2.5.

2.4 Quantitative Confidence and the Bayes formula

The quantitative approach to the expression and measurement of confidence levels is grounded on probability theory. Probability is a number between 0 and 1 that is used to express the likelihood that an event will happen at a given time in the future, based on the number of times it has happened in the past under similar circumstances. For example, when we throw a coin, we know there is roughly 50% probability of getting a heads and the same probability of getting a tails, because if we throw the coin tens or hundreds of times and count the outcomes, the number of times we get heads is roughly equal to the number of times we get tails. Probability distributions show the frequency with which a certain outcome materialises; in our work we will mostly not use distributions and work with the expected, average, or central value. If the event hasn't materialised in the same way in the past or we don't have sufficient empirical data to estimate its probability, we rely on expert assessment or so called "subjective" probability.[11]

[11] A third possibility is to use computer-simulated frequencies as we do in sections 3.1.4 and 3.3.2.

Box 1: Different ways of expressing probability judgements

Probability: "there is a 0.05 probability of observing (this amount of) matching text under such conditions if the influence theory is not true."

Percentage: "there is a 5% chance of observing (this amount of) matching text under such conditions if the influence theory is not true."

Relative frequency: "under such conditions, we would observe (this amount of) matching text one in 20 times the influence theory is not true"; or "we would observe (this amount of) matching text 50 times every 1000 times that the influence theory is not true".

Odds: "if the influence theory is not true, the odds against observing (this amount of) matching text are 19 to 1."

Natural frequency: "from a sample of 100 similar policy processes where two organisations did not influence each other, (the same amount of) matching text between their products was observed 5 times."

Adapted from O'Hagan, et al. (2006)

The literature on eliciting probabilities from experts (O'Hagan, et al., 2006) (Cooke, 1991) (EFSA, 2014) (Gosling, 2014) (Oakley & O' Hagan, 2016) covers elicitation of probability distributions; different ways to express probability judgments and their implications; and how to extract, calibrate, and assemble expert judgments. O'Hagan et al. (O'Hagan, et al., 2006) provide a useful list of different quantitative ways statements about chance and uncertainty can be expressed: as probabilities, percentages, relative frequencies, odds, or natural frequencies. Box 1 illustrates how we would express uncertainty about observing matching text as a piece of evidence while investigating a theory of policy influence.

How do we come up with these numbers? There are several ways of doing it. The simplest is perhaps the evaluator using their own judgement; or the evaluator interacting with a team of colleagues or stakeholders. To be defensible, the estimation needs to be convincing and that will depend on whether the evaluator – with or without a team – is able to build solid arguments in its support. It is possible to use formal procedures that have been tested to elicit judgement from experts: the best-known ones differ mainly on whether and how experts are allowed to interact and exchange views: no interaction for Cooke's method (mathematical aggregation) (Cooke, 1991); full interaction for the SHELF method (Oakley & O' Hagan, 2016); and a more limited, controlled interaction for Delphi, a middle ground between the first two methods (EFSA, 2014). The above-mentioned literature addresses typical biases involved in eliciting probabilities that apply in evaluation settings as well as anywhere else: overconfidence, anchoring and adjustment, availability, and representativeness.

Although we haven't tested it fully at the time of writing, we believe the SHELF method could be easily adapted to evaluation processes: the official website http://www.tonyohagan.co.uk/shelf/ comes with templates for sets of slides that could be used in elicitation workshops, where participants are invited to establish boundaries for values (minimum and maximum) as well as tertiles or quartiles.

We can thus use either formal or informal procedures to elicit probabilities in participatory settings; but even when the procedure is informal, with this method we won't be able to avoid scrutiny because being transparent on the estimates means we are somehow forced to justify them!

2.4.1 The Bayes formula

The Bayes formula was first introduced in mid-18th century as a rule to calculate the probability of an event A happening once it is known that another event B has happened; provided we know the

probability of B and the joint probability of the two events A and B happening at the same time. In symbols:

$$P(A|B) = P(A \cap B) / P(B)$$

We can express the joint probability $P(A \cap B)$ in two ways: as the probability of B times the probability of A knowing that B has happened – $P(B)*P(A|B)$ – or as the probability of A times the probability of B knowing that A has happened: $P(A)*P(B|A)$. We can thus replace the joint probability in the expression above and rewrite as:

$$P(A|B) = P(A)*P(B|A) / P(B)$$

We can intersect an event B (imagine it as a set) with two halves of a space, say A and its opposite ~A; and reframe event B as the union of the intersections of B with the two halves of the space $(B \cap A) \cup (B \cap \sim A)$. So the probability of B can be expressed as the probability of the union, which is the sum of the probabilities of the two intersections: $P(B \cap A) + P(B \cap \sim A)$. If we replace P(B) in the formula above with this expression we obtain:

$$P(A|B) = P(A)*P(B|A) / [P(B \cap A) + P(B \cap \sim A)]$$

If we expand the joint probabilities as above, we obtain the following "long form" of the Bayes formula:

$$P(A|B) = P(A)*P(B|A) / [P(A)*P(B|A) + P(\sim A)*P(B|\sim A)]$$

If we assume that event A is the existence of a theory (T) or the fact that a certain process has taken place to lead to an outcome; and that event B is the observation of evidence (O) that leads us to believe that T has taken place, the formula becomes:

$$P(T|O) = P(T)*P(O|T) / P(O)$$

In other words, the formula calculates how much bias-free [12] confidence in the theory we should have after having observed empirical data. P(T) is known as the "prior confidence" in the theory and P(O|T) as the "sensitivity" (or the probability of evidence O once we know that theory T is true). If we want to measure the strength of evidence O for theory T (see also section 2.1.2) we need to use the long form:

$$P(T|O) = P(T)*P(O|T) / [P(T)*P(O|T) + P(\sim T)*P(O|\sim T)]$$

The formula has long-standing applications in several fields such as physics, finance, medicine, law, engineering, computer science, and crime investigation. In medicine, it can be used to formalise confidence that the diagnosis is correct on the basis of observed symptoms and results of medical tests. In evaluation, it gives us an estimate of the confidence we should have that the theory is true after having completed data collection and analysis. More formally, it calculates the posterior confidence, indicated as P(T|O): or our confidence in the theory *after* observation of empirical data. The formula needs to be fed three values:[13]

1. The **prior confidence**: our degree of belief in the existence of the theory *before* observation of empirical phenomena, indicated as **P(T)**. It's also known as the "base rate"; for example, in medical diagnosis, it would be the prevalence of the disease in a group of people similar to the patient. In evaluation, it embodies

[12] Specifically, free of conservatism bias.

[13] In statistical tests, Type I error is defined as the probability of wrongly rejecting the null hypothesis (often denoted as H_0). In our case H_0 would be the hypothesis that the theory is false (~T), while the hypothesis that the theory is true (T) corresponds to the alternative hypothesis H_1. O is evidence leading us to believe that theory T is true, or to believe in the alternative hypothesis H_1. Therefore, wrongly rejecting H_0 means that H_0 is true and our theory is false (~T); and at the same time that we observe O which leads us to believe that T is true and reject H_0 (the null). The probability of wrongly rejecting the null thus becomes the probability of observing O under the hypothesis that the theory is false (~T) (or that the null is true), which is exactly how we have defined the Type I error.

prior knowledge on the plausibility of the theory, which we may have from past evaluations or systematic reviews. If we don't know want this prior knowledge to affect our posterior estimate, and we want to let the evidence "speak for itself", we can set it at 0.5 (exactly in the middle of the confidence spectrum: see Table 3 below).

2. The **sensitivity**: the probability of making a specific observation O under the hypothesis that the theory is true: a conditional probability indicated as **P(O|T)**. In medical diagnosis, it is the probability that the patient will present symptoms and/or test results leading us to believe the patient has a certain condition T, if that condition is actually present (see the confusion matrix in section 2.1.3). In evaluation, it's the probability that given documents or interviews or timelines or surveys etc. will present specific features if the theory T under investigation is true. For example, if a policy influence theory is true, the Sensitivity could be the probability that the text of two documents by different authors, one of which is supposed to have been influenced by the other, would present similar features; or that a person involved in the intervention who has stakes in providing a positive picture of it does indeed provide it in an interview if the positive impact theory about the intervention is true. If the sensitivity of an observation is high, it means we expect to see it if the theory is true, and the theory is weakened if we don't see it (Befani & Stedman-Bryce, 2017).
3. The **Type I error**: the probability of making a specific observation O under the hypothesis that the theory is false, a conditional probability indicated as **P(O|~T)**. In medical diagnosis, it is the probability that the patient will present symptoms and/or test results leading us to believe the patient has a certain condition T, when that condition is actually absent (hence the word "error"). In evaluation, it's the probability that given documents or interviews or timelines or surveys etc. will present specific features if the theory T under investigation is false. For example, if a policy influence theory is false, the Type

> I Error could be the probability that the text of two documents by different authors would still present highly similar features that would normally lead us to believe that the theory is true; or that a person involved in an intervention whom we assume has an incentive to portray it in a negative light would provide a positive picture of it in an interview when the positive impact theory about the intervention is actually false. If the Type I error of an observation is low, it means we don't expect to see it if the theory is false, so if we see it the theory is strengthened. The inverse of the Type I error (called the specificity) is also a useful concept (see section 2.1.3): it's the probability that we will NOT make the specific observation if the theory is not true, and hence we won't be misled into believing that it's true. Notice that if the specificity is high (which is the same as saying that the Type I error is low), or in other words the theory being false implies that it's very unlikely we will make the observation, the logical implication is that if we do make the observation then the theory must be true.[14]

To provide a quick example of how this formula works, let's look at Table 2. It presents three theories, each associated with a different value of the prior. Let's assume we are assessing an empirical observation O and estimating the Sensitivity and Type I error in relation to it. The fourth column shows the value of the posterior obtained from feeding the values of the first three columns into the formula[15]. Notice how the posterior (the confidence that the theory is true) is higher when type I error is lower and how – when the positive observation is actually made – the sensitivity values are not as impactful on this confidence as the type I error values. That is because Sensitivity is mostly relevant for conclusiveness (namely, for disconfirmation) when the positive observation is not made.

---

[14] Formally: IF non T => non O, THEN O => T

[15] All our calculations are made using the freely available Bayesian Updating tool (Befani, Bayes Formula Confidence Updater, 2017) https://www.cecan.ac.uk/wp-content/uploads/2017/03/bayes_formula_confidence_updater.xlsx

Research on conservatism shows how important this formula is: not using it means incurring into a systematic underestimation of the strength of empirical evidence (Kahneman, 2012).

2.4.1.1 How confirmation bias affects probability estimates

In the process of identification of numerical values, both Sensitivity and Type I error are subject to various forms of Confirmation Bias (see Section 1.2.1). For example, for Type I error, we can fail to recall events linked to alternative explanations or theories that can produce the assessed observations as much as the theory under investigation. Failing to remember the existence of an alternative knowledge product, perhaps published by another organisation, with a similar content that the institution could have been influenced by, could be an example of Memory Confirmation Bias. An example of Search Confirmation Bias is when, if an institution has had contacts with several think tanks in the lead up to the strategy formulation, the investigator would normally not devote the same energy to research and analyse the linkages between the institution and all of these organisations, prioritising one preferred organisation. Furthermore, Analysis Confirmation Bias could manifest itself when the evaluator is exposed to a (theoretically influenced) stakeholder that they know has good relations with the (theoretically influencing) organisation, and the stakeholder is claiming that their institution was influenced by the former. A victim of confirmation bias might tend to think of that evidence as a smoking gun (see section 2.6), downplaying or even failing to consider any external motivation the stakeholder might have had; while another, more cautious evaluator would consider alternative motivations more seriously and downplay the confirmatory power of that observation.

As for sensitivity, low values can indicate CB (in all its forms) because high values make it easier for the theory to be rejected if hoop tests are not observed (or "passed", which is something we

might unconsciously fear). Search bias makes us reluctant to seek Hoop Tests and Analysis bias makes us underestimate sensitivity values so that, even if fail to observe Hoop tests, we can still hold strong hopes that the theory is true.

In summary, for theories that commissioners have a stake in confirming, higher values of both Sensitivity and Type I error should be trusted more than lower values, while the latter should raise suspicions of CB.

2.4.2 Measuring evidence "strength"

In addition to updating confidence, calculating the values to feed into the Bayes formula provides the opportunity to measure the evidence "strength" or "probative value" in at least three ways (Friedman, 1986; Kaye, 1986):

1. The **difference** between Posterior and Prior: **P(T|O) – P(T)**
2. The **ratio** between Sensitivity and Type I error (a.k.a. the likelihood ratio: **P(O|T) / P(O|~T)**
3. The **logarithm** of the likelihood ratio (a.k.a. "weight of evidence"): **log [P(O|T) / P(O|~T)]**

Table 2: Example of confidence levels in statements/theories before and after data collection

Theory	Prior P(T)	Sensitivity P(O|T)	Type I Error P(O|~T)	Posterior P(T|O)	Posterior minus Prior	Likelihood Ratio LR (Sensitivity/Type I Error)	Weight of Evidence (log LR)
One	0.50	0.90	0.10	0.90	0.40	9	2.20
Two	0.40	0.90	0.01	0.99	0.49	90	4.50
Three	0.60	0.60	0.30	0.75	0.15	2	0.69

Perhaps the most used measure is the Likelihood Ratio; however, the latter quickly skyrockets into very high values for very small values of the Type I error; which creates a need for the weight of evidence. You can use these measures to compare how strongly the same evidence supports different theories, or to compare how strongly different pieces of evidence support the same theory. In Table 2, observation O most strongly supports Theory Two and you can see how the LR in that case is ten times higher (90) than for Theory One (9); while the weight of evidence is roughly double (4.5 to 2.2). Perhaps the LR is easier to interpret, because it's an odds ratio: it tells you how much likelier it is to make that observation if the theory is true compared to a situation where the theory is false.

And observing O is 90 times likelier if Theory Two is true than if it's false. For Theory Three the odds are merely doubled: so the same observation feebly indicates that theory three might be true, too, but in a much weaker way: observing it is only twice as likely if theory three is true than if theory three is false.

In sum, the higher the difference between Posterior and Prior, and between Sensitivity and Type I error, the higher the strength or probative value of the evidence. The values required by the Bayes formula plus the likelihood ratio are part of the unaptly named "confusion matrix", which clarifies the fundamental concepts of diagnosis by laying out its basic elements as well as their interrelationships, and systematically mapping the relationship between (ontological) theory and (empirical) evidence.

2.4.3 The confusion matrix

The "confusion matrix" (or, as should be called more fittingly, the "clarity matrix"), systematically maps theory against empirical evidence, structuring their inter-relationships (Figure 2). It's a relatively simple 2x2 matrix, where the columns represent two opposite states of ontological reality (whether the theory is true or not), and the rows represents two opposite states of observable reality (whether empirical data denoting the existence of the theory is observed or not). We could consider a simple evaluation situation where the theory under investigation is that an institution's policy product has been influenced by another organisation's knowledge product. The observable reality we could focus on is the presence or absence of similar features between the policy document and the knowledge product.[16]

[16] Three of the real-life examples we present in detail in chapter 3 are similar to this "stylised" or simplified example.

Figure 2: The confusion matrix

		Theory (ontological reality)		
		The proposition / statement / theory is TRUE	The proposition / statement / theory is FALSE	
Empirical observation O leading us to believe that the proposition / statement / theory is true (observable reality)	Evidence (O) is OBSERVED	True Positive (TP)	False Positive (FP)	**Positive Predictive Value** = TP / (TP + FP)
	Evidence (O) is NOT OBSERVED	False Negative (FN)	True Negative (TN)	False omission rate = FN / (FN + TN)
		True positives rate (TPR) = **Sensitivity** = 1 – Type II error = TP / (TP + FN)	False positives rate (FPR) = 1 – Specificity = **Type I error** = FP / (FP + TN)	**Likelihood ratio** = TPR / FPR = Sensitivity / Type I error
		False negatives rate (FNR) = **Type II error** = 1 – Sensitivity = FN / (TP + FN)	True negatives rate (TNR) = **Specificity** = 1 – Type I error = TN / (FP + TN)	

The values of the four central cells represent four possible states, obtained by crossing the two possible states of ontological reality (theory true or not true) with the two possible states of empirical reality (empirical data showing that the theory is true observed or not observed).

The first cell on the top left (true positive) represents a situation where evidence leading us to believe that the theory is true is observed, and the theory is indeed true. These cases are named "true positives" because the "presence" of the evidence is not misleading. In our evaluation example, we could posit that, if influence has taken place, we would observe that the policy document is similar to the knowledge product; so the true positive situation would be when the policy has actually been influenced and we observe similar features in the two documents. The second cell on the top right (false positive) represents a situation where empirical data leading us to believe that the theory is true is observed, but the theory is false. These are termed "false positives" because we are misled by what we have observed to think that the theory is true, while it isn't. In our

evaluation example it would be when we observe similar features in the two documents, despite the policy not being influenced by the knowledge product. The third cell on the bottom left covers cases where the theory is true but the empirical data leading us to believe so is not observed, as in a "false negative": we are led to believe the theory is false while in fact it is true, just like when that particular kind of influence has actually taken place but there are no similar features in the two documents. Finally, the fourth cell on the bottom right is the true negatives cell, where we do not observe the evidence and are correctly led to believe that the theory is false (no influence has taken place, no similar features).

When we apply this method, we assess our empirical material on its ability to indicate reality correctly, or to change our beliefs about (ontological) reality in the right direction; in other words we want our data to have high sensitivity and high specificity and low values for both types of errors. It's important to notice that the ability of data to make the correct suggestion is asymmetric, which means it can be different depending on whether the theory is true or not; and on the whether the suggestive data is observed or not. The ability of an observation to correctly show that the theory is true when made, is not the same as the ability of an observation to correctly show that a theory is false when not made. Most empirical material is asymmetric, that is it's able to weaken the theory if not observed but not necessarily to strengthen the theory if observed; or vice versa (it is able to strengthen the theory if observed but not weaken it if not observed). For example, observing identical features between the two documents by different authors is highly confirmatory for the influence theory; but not observing this does not hold great power to weaken or reject the theory.

These abilities are measured by the number of times that tests are right and lead us to a correct belief[17]; but because they are asymmetrically powerful in showing presence or absence, the reliability in showing presence has a different name than the ability to show absence. If the theory is true (or the disease is present), the number of times the positive observation is made out of the total number of attempts is named "the true positives rate", or "sensitivity" (remember this term in the Bayes formula). It's the probability of making the positive empirical observation under the hypothesis that the theory is true. It's also the probability of observing similar features in the two documents if influence has actually taken place.

If the theory is false, the number of times the test correctly identifies absence of the condition by showing as negative out of the total number of times the test is performed in a "theory is false" scenario (or the number of times we do NOT observe the data if the theory is false), is termed "true negatives rate", or "specificity". It indicates the empirical material's ability to correctly identify cases where the theory is false. In our evaluation example, it's the probability of not observing similar features between the two documents if influence has not taken place.

Ideally, we want empirical observations to be both sensitive and specific for our theories of interest: we want them to tell us that the theory is true when it is so, and that the theory is false when it is so; rather than them misleading us to hold false beliefs in either scenario. However, unfortunately, evidence tests are not always correct, and do mislead us at times. Their error rates are asymmetrical, too, just like their abilities to be right: they have different names and can be different depending on whether the theory is true or not.

[17] The beginning of section 2.4 provides more details on how to estimate probabilities: but as an example, in medical diagnosis sensitivity and specificity are estimated using empirical frequencies.

If the theory is true (or the disease is present), the number of times the test is misleadingly negative out of the total number of times the test is performed, is termed the "false negatives rate", or Type II error. It's the probability of being misled into believing that the theory is false when it isn't. If the theory is false, the number of times the test is positive and we are misled to believe that the theory is true, is called the "false positives rate", or Type I error. It's the probability of being misled into believing that the theory is true when it isn't. Needless to say, we want both errors to be low.

Remember that the Type I error is included in the Bayes formula, so we'll use it more than the Type II error. In statistical tests, the Type I error is considered more serious than the Type II error because it means we are rejecting a hypothesis which normally represents current knowledge (or received wisdom if you want) in favour of something which is usually new: we claim to have made a discovery while we actually haven't and are abandoning tried and tested ways for something that will not work. In comparison, a situation where we hold on to past beliefs and fail to acknowledge a new discovery that is actually true (the Type II error) is still bad but sounds less disruptive. In evaluation, imagine taking a policy decision that will affect the public on the basis of a wrong theory that you think is true (Type I error), compared to a situation where you don't act on a true theory because you think it's false or because you simply ignore its existence.

In the confusion matrix, the two errors are linked to the two "abilities to be right": the Sensitivity plus the Type II error add up to one, as do the Specificity and the Type I error. So we only need to know the Sensitivity to know the Type II error (or vice versa); and we only need to know the Specificity to know the Type I error (or vice versa). Which is why it makes sense to focus on one value only for each column, as the Bayes formula does.

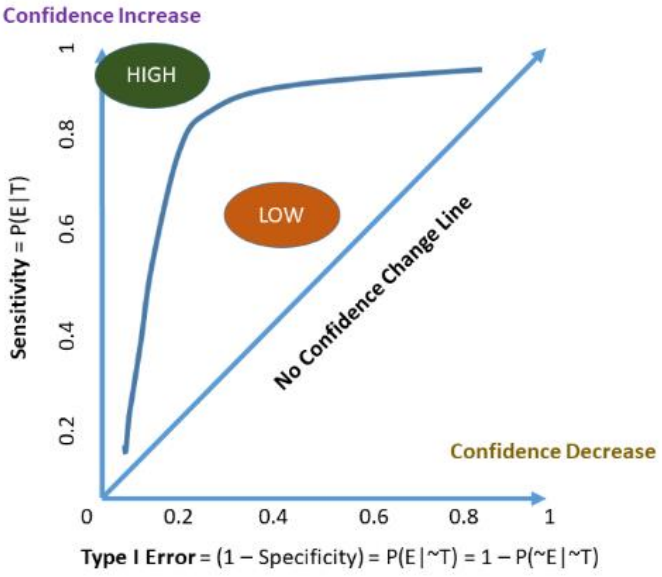
We hope the Confusion Matrix clarifies the relationship between hypotheses on theoretical realities, evidence, strength of evidence, and the asymmetrical power of the evidence to strengthen or weaken

the theory. Perhaps attention should be also paid to where the posterior and the likelihood ratio (a measure of strength or probative value) are located in the matrix and which other values they are obtained from. In the hope of clarifying these conceptual underpinnings further, we now address how observations or pieces of evidence can be visualised on a chart.

2.4.4 Plotting observations on the likelihood ratio chart

An interesting tool offered by the Bayesian formalisation is the likelihood ratio chart, which is a X-Y axis chart where the X-axis shows the value of the Type I error and the Y-axis shows the value of the Sensitivity. For each piece of empirical material (and each theory under investigation), we can plot these values in the chart (Figure 3) (Bennett, 2014; Befani, 2020). Observations falling on the diagonal line do not alter our prior confidence: the probability of making them if the theory is true (Sensitivity) is the same as the probability of making them if the theory is false (Type I error); hence making this kind of observation is not informative on the truth of that theory. The probative value of the observations lying along the diagonal line is close to zero on the first measure of probative value; close to one for the likelihood ratio; and again, close to zero for the weight of evidence.

Figure 3: How confidence is affected by sensitivity and specificity



As we move away from the diagonal line, however, the probative value changes. The closer we move to the top left corner, the higher it gets. If the observation is made, being close to the Y-axis indicates strong confirmation (not necessarily being close to the top left corner). If the observation is not made, being at the top of the chart indicates strong disconfirmation (not necessarily being in the top left corner). The top left corner, however, is the place for powerful empirical material, that is always informative for the theory under investigation, no matter if it's observed or not. It's the area where sensitivity is highest and Type I error is lowest (and thus the LR is highest); and the test is both highly sensitive and highly specific. We'll address this chart again below (Figure 5) but if you as the

evaluator are mapping your pieces of evidence against a theory (or the same evidence against different theories) in this visual chart, the top left corner is the best possible placement.[18]

2.5 Expressing confidence qualitatively

Expressing confidence levels (or probabilities) is key to the application of the approach: as we've seen above, in order to understand how observations should affect our confidence that the theory is true, we need three types of confidence/probability values: the prior (the initial degree of belief in the theory); the sensitivity (how likely it is that we will make that observation if the theory is true) and the Type I error (how likely it is that we will make that observation if the theory isn't true).

While quantifying confidence levels is typical practice in standard (frequentist) statistics – where three levels of confidence are usually considered: very good (0.99+), good (0.95+) and not so good (0.90+) – we don't necessarily need to use numbers to assess confidence. We can use qualitative descriptors of confidence levels, and assign them to quantitative intervals or numerical ranges. Table 3 outlines our recommended way of describing confidence levels qualitatively, with options ranging from neutrality to practical certainty; and covering intermediate degrees like "more confident than not", "cautiously confident", "highly confident", and "reasonably certain". Our scale is relatively fine-grained, offering 5 possibilities on each side of the spectrum plus a central neutral point. The distribution of confidence levels between the two extremes incorporates the logarithmic shape of the human sensory perception curve (Befani, 2020): you can see that the central ranges are the largest (0.20), and that intervals gradually narrow (0.15, 0.10, 0.04) as

[18] The bottom right corner is equally useful, but with a reversed meaning: if you make an observation that's sitting in the bottom-right corner, the theory is ruled out (or its opposite is confirmed); while if you don't make that observation, the theory is confirmed (or its opposite is ruled out).

confidence increases, becoming the smallest (0.01) for the highest levels of confidence (Figure 4). You might also notice that the scale is symmetrical on the two opposite sides of the true/false spectrum.

As long as there is transparency over which intervals are associated with which qualitative descriptors, other scales can also be used; however, other scales we have come across tend to be less fine-grained, more linear, or stricter on the extremes, while this scale is comparable to the standards currently used in quantitative social science; including, in particular, the 95% and 99% confidence thresholds.[19]

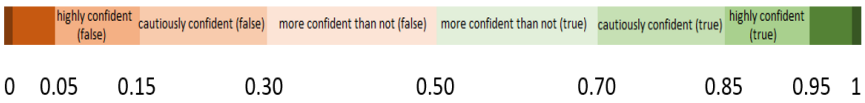
The values in Table 3 can be visualised along a "confidence spectrum", where confidence ranges from practical certainty that the theory is false (dark orange left end in Figure 4), to practical certainty that theory is true (dark green right end). The benefits of the Bayes formula can thus be harnessed from a qualitative perspective: the next section will get into more detail on how this can be accomplished.

[19] Some authors propose the adoption of the decibel scale (Fairfield & Charman, 2017) used in physics and assert that 30 db is the gold standard after which evidence "speaks loudly". To us, this looks like uncritical borrowing of a concept from the physical sciences because 30db corresponds to 10*Log10(LR), or a likelihood ratio of around 1140; which means a posterior of 0.999 starting from a prior of 0.5. The most widely used standard in social statistics is 95% confidence, which corresponds to a likelihood ratio of 19 and a dB value of less than 13. It makes no sense to set a qualitative evidence standard for our method that is 60 times stronger – or 17 dB "louder" – than the standard used in quantitative research.

Table 3: Translation between confidence levels and ranges/ numerical intervals

Qualitative descriptor of confidence level	Low end	High end	Middle Value	Range length
Practical certainty that () is true/observed	0.99	1	0.995	0.01
Reasonable certainty that () is true/observed	0.95	0.99	0.970	0.04
High confidence that () is true / observed	0.85	0.95	0.900	0.10
Cautious confidence that () is true/observed	0.70	0.85	0.775	0.15
More confident than not confident that () is true/observed	0.50	0.70	0.600	0.20
Neither confident nor not confident that () is true/observed (or false/not observed) – no idea	0.50	0.50	0.500	0
More confident than not confident that () is false/not observed	0.30	0.50	0.400	0.20
Cautious confidence that () is false/not observed	0.15	0.30	0.225	0.15
High confidence that () is false/not observed	0.05	0.15	0.100	0.10
Reasonable certainty that () is false/not observed	0.01	0.05	0.030	0.04
Practical certainty that () is false/not observed	0	0.01	0.005	0.01

Figure 4: The confidence spectrum



2.5.1 Qualitative Bayesian Updating

Feeding qualitative judgements into the Bayes formula is not completely different from feeding it numerical point estimates: we still have to assess the three likelihoods (Prior, Sensitivity, and Type I error). However, instead of having to settle on a single number, we can produce one of the qualitative judgements of table 3 and use the two extremes of the corresponding numerical range for updating the formula. Since the range has a middle value, we can also associate a single value to the range, which will be practical in several circumstances (see Chapter 3). However, using a single number is not necessarily required, and – assuming our initial confidence is neutral (0.5) – we can focus on the ranges for Sensitivity and Type I error. Estimating these two values qualitatively means implicitly selecting two numerical intervals with a total of four extremes. For example, for sensitivity we might be "highly confident" that we'll make the observation if the theory is true (0.85-0.95); and for Type I error, we might be "reasonably certain" that we won't make it if the theory is false (0.01-0.05). These two ranges have four extremes and create the four reference scenarios illustrated in Table 4, with infinite possibilities in-between (Befani, 2020).

The first scenario[20] is associated with the maximum strength or probative value of the evidence: the sensitivity value is assumed to be the highest possible in the selected range, while the Type I error the lowest. At the other end of the spectrum, the last scenario is associated with the lowest strength or probative value, and the sensitivity is assumed to be the lowest end of the range, while the type I error the highest. The second and third scenarios sit in the middle of the continuum: in the second, the S is low and T1E high, while in the third S is high and T1E low.

[20] As we can see from the probative value measures, as well as from the posteriors, assuming the observation has been made, the first scenario represents the strongest evidence (LR of 95.00) and updates the prior to 0.99, while the fourth scenario is the weakest (LR of 17.00) and updates the prior to 0.94.

Table 4: Four scenarios created by estimating confidence qualitatively

	Sensitivity	Type I error	Probative Value: Likelihood Ratio (LR)	Probative Value: log (LR) or Weight of Evidence	Posterior (up from a Prior of 0.5)
Highest Probative Value (highest S, lowest T1E)	0.95	0.01	95.00	4.55	0.9896
Middle Scenarios	0.85	0.01	85.00	4.44	0.9884
	0.95	0.05	19.00	2.94	0.9500
Lowest Probative Value (lowest S, highest T1E)	0.85	0.05	17.00	2.83	0.9444

This means that, by making two qualitative confidence assessments on the Bayes formula values, and a neutral prior, we are able to establish boundaries for the posterior confidence: and claim that it ranges from 0.94 to 0.99. It largely overlaps with "reasonable certainty"; or in other words, if we judge the sensitivity "highly confident" and the type I error "reasonably certain", from a neutral prior, our posterior confidence almost entirely overlaps with "reasonable certainty" that the theory is true.

If the numerical range resulting from the confidence update covers two qualitative descriptors, we could choose the one that the range overlaps with more extensively: for example, if we had 0.87-0.97, we could pick "high confidence that the theory is true". Or, if we want to be more conservative, we can pick the lowest qualitative descriptor overlapped by the range. Table 5 presents different theories, for all of which we have a prior of 0.5; and lists different qualitative judgements of sensitivity and type I error for an observation O in relation to those theories. The last column presents the resulting posterior range and corresponding qualitative confidence level.

A freely available tool for Bayesian Updating (Befani, 2017) currently offers the opportunity of working with qualitative levels of confidence for Sensitivity and Specificity: it converts them into numerical ranges and then computes the corresponding range for the updated confidence or posterior.[21] The numerical range can then be converted back into the qualitative descriptor.

A simpler alternative is to work with single point estimates using the midpoint of the range associated with the qualitative judgement (Table 3). Continuing the above example, we could use 0.90 for sensitivity (the middle of the high confidence interval from 0.85 to 0.95) and 0.03 for Type I error (the midpoint of the negative reasonable certainty interval from 0.01 to 0.05). Our posterior would be 0.97, exactly in the middle of the "reasonable certainty" range.

Table 5: Bayesian Updating with qualitative statements[22]

Theory	Prior	Sensitivity	Type I Error	Posterior
Theory One (T_1)	No idea – it could be either true or false (0.5)	High Confidence that O is observed (0.85-0.95)	Reasonable Certainty that O is not observed (0.01-0.05)	0.94-0.99: Reasonable certainty that T_1 is true[23]
Theory Two (T_2)		More confident than not that O is observed (0.50-0.70)	Cautious confidence that O is not observed (0.15-0.30)	0.62-0.82: Cautious confidence that T_2 is true[24]
Theory Three (T_3)		More confident than not that O is	Practical certainty that O is not	0.97-1:

[21] This is done by carrying out four different updating calculations, one for each of the following scenarios: min S, min T1E; min S max T1E; max S min T1E; and max S max T1E. Two of these represent the extremes for probative value: min S max T1E at the lower end, and max S min T1E at the higher end.

[22] An interval for the prior could also be created: that would increase the number of scenarios required to eight instead of four.

[23] or High Confidence using the worst case scenario method

[24] or barely more confident than not using the worst case scenario method

Theory	Prior	Sensitivity	Type I Error	Posterior
		not observed (0.30-0.50)	observed (0-0.01)	Reasonable or Practical certainty that T_3 is true
[That particular type of] influence took place		No idea – O could be either observed or not (0.5)	High Confidence that O is not observed (0.05-0.15)	0.77-0.91: Cautious / High confidence that T is true
		No idea – O could be either observed or not (0.5)	Reasonable Certainty that O is not observed (0.01-0.05)	0.91-0.98: Reasonable certainty (high confidence) that T is true

2.6 Metaphors and typologies: the Process Tracing tests

Figure 5 visualises where observations can potentially lie on a bi-dimensional space, where the x-axis is the Type I error and the y-axis the Sensitivity (Humphreys & Jacobs, 2015). Bayesian Updating and the Confusion Matrix allow observations to occupy any potential space on a continuum of values within that square. On the other hand, the qualitative research method Process Tracing (Beach & Pedersen, 2013; Bennett, 2010; Bennett, 2008; Bennett, Checkel, & (eds), 2014; Collier, 2011; Van Evera, 1997), which is based on an informal instead of formal Bayesian logic, divides that space in roughly four blocks and assigns one of its four famous metaphors (Smoking Gun, Hoop test, Doubly Decisive, and Straw-in-the-Wind) to each block. In a way, Process Tracing "crispifies" the continuous space of possibilities; it makes the space discrete, by dividing it into categories which can merely be conclusive or not in

terms of evidence strength.[25] Box 2 lists the main characteristics of the four Process Tracing tests, indicating where they fit in the confusion matrix. Notice the difference between not observing a smoking gun, which is absence of evidence, and not observing a hoop test, which indicates evidence of absence.

Box 2: the four Process Tracing tests and their relation to the confusion matrix

Smoking Gun: if a specific empirical observation is made, the theory is confirmed (think a suspect found to be holding a smoking gun over someone who was just shot). If that is not made, the theory is neither confirmed nor rejected (this could also be referred to as "absence of evidence"). High specificity or true negatives rate; average or low sensitivity.

Hoop Test: if a specific empirical observation is made, the theory is neither confirmed nor rejected. If it is not made, the theory is rejected (it didn't make it through the hoop. This can also be referred to as "evidence of absence"). High sensitivity or true positives rate; average or low specificity.

Doubly Decisive: if a specific empirical observation is made, the theory is confirmed. If it is not made, the theory is rejected. High sensitivity and high specificity. This test is always useful and conclusive whether we make the observation or not.

Straw in the Wind: the theory is never confirmed nor rejected, though it can be slightly strengthened or weakened. The test is always inconclusive whether we make the observation or not. Average or low sensitivity; and average or low specificity.

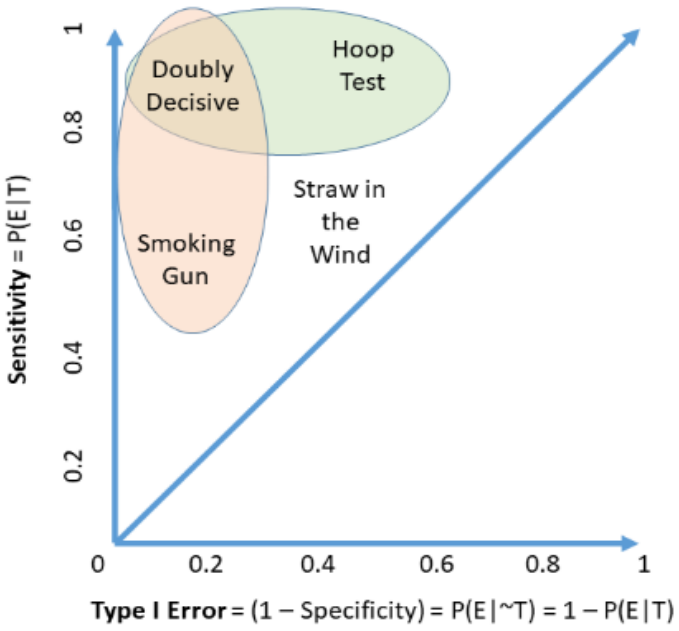
Adapted from Befani, D'Errico, Booker & Giuliani (2016)

[25] This way of thinking has been mitigated in the second edition of a major textbook (Beach & Pedersen, 2019).

Considering the connections between Process Tracing metaphors and the confusion matrix, we can claim that Smoking Gun tests are highly specific but not sensitive; that Hoop tests are highly sensitive but not specific; that Doubly Decisive tests are both sensitive and specific; and that Straw in the Wind tests are neither sensitive nor specific. In other words, Smoking Guns have a low Type I error and Hoop Tests have a low Type II error.

We can add the Process Tracing test "labels" to Figure 3 and obtain Figure 5, where the strengthening or confirmatory area is the orange egg on the left handside, and the weakening or disconfirmatory area is the green egg on the top of the chart.

Figure 5: Relation between sensitivity, specificity, and Process Tracing tests



It's worth remarking how probative value, or test "strength", is a different concept from test direction, or the ability to confirm or disconfirm: an observation can confirm strongly or weakly; it can disconfirm strongly or weakly; it can do one strongly and one weakly, both strongly, or both weakly (Table 6).

Table 6: Distinction between test direction and test strength in relation to a particular theory

		Test direction	
		Confirmation	Disconfirmation
Probative value (test strength)	Strong	Smoking Gun, Doubly Decisive	Hoop Test, Doubly Decisive
	Weak	Hoop Test, Straw-in-the Wind	Smoking Gun, Straw-in-the Wind

Information about probative value only, that ignores the test direction (or whether that strong ability is confirmatory or disconfirmatory), is incomplete. Likewise, information about test direction that ignores test strength is also incomplete: both types of information are needed to characterise and use empirical tests in support of or against theories.

As an evaluator or a researcher, you can use a diagnostic lens, or a formal Bayesian approach, to add nuance, transparency, and traceability to the way you apply Process Tracing and make judgements on the strength and direction of evidence tests.

3. How to apply Bayesian TBE in practice

CHAPTER SUMMARY. In this chapter we present examples of the method's application, organised along three steps. In the first step, we outline the form or shape that theories can take in order to be compatible with this method; how they are represented, illustrated, and formulated. We can apply the method to a wide range of theories: process mechanisms, CMO configurations, and models of complex systems; in a wide range of policy sectors: our examples cover policy influence in development, energy policy, forestry policy, and public health. In the second step we show how data collection can be designed having in mind the notion of strong or conclusive evidence and seeking it. We illustrate how the evidence is weighed or assessed against particular theories of change and the transparent steps we follow to come to the conclusion that some theories are more strongly supported by the evidence than others. Finally, in the third step we get into detail about more complex but relatively common situations and show how confidence can be updated when the evidence is multidimensional, mixed, or even contradictory.

Now that we've covered the conceptual and theoretical aspects of diagnostic/Bayesian evaluation, we are ready to put the latter to the test. In this chapter we discuss how the method's application looks like in practice, following a logical sequence of three steps: a) developing a testable theory; b) identifying diagnostic tests and designing data collection; and c) updating confidence using the Bayes formula (quantitatively or through the qualitative translation of confidence intervals). The sequence can be thought of as a "cycle" because it is often repeated for different theories that seem worthy of testing as new evidence emerges. Each step includes a discussion of its main objectives, challenges, and solutions which are exemplified with real life applications.

3.1 Step One: Developing a testable theory

This stage requires us to have at least a sketch of a theory (which can take the form of a proposition, statement, mechanism, explanation, claim, process, model, etc) that we want to empirically test. It can be developed as possible answers to evaluation questions, like for example the following:

1. What role did the intervention (and/or other factors) play in achieving the outcome?
2. How did the intervention (and/or other factors) make a difference?
3. How did the intervention (and/or other factors) contribute to the outcome?

The idea of "testability" is linked to standards of scientific quality like falsifiability and demonstrability: a falsifiable theory can be rejected by potentially observable evidence. In our context, the notion of falsifiability is not to be intended in a deterministic or strictly Popperian sense; but it's relevant because, for practical purposes, we can consider probabilistic results associated with high levels of confidence, like for example 0.99999 or higher, to approximate determinism. It is well known that not all theories are falsifiable (for example the existence of God or some conspiracy theories) but the important distinction for us is between theories for which conclusive evidence can potentially be found, and theories for which it cannot; in other words, between theories we can reasonably prove or substantially increase our confidence about, and theories which – at best – we can only find weak and inconclusive evidence for.

For example, "the programme had benefits" or "the intervention had an influence" are too ambiguous and vague to be demonstrated conclusively, even by probabilistic standards, because even if the programme has been a disaster by most accounts, most likely it will have had some benefits for at least one person. Theories need to be reformulated more specifically: for example, "the programme has had

at least some benefits". However, the latter is not interesting enough for us and we will want to know which benefits it had, how they were produced, who it had benefits for, and so on.

Using Cartwright's distinction between high-level, middle-level, and low-level theory (Cartwright, 2020; Cartwright, Charlton, Juden, Munslow, & Williams, 2020), we anticipate that this method is mostly applicable to low-level theories (concrete, particular, and local) unless enough similar cases can be merged to obtain a more abstract, more general, and/or more global proposition tending towards mid-level. This is in line with the method's relevance for evaluation because, in our experience, low-level theories are the ones that most within-case evaluations actually target, when not enough cases are available or can be investigated to test a higher-level theory.

In terms of which form the theory takes, we choose not to be overly prescriptive and offer the Bayesian Updating option to a wide range of theoretical statements. While some will describe a causal process, others will focus on snapshot mechanisms triggered at a particular point in time; yet others will merely refer to conditions or the unobservable presence of properties (like for example for medical diagnosis or simulation of complex adaptive systems). This method is not merely a variant of Process Tracing, but rather an extension of Theory-Based Evaluation and is compatible with all the forms that TBE takes.

While we would often start from general propositions or higher-level theories, in the first application step we articulate our theories and increase their conceptual precision, bearing in mind whether empirical evidence can potentially exist that can be linked to the theory as directly as possible. Our statement that our organisation has had an influence on another organisation's policy or strategy is initially untestable and requires zooming in on specific forms of influence that can be clearly identified and tested empirically. In the following sections we discuss six theories from real life evaluations that were initially untestable and illustrate the process through which they became testable.

3.1.1 Tackling urban crises in Amman

In this example (Befani & D'Errico, 2020) the initial idea was that a knowledge product, the IIED Urban Context Analysis Toolkit (Sage, Meaux, Osofisan, Traynor, & Jove, 2017), had positively influenced the Greater Amman Municipality. At the time it was unknown to the evaluator what this influence consisted of, and what GAM activity in particular had been influenced by the toolkit. After a few preliminary interviews, the theory was refined to indicate that the object of influence was specifically the Urban Resilience Strategy. A quick check of the timelines (the publication dates of the two documents) puzzled the evaluation team because the supposedly "influencing" document was published one month after the supposedly "influenced" one. This led to further inquiry which established that the influencing document had "acted" while still in draft form. When the evaluator probed for more details as to which part of the influencing document had influenced which part of the influenced one, the theory was further refined to explain that the toolkit had influenced the methodology used by the International Refugee Council in a research which produced recommendations that were incorporated in the urban resilience strategy. At this point the evaluation team was given access to the recommendations brief that arose from said research and could verify that it matched parts of the officially published resilience strategy. However, the link between (draft) toolkit and recommendations brief wasn't quite clear yet, as the brief only included a short section on the methodology with few details. After further probing, the evaluator obtained the interview protocols that had been used during the research that eventually led to the production of the recommendations brief; and could match them to the draft toolkit.

The list below illustrates the evolution of the explanatory mechanism or "theory of change" from the first to the last stage.

1. Knowledge product has influenced local government.
2. Knowledge product has influenced local government's resilience strategy.
3. Draft knowledge product has influenced local government's resilience strategy.
4. Draft knowledge product has influenced research methodology of organisation whose research informed recommendations to the local government's on content of their resilience strategy.
5. Draft knowledge product has influenced research methodology's interview protocols used by organisation to produce research which eventually informed recommendations to the local government's on content of their resilience strategy. In particular, the team used the templates suggested in the toolkit to design the interview protocols.

You can see how the claim has become increasingly specific and linkable to observable, empirical evidence. This example is discussed again in sections 3.2 and 3.3.

3.1.2 Informing environmental policy in Malawi

In the second example (Annex L of (LTS International and the Centre for Development Management, 2017), the commissioner wanted to understand the role of CEPA (Centre for Environmental Policy and Advocacy), an organisation they were funding, in the process that led to the formation of the national environmental policy in Malawi. An initial series of exploratory interviews indicated that CEPA had produced two documents that had allegedly played an important role in shaping the policy: a Policy Review summarising the content of environmental policies of similar countries, and a Position Paper advocating for specific content to be included in the country's forthcoming policy. The government had mostly been

advised by a team of local consultants, mainly from academia, who had co-authored an Issues Paper, a White Paper and eventually a draft policy. The evaluation team then sought to understand what role the different documents played – including an alternative policy review produced by UNDP early on. Eventually, the existence of a temporal gap between the Issues Paper and the White Paper became apparent, which the team also investigated. It was discovered that a consultation process between, on one hand, MPs and the team in charge of writing the policy and, on the other hand, communities and civil society, had produced feedback on the Issues Paper and allegedly provided content to be included in the White Paper (which was very similar to the draft policy).

The case's theory development is summarised below, focusing on the claim's increasing level of detail.

1. The organisation has affected environmental policy formation in the country.
2. The organisation has produced a Policy Review and a Position Paper that have affected environmental policy formation in the country.
3. The organisation's Position Paper, which was grounded on their own policy review, influenced the Issues Paper authored by the government's consultants; who eventually produced a White Paper which was very similar to the draft policy (the latter virtually identical in content to the approved legislation).
4. The organisation's Position Paper, which was grounded on their own policy review, influenced the Issues Paper authored by the government's consultants; at this point the organisation set up or coordinated a series of events and activities aimed at collecting feedback on the Issues Paper which fed into the White Waper, which was very similar to the draft policy (and eventually the approved legislation).

3.1.3 Curbing deforestation in Brazil

The goal of this study was to explore and test the factors contributing to a sudden drop in deforestation trends in 2008-2011; as well as an additional decrease in 2012-2013, followed by a slow but steady reversal of the trend which became a pronounced increase after 2016 (Brandao & Befani, 2021). A wide range of data, primary and secondary, which was collected and analysed (from official reports and statistics to exploratory interviews and media) led to the hypothesis that three interventions implemented between 2008 and 2009 had a major impact on the initial drop; that the additional drop that followed could be explained by the three interventions continuing to work (some seemingly becoming more effective) and by a new intervention successfully engaging a group of deforesters that had previously been missed; and that the reversal of the downward trend could be attributed to the loss of effectiveness of almost all of these interventions, combined with the poor effectiveness of a new one that was supposed to be a replacement.

An intermediate version of the theory read as follows:

1. Improved environmental monitoring and enforcement, Credit restrictions, and the Cattle Agreements, instituted sequentially between 2007-08 and 2008-09 acted to trigger behavioural change of local actors involved in deforestation, leading them to substantially reduce deforestation activities. The interventions mostly targeted and hence affected medium and large actors which contributes to explain the major drop.
2. Previous interventions continued to work and were further strengthened by a multi-stakeholder zero-deforestation pact which managed to significantly engage and reduce deforestation among smallholders between 2010-11 and 2013-14.
3. Deforestation increases since 2015 can be explained by the end of the multi-stakeholder pact in 2014 and the gradual reduced effectiveness of the first and third interventions of the first

group. Value chain projects (a new intervention) did not manage to reduce the incentives to deforest like the previous policies had done.

When we started refining the theory in component one, we assumed that the three interventions, somehow taken together, were causally responsible for the decisive downward deforestation trend. We thought of their interaction as an obstacle race where the first obstacle represents the practice of law enforcement (the Boi Pirata); the second obstacle represents the restrictions to credit; and the third obstacle is the market pressure brought about by the reduction in demand for beef not complying with the law (the Cattle Agreements). Ranchers' incentives to deforest were attacked on all fronts: if they could survive law enforcement, they couldn't access credit; if they managed to survive both, the slaughterhouses embargo would get them.

Since we had to imagine what the opposite of the theory would look like to estimate its implications on empirical observations, we used set theory and imagined the theory as a logical union: as in, we assumed that at least one of these three interventions if not more were causally responsible for the widespread behavioural change that caused the downward trend in the outcome. In other words, we don't know if the runner stopped at the first obstacle, or at the last one, or if they stopped for a moment at the first and then continued and stopped again at the second; but we believe that at least somewhere along the way one of these obstacles changed their behaviour and was consequential.

For the second component, an intermediate version of the theory is that the three previously mentioned interventions continuing to work ensured that the deforestation rate trend did not lose momentum and did not start to increase again; and that the additional decrease (which was not as steep as in the previous period) was caused by either or both of two factors: a) the marginal toughening of the rules started in the previous set of interventions and b) the behavioural change observed in smallholders due to the

multi-stakeholder engagement process. In this period there is still no evidence of behavioural change meant to evade the new law enforcement, which will be found in the third period when the trend reversed.

We'll see in the next sections our analysis of the empirical implications of these theory components being true or false, and our assessment of how strong the evidence found was for them (sections 3.2 and 3.3).

3.1.4 Adopting protective behaviour in a flu pandemic

This study attempted to understand the factors behind the adoption of protective behaviour during a flu pandemic and assess the impact of various protective behaviours on infection rates, taking account of personal attitudes, social norms, and perceived threat. The tool chosen to represent this complex web of causal influences with individuals and their neighbours at the centre was Agent Based Modelling. An Agent Based Model was thus created and calibrated to represent and test various hypotheses. The question we chose to answer with Bayesian Updating was, what kind of protective behaviour lies behind which change in infection rates?

In this case the space of possible theories and explanatory propositions is represented by the possible settings of a simplified version of the TELL ME model of communication about influenza (Badham & Gilbert, 2015).[26] The model is made of two interacting layers, with the first consisting of simulated individuals that perceive their situation and make decisions about whether to adopt protective

[26] The model is also available as a NetLogo tutorial (Badham, 2019) (model Version 3).

behaviour. [27] In our specific example we decided to look at vaccination efficacy, and consider three values: 80% (standard efficacy, where 80% of the vaccinated population is protected); 90% (better efficacy, where 90% of the vaccinated population is protected); and 100% (ideal efficacy).

We might be in a position where we can't measure the efficacy of protection measures (or in this case, of the vaccine) and therefore we wouldn't be able to know which theory on protection efficacy is true. But if we can set the model to represent a number of theoretical hypotheses (for example, these three), we can study their implications in terms of empirical observations and – applying a diagnostic/Bayesian evaluation lens – we can estimate the probability that each of the theories is true after observation of various infection rates (Befani, Elsenbroich, & Badham, 2021). We discuss this example again in section 3.3.2.

3.1.5 Improving the governance of national parks in Uganda

The initial theory in this evaluation posited that an IIED-led network, the Poverty and Conservation Learning Group (PCLG), had influenced the Ugandan Wildlife Authority's decision making (Befani, D'Errico, Booker, & Giuliani, 2016; D'Errico, Befani, Booker, & Giuliani, 2017). Exploratory interviews focused on the mountain gorilla tourist permit fee at the Bwindi Impenetrable National Park (BINP), which is a fee levied on tourists who access the area to watch the local gorillas. Part of this fee is redistributed to the local community and the UWA had decided to increase the share from $5 to $10.

[27] One example of protective behaviour is vaccination, but the behaviour in the model is generic enough to be able to represent reducing contacts or improving hand hygiene.

Was this policy change influenced by the Uganda PCLG? The group had undertaken research (together with two other partners) to understand why the park resources continued to be used illegally despite many years of integrated conservation and development interventions. It was discovered that a key driver of the local communities' (illegal) behaviour was the perception of unfairness concerning the distribution of conservation resources. As a consequence, the group had advocated for increasing the shared fee. In particular, the group's chairman had written a letter to the UWA, specifically asking to increase the share from $5 to $10. The change was championed at the UWA level by a PCLG member who was also a UWA member.

The list below summarises the evolution of the claim from an ambiguous, vague statement to a detailed and testable proposition.

1. PCLG influenced conservation-related decision making in Uganda.
2. PCLG influenced decisions made by the Uganda's Wildlife Authority.
3. PCLG influenced the UWA's decision to increase the share of the Gorilla fee permit from $5 to $10
4. PCLG undertook research with partners and discovered that continued illegal activities in the park were rooted in dissatisfaction with distribution of conservation resources; which prompted them to write a letter to UWA requesting that the shared fee be raised from $5 to $10; a change that was championed by a PCLG member who was also a member of the UWA.

We discuss this theory again in sections 3.2 and 3.3.

3.1.6 Encouraging supply/demand balance in the decarbonised electricity grid

In this evaluation (Anderson, Ahmed, Befani, & Michaelis, 2020; BEIS, 2018), there were initially two high level questions around the contribution of a programme called "transitional arrangements", which aimed at encouraging a balance between supply and demand in a decarbonised electricity grid, through the development of so called "demand side response": or the reduction of imported electricity below an established baseline, by means other than a permanent reduction in electricity use. Under this definition, DSR may be achieved through any combination of onsite generation, temporary demand reduction or load-shifting. More concretely, the programme consisted of two auctions for specific types of capacity within the Energy Capacity Market, the first for delivery of capacity in the 2016/17 delivery year, held in January 2016, and the second for delivery of capacity in 2017/18, held in March 2017. The second TA had two main objectives: to encourage the last two types of DSR and to contribute to the development of flexible capacity for the future Capacity Market.

The high-level questions were:

1. What outcomes can be attributed to the (second) TA and were they as intended by the policymaker? What outcomes occurred for whom and under what circumstances?
2. Through what levers and causal mechanisms has the (second) TA contributed to these outcomes and the variation by group and circumstance?

While most cases described in this section, with the exception of 3.1.4, are theories expressed in the form of a causal process, in this case the explanation is more compact and was expressed with Context-Mechanism-Outcome (CMO) configurations. Two propositions were initially formulated, one conveying that the TA had been additional to the outcome and the other that it hadn't been so:

1. The TA has been additional in contributing to more and/or more competitive flexible capacity for the capacity market in 2018-19 and subsequent years.
2. The TA has made no difference to the capacity available to the CM in 2018/19 and subsequent years.

Thus formulated, the statements aren't quite testable yet; they were made more specific by creating a series of Context-Mechanism-Outcome (CMO) configurations that explained how and why the outcome was achieved or wasn't and what the TA's role was. We report the M (mechanism) component for some of those configurations:

For the "additionality" theory:

1. Our experience of participating in the second TA means the capacity market seems less risky.
2. In order to participate in the second TA, we invested in capacity or the ability to provide capacity which will make us better positioned to participate in the main CM.
3. (or new entrants) In order to participate in the second TA we have built a customer base and so now we want to continue with the CM.

For the "no difference" theory:

- (for existing aggregators) We have always intended to participate in the CM and the TA did not help us to grow our flexibility business.
- (for new entrants) We are a new entrant to flexibility in the CM but would have started participating with flexible capacity in the CM at the same level anyway, because of other changes, not the TA.

Notice how the theory is expressed in the form of a "snapshot mechanism", instead of a causal process mechanism. We discuss these theories again in sections 3.2 and 3.3.

3.2 Step Two: Identifying diagnostic tests and designing data collection

In the second step, the process of data collection (and to some extent analysis) is designed bearing in mind the potential existence of conclusive tests: that is, seeking to observe strong or high-probative-value evidence that can convincingly strengthen or weaken the theory. Observations need to be assessed according to their power to alter our confidence in the theory. Different observations will have different implications for the confidence in different theories. In order to understand how these different pieces of evidence change our confidence in the theory, we need to focus on the Bayes formula components: Sensitivity and Type I error; in other words, we need to estimate the chances of making a given observation under the two opposite assumptions that the theory is true and that it isn't; or to assess the implications of a theory being true (or false) on the chances of observing the pieces of evidence (see Chapter 2). We can then estimate the probative value using two of the measures introduced above (Likelihood Ratio and Weight of Evidence) (see Section 2.4.2).

In this section we reprise the theories of Section 3.1 and explain how we mapped them (or their components) against different observations. For each theory and each observation, we estimated the likelihood of making that observation under the hypothesis that the theory is true (a.k.a. a Sensitivity estimate) and the same likelihood under the opposite hypothesis that the theory is false (Type I error estimate). This is also illustrated theoretically in Table 7: each cell represents the association between one theory and one observation, which produces four different values. For each combination of observation and theory we have a sensitivity, a Type

I error, a likelihood ratio, and a weight of evidence. In this phase, however, we mostly stop short of making numerical estimates and merely establish the direction of confidence and some measure of strength, using the Process Tracing metaphors.

3.2.1 Helping local governments deal with urban crises

Table 8 shows how ontology or theory "squares" against empirics in the urban crises evaluation. The first, broad theory (the published knowledge product has influenced the local government's resilience strategy) does not have interesting implications for any of the four pieces of evidence considered, because they refer to specific parts of the process; the sensitivities are roughly the same for all (and not very high because the influence could have taken place in a number of ways); similarly, the Type I errors are relatively high because all those things could have materialised/happened even if influence had not taken place[28].

The second theory being true (the draft knowledge product has influenced interview protocols) would make it likely to find some close match between the draft toolkit and the interview protocols; some general alignment would be required (so be a hoop test) to confirm theory but a word-by-word matching would not be required for the same purpose: influence could have taken place even without observing, let's say, copy-pasting. For this reason, the sensitivity can be estimated to be somewhere in the middle of the probability spectrum. At the same time, a word-by-word matching, particularly of a considerable extent, is very unlikely unless the theory is true. Therefore, the Type I error is low.[29] When assessing the chances of

[28] As we would discuss in step three, we are not considering more precise estimations for this theory and we can posit that (as a result of these values fed into the formula) the prior is roughly the same as the posterior.

[29] This means that, upon making the observation, the theory would be considerably strengthened.

making the other observations, we find that they are mostly uninformative for the second theory and can thus be considered straw-in-the-wind tests.

The fourth theory component (recommendations brief influenced local government's resilience strategy) follows a similar pattern: only one observation (word-by-word matching between recommendations brief and resilience strategy) is relevant; and a smoking gun, too.

The third part of theory (interview protocols have been instrumental to research resulting in recommendations brief) is different and perhaps more interesting; while some word-by-word matching between the two documents is again a smoking gun and would have considerably strengthened the theory if observed, in reality it was not observed and the evaluators had to look elsewhere for strong evidence. We were given access to the shared file system and noticed that the protocols were last accessed shortly before the fieldwork was said to have begun. If they used said material during the fieldwork, there's a chance they saved the files on their laptop and no longer had to access them for the duration of it, or certainly after the fieldwork had been completed, when the material was no longer needed. So, whatever they used those protocols for, it happened around the time they actually did the fieldwork. If the theory is true and they used it, we would expect to see access around that time. In theory they could have accessed the material again for a similar study, but we are not aware of any such study. This brings our estimation for the sensitivity between 0.7 and 0.8, or in other words we are cautiously confident that we would have made the observation if the theory was true. If the theory isn't true, what's the chance of the researchers having accessed the documents for the last time around the fieldwork start time? On the basis of our experience carrying out similar work, we would agree that there's a very good chance that the same people taking part in the research would not consult material during or shortly before the fieldwork that they actually did not use during it. The chances are probably not extremely high but at least

on the basis of personal experience we are confident (something between cautiously and highly confident) that this would not have happened if the theory were not true, levelling out to around a 0.2 value of the type I error[30] (see section 2.4 for more information on these numerical estimates and section 2.5 for the qualitative estimates).

[30] With these values our confidence in that theory component would increase to 0.79 upon observation.

Table 7: Bayes formula/confusion matrix values for different observations and different theories

Theory	Sensitivity, Type I error, LR, and Weight of Evidence of O_1	Sensitivity, Type I error, LR, and Weight of Evidence of O_2	…	Sensitivity, Type I error, LR, and Weight of Evidence of O_n
One (T_1)	S of O_1 for T_1 (e.g., 0.9) T1E of O_1 for T_1 (e.g., 0.1) LR of O_1 for T_1 (e.g., 9) Log (LR) of O_1 for T_1 (e.g., 2.2)	S of O_2 for T_1 (e.g., 0.5) T1E of O_2 for T_1 (e.g., 0.8) LR of O_2 for T_1 (e.g., 0.6) Log (LR) of O_2 for T_1 (e.g., -0.5)		S of O_n for T_1 (e.g., 0.05) T1E of O_n for T_1 (e.g., 0.5) LR of O_n for T_1 (e.g., 0.1) Log (LR) of O_n for T_1 (e.g., -2.3)
Two (T_2)	S of O_1 for T_2 (e.g., 0.6) T1E of O_1 for T_2 (e.g., 0.15) LR of O_1 for T_2 (e.g., 4) Log (LR) of O_1 for T_2 (e.g., 1.4)	S of O_2 for T_2 (e.g., 0.1) T1E of O_2 for T_2 (e.g., 0.7) LR of O_2 for T_2 (e.g., 0.14) Log (LR) of O_2 for T_2 (e.g., -1.9)		S of O_n for T_2 (e.g., 0.9) T1E of O_n for T_2 (e.g., 0.4) LR of O_n for T_2 (e.g., 2.2) Log (LR) of O_n for T_2 (e.g., 0.8)
…	…	…	…	…
Theory P (T_p)	S of O_1 for T_p T1E of O_1 for T_p LR of O_1 for T_p Log (LR) of O_1 for T_p	S of O_2 for T_p T1E of O_2 for T_p LR of O_2 for T_p Log (LR) of O_2 for T_p	…	S of O_n for T_p T1E of O_n for T_p LR of O_n for T_p Log (LR) of O_n for T_p

Table 8: Ontological objects (theories) vs. empirical observations in the urban crises example

	Word-by-word matching between draft toolkit and interview protocols	Word-by-word matching between interview protocols and recommendations brief	Word-by-word matching between recommendations brief and resilience strategy	Timeline of access to interview protocols (last accessed around the time field work has been said to begin)
Published knowledge product has influenced local government's resilience strategy	S – ; T1E – Irrelevant, Straw-in-the-Wind (Prior roughly the same as posterior)	S – ; T1E – Irrelevant, Straw-in-the-Wind (Prior roughly the same as posterior)	S – ; T1E – Irrelevant, Straw-in-the-Wind (Prior roughly the same as posterior)	S – ; T1E – Irrelevant, Straw-in-the-Wind (Prior roughly the same as posterior)
Draft knowledge product has influenced interview protocols	S – T1E LOW Smoking Gun (Posterior considerably higher than prior)	S – ; T1E – Irrelevant, Straw-in-the-Wind (Prior roughly the same as posterior)	S – ; T1E – Irrelevant, Straw-in-the-Wind (Prior roughly the same as posterior)	S – ; T1E – Irrelevant, Straw-in-the-Wind (Prior roughly the same as posterior)

	Word-by-word matching between draft toolkit and interview protocols	Word-by-word matching between interview protocols and recommendations brief	Word-by-word matching between recommendations brief and resilience strategy	Timeline of access to interview protocols (last accessed around the time field work has been said to begin)
Interview protocols have been instrumental to research resulting in recommendations brief	S – ; T1E – Irrelevant, Straw-in-the-Wind Prior roughly the same as posterior	S – T1E LOW Smoking Gun Posterior considerably higher than prior	S – ; T1E – Irrelevant, Straw-in-the-Wind Prior roughly the same as posterior	S 0.75? T1E 0.2? Between Straw-in-the-Wind and Doubly Decisive If prior is 0.5, posterior increases to 0.79
Recommendations brief influenced local government's resilience strategy	S – ; T1E – Irrelevant, Straw-in-the-Wind Prior roughly the same as posterior	S – ; T1E – Irrelevant, Straw-in-the-Wind Prior roughly the same as posterior	S – T1E LOW Smoking Gun Posterior considerably higher than prior	S – ; T1E – Irrelevant, Straw-in-the-Wind Prior roughly the same as posterior

3.2.2 Improving wildlife management in Uganda

Table 9 reports five different theories/theory components and six different empirical observations. For the first (sub)claim, "PCLG undertook research with partners and discovered that continued illegal activities in the park were rooted in dissatisfaction with distribution of conservation resources", the only two relevant pieces of evidence are the first (observation of the research report and related matching content) and the second (acknowledgement of report and matching report content in conversations tracked by emails or meeting minutes). The existence of the report is a doubly decisive test – we expect to see it if the theory is true, and it wouldn't exist unless the latter was true. Observations of conversations mentioning the research and its content can be Smoking Guns if the matching is specific (they wouldn't happen without the research); but we wouldn't necessarily expect to observe them just because the research has been carried out.

The second theory component (PCLG used above findings to request to UWA that the shared fee be raised from $5 to $10) requires that the research is carried out so it would be considerably weakened by the observation of an alternative research report presenting the same findings, or by the mentioned research report not being found (hoop test). The acknowledgement of the research in conversations is relatively disconnected from this claim – if the latter is true, we wouldn't necessarily expect that these findings were discussed in conversations although that would perhaps be more likely than not; similarly, if the theory were not true, we could still find evidence of these discussions: the findings would have been discussed without the request being made. A convincing confirmation of this claim would be the observation of the actual letter proposing and motivating this change (smoking gun because, while we wouldn't observe this if the theory weren't true, we wouldn't necessarily expect to see it; or the request be made in this way if the theory were true). Access to UWA board meeting minutes

where a PCLG member proposes the specific change and ideally motivates it on the basis of the research findings would also be a smoking gun (would not be observed if the theory weren't true; and might not be observed if the theory is true because the request might have come in the form of a letter that is subsequently not discussed).

The third claim states that PCLG contributed to the decision by influencing its content and perhaps its timing. The existence of the research report in itself is not directly related to the claim, which means this test is a straw-in-the-wind for this particular claim. Conversations acknowledging the research content in relation to the decision (for example how the two were related) would strengthen the theory; they are not particularly expected if the theory is true, but the latter would be harder to dismiss if they were observed because the chance of observing this if the theory isn't true is low. This would make the test some sort of weak smoking gun or strengthening straw. The existence of an alternative research report with similar conclusions, might weaken the theory a bit but not much: it would still be a straw-in-the-wind because, taken by itself, if the theory is true the chance of observing it is perhaps lower than 0.5 but not too much; and if the theory isn't true the report could very well exist although the chance wouldn't be high (perhaps > 0.5 but not much). The two key observations for this claim are the letter from PCLG to the UWA board in which PCLG formally requests the $5 increase in the fee share and meeting minutes reporting the proposal on behalf of a PCLG member during the UWA board meeting. We're not expecting to observe the letter if the theory is true, but if it isn't, the chances of finding such formal request of exactly the change that has been implemented are considerably lower than 0.5, making the observation almost a smoking gun for this theory or at least evidence that is considerably strengthening. In itself, it still does not confirm but it is strong evidence. Even stronger is the formal proposal put forward during the meeting, as recorded in the minutes; we wouldn't necessarily expect PCLG to formally make the request during the meeting although – knowing they have a seat at the table and attending – it would be suspicious if no formal proposal had been

made (so the sensitivity is higher than 0.5 although not much higher). At the same time, if the theory isn't true, it's very unlikely PCLG would be proposing exactly the same change that is eventually approved, although in itself this is still not definitely conclusive if an alternative influence path stemming from parallel research exists (so in itself this would be almost a smoking gun).

The fourth claim, "research used was produced by someone else, not PCLG" would be compatible with the existence of the PCLG report (for which it would be a straw-in-the-wind); but the acknowledgement of PCLG research in conversations, particularly if positive, would strongly weaken the theory. At the same time, if the board didn't use alternative research, (observation of) those conversations would be possible but not expected. The existence of the letter by itself would weaken the theory a bit but not strongly (PCLG could in theory be using that research themselves). The fact that PCLG research is discussed and used to make the request during the UWA board meeting, while no other research is mentioned, would on the other hand, quite convincingly weaken the theory: it would be quite unlikely an occurrence if the theory were true. At the same time if the theory isn't true, there is a small probability of this happening. Finally, the observation of an alternative research report would be a mild hoop test for the theory – not very strong because we might not be guaranteed or expected to find the report if it existed. And if the theory wasn't true and the report had not been used, it could still have existed and be seen, so its observation doesn't confirm the theory either. In order to confirm the theory, we would need to see a positive discussion of this alternative report in conversations and particularly in board meeting minutes.

Finally, for the fifth claim "UWA was considering a similar decision before PCLG completed the research and started lobbying but they had not quite worked out the details and didn't know how urgent it really was", the observations discussed so far are all largely irrelevant because they discuss PCLG actions and involvement. The last observation, "written evidence that the UWA board was considering

a similar change but had not quite worked out the details nor was aware of how urgent it was", on the other hand, is strong confirmatory evidence for the claim. While we wouldn't necessarily expect to find it if the claim is true, if it isn't the observation would be very unlikely, making it a smoking gun.

It should be clearer at this point how different observations have different implications for different theories and theory components; and how different theories (being true or not) have, in turn, different implications for the chances of observing different pieces of evidence. We reprise this example again in section 3.3.

Table 9: Ontological objects and empirical observations in the Uganda wildlife example

	Observation of the Research report and related content	Acknowledgement of research (matching content) in conversations tracked by email or meeting minutes	Letter from PCLG to UWA with specific suggestion	Observation of an alternative research report with similar conclusions	Meeting minutes where PCLG member champions change in the UWA board (and does not mention alternative research as a significant influence)	Written evidence that the UWA board was considering a similar change before the lobbying but had not quite worked out the details nor was aware of how urgent it was
PCLG undertook research with partners and discovered that continued illegal activities in the park were rooted in dissatisfaction with distribution of conservation resources	S high T very low (Doubly Decisive)	S – T very low (Smoking Gun)	Irrelevant S – T – (Straw-in-the-Wind)	Irrelevant S – T – (Straw-in-the-Wind)	Irrelevant S – T – (Straw-in-the-Wind)	Irrelevant S – T – (Straw-in-the-Wind)

	Observation of the Research report and related content	Acknowledgement of research (matching content) in conversations tracked by email or meeting minutes	Letter from PCLG to UWA with specific suggestion	Observation of an alternative research report with similar conclusions	Meeting minutes where PCLG member champions change in the UWA board (and does not mention alternative research as a significant influence)	Written evidence that the UWA board was considering a similar change before the lobbying but had not quite worked out the details nor was aware of how urgent it was
PCLG used above findings to request to UWA that the shared fee be raised from $5 to $10	S very high T – Hoop test	S not much > 0.5 T not much < 0.5 Straw-in-the-Wind	S not bad T very low (Smoking Gun)	S low T – (would weaken the theory if observed)	S not bad T very low (Smoking Gun)	Irrelevant? S – T – (Straw-in-the-Wind)

	Observation of the Research report and related content	Acknowledgement of research (matching content) in conversations tracked by email or meeting minutes	Letter from PCLG to UWA with specific suggestion	Observation of an alternative research report with similar conclusions	Meeting minutes where PCLG member champions change in the UWA board (and does not mention alternative research as a significant influence)	Written evidence that the UWA board was considering a similar change before the lobbying but had not quite worked out the details nor was aware of how urgent it was
PCLG contributed to the decision taken by UWA, by influencing the content and perhaps the timing	irrelevant S – T – (Straw-in-the-Wind)	S low T mid-low Strengthening straw?	S – T relatively low Almost a smoking gun	Not very relevant in itself S perhaps a bit low (<0.5 but not much) T not very high though (perhaps >0.5 but not much) Straw-in-the-Wind	Did it cite PCLG research? If yes: S not bad T quite low Weak Doubly Decisive	S – T – Straw-in-the-Wind

	Observation of the Research report and related content	Acknowledgement of research (matching content) in conversations tracked by email or meeting minutes	Letter from PCLG to UWA with specific suggestion	Observation of an alternative research report with similar conclusions	Meeting minutes where PCLG member champions change in the UWA board (and does not mention alternative research as a significant influence)	Written evidence that the UWA board was considering a similar change before the lobbying but had not quite worked out the details nor was aware of how urgent it was
Research used was produced by someone else… this is tricky… it was clear they used their research, but what if PCLG had copied the research from somewhere else?	Irrelevant S – T – (straw-in-the-wind)	S low T – Absence is almost a Hoop test	Not very relevant S not much < 0.5 T not much > 0.5 (Straw-in-the-Wind)	S high T – Almost Hoop test	If PCLG hadn't copied the research from someone else: S very low T – Observation strongly weakens the theory	Some kind of Straw-in-the-Wind

	Observation of the Research report and related content	Acknowledgement of research (matching content) in conversations tracked by email or meeting minutes	Letter from PCLG to UWA with specific suggestion	Observation of an alternative research report with similar conclusions	Meeting minutes where PCLG member champions change in the UWA board (and does not mention alternative research as a significant influence)	Written evidence that the UWA board was considering a similar change before the lobbying but had not quite worked out the details nor was aware of how urgent it was
UWA was considering a similar decision before PCLG completed the research and started lobbying but they had not quite worked out the details and didn't know how urgent it really was	Irrelevant S – T – (Straw-in-the-Wind)	Almost irrelevant S – T relatively high (Straw-in-the-Wind)	Not very relevant S kind of low T – (straw-in-the-wind) Low sensitivity weakens the hypothesis if evidence is observed	Irrelevant S – T – (Straw-in-the-Wind)	S quite low T not very low (straw-in-the-wind at best) Low sensitivity weakens the hypothesis if evidence is observed	Smoking Gun

3.2.3 Explaining deforestation trends

We continue to illustrate how evidence is assessed against theories and theory components with the forestry policy example (Brandao & Befani, 2021). For brevity we only report on the first theory component here, which reads as follows: "Improved environmental monitoring and enforcement, Credit restrictions, and the Cattle Agreements, instituted sequentially between 2007-08 and 2008-09 acted to trigger behavioural change of local actors involved in deforestation, leading them to substantially reduce deforestation activities. The interventions mostly targeted and hence affected medium and large actors which contributes to explain the major drop".

The theory has various subcomponents which we test one by one, as outlined below. We do not create a matrix here because each observation is relevant for one theory only. The first line of indentation describes the evidence and the second indented line includes some considerations on its strength and direction.

1. BOI PIRATA (and subsequent similar monitoring and enforcement operations over two or three years) changed the behaviour of large and medium holders because it made deforesting riskier and substantially contributed to the drop
    a. The Ministry of Environment of Brazil claimed in a public interview that Boi Pirata was the first operation held in the territory after the law revision in 2008 and the first one seizing cattle. Similar statements have been made by the Brazilian Ministry of Environment.
        i) Cautious confidence if true, but it's possible that they would want to have the credit and didn't consider alternatives – more confident than not they wouldn't say that if theory is not true. Seems largely a straw-in-the-wind.

b. Several media sources reported that several ranchers removed cattle from their properties at APATX during the Boi Pirata operation, which showed they were wary of its consequences.

    i) Seems likely if theory is true, and unlikely that several ranchers at the same time would remove their cattle if the theory is not true – the long lines of cattle running away were a fairly unprecedented sight. Perhaps not a doubly decisive, but since it's quite symmetrical it seems a strong straw-in-the-wind.

c. The Operation Boi Pirata is the only individual field base operation mentioned by interviewees as relevant in answer to a question about the role of individual interventions to reduce deforestation in the region.

    i) Possibly more likely than not if the theory were true; and more unlikely than not if it weren't true – not lower as the informants might be just relaying general wisdom or common sense. Seems a typical, slightly strengthening straw-in-the-wind.

d. The correlation between the number of embargoes and deforestation rates between 2005 and 2019 is relatively high (Adjusted R-squared: 0.5453). For every additional embargo, 459 hectares of forest seems to have been saved ($p = 0.01545$**).

    i) There would be an expectation if the theory is true, but correlation does not imply causation, so this could be some kind of hoop test.

2. CREDIT RESTRICTIONS were effectively providing a major incentive for large holders not to deforest, as credit was successfully restricted

    a. Data showing a major decline in credit allocation for large holders in 2009.

i) Expected if the theory is true; if the theory is not true it's difficult to imagine that deforesting behaviour could have remained unaltered while credit was so fundamentally restricted, so the strengthening power is relatively high. This is a candidate for a doubly decisive.

b. Bank representatives said the law changed the way they worked: monitoring was enhanced and more information requested; they explained how it became harder to lend to large holders

   i) This is expected if the theory is true, and it's unlikely to happen if the theory is not true, even though the bankers might have been giving a politically correct answer while trying to find loopholes. The data shows that credit to large holders was effectively reduced. A Hoop test with some confirmatory power, but not quite a doubly decisive in itself.

c. The correlation between credit (number of contracts per year) and deforestation rates between 2005 and 2019 is relatively high (Adjusted R-squared: 0.7487). For every additional credit contract, 8 hectares of forest have been chopped down (p: 3.93e-05***).

   i) This is a stronger hoop test than the one above

d. Credit to medium-large holders was more severely restricted than smallholder credit.

   i) We would expect this If the theory is true; if the theory not true, it would be relatively unlikely to observe such strong restrictions but especially more specific restrictions on large holders. Seems like a hoop test with a bit of strengthening power.

3. The CATTLE AGREEMENTS pressured slaughterhouses not to buy cattle from deforesters and substantially reduced the deforesting behaviour of large holders when they came into effect.

a. All major slaughterhouses buying cattle from SFX signed the TAC Agreement.

    i) This is expected if the theory is true, but it's almost as likely if the theory is not true because they could have signed for political correctness or to avoid the legal consequences. Seems largely a straw-in-the-wind.

b. Several statements from ranchers to the effect of "after the Cattle Embargo there was no way to continue as before", they didn't say the same about other policies.

    i) Mostly expected if the theory is true, more unlikely than not if the theory isn't true. Again, a Hoop with a bit of strengthening power.

c. Slaughterhouses started to invest money in the regularization of their supply-chain (more than 1M $USD were invested by companies in that period), initially to implement an information system (this was a requirement that was introduced by the Cattle Agreement). But it's more than a mere signature: it might signal that they experienced a shortage of supply and needed to invest in regularization to continue to be able to buy.

    i) Again, this is expected if the theory is true, and it's unexpected if the theory is not true. It's unlikely that the companies would have invested if they didn't need it to keep their business alive. Not quite a doubly decisive but a Hoop test with some solid strengthening power.

d. Two important municipal meetings, one before and one after the intervention, seem to signal the pivotal role of the Cattle Agreements. By 9th April 2009 (before the intervention), a first meeting took place gathering hundreds of representatives of local organizations and representatives of external actors such as NGOs and federal agencies. The meeting turned out to be an assembly against environmentalists, NGOs, and anti-deforestation efforts in general. However, one month later and after the Cattle

Embargo on the 1st of May (the first step leading to the Cattle Agreement), a second meeting took place and by that moment the situation had completely changed. Local organizations, in particular rancher representatives, (peacefully) took part in discussions about the anti-deforestation agenda. There was only one month between the two meetings.

   i) We wouldn't necessarily expect to see this if the theory is true, but if it's not true, such a big difference in only a month's time seems quite unlikely. This is a candidate for a smoking gun.

4. We also tested if other factors could have substantially contributed to the drop as follows: "the decrease of beef prices substantially contributed to the drop"
   a. There is a relatively strong negative correlation between beef prices and deforestation rates between 2005 and 2019 (Adjusted R-squared: 0.3376, p-value: 0.01723**).
      i) If the theory is true, we would expect to see a positive correlation rather than a negative one, while correlation doesn't strengthen the causation hypotheses in itself. It's a hoop test but we don't observe the expected evidence, so it fails.
5. Temperature change substantially contributes to the drop
   a. Similarly to the above, there is no significant correlation between temperature and deforestation rates between 2005 and 2019
      i) If the theory were true, we would expect to see a correlation – failed hoop test as above.
6. Precipitation substantially contributes to the drop
   a. Again, no significant correlation between precipitation and deforestation rates between 2005 and 2019
      i) As above, another failed hoop test.

In section 3.3.3.2 we outline how we formally estimated confidence levels for the theory components and the theory as a whole.

3.2.4 Improving the energy capacity market

In this evaluation, the final form of the theory was a series of CMO configurations (Table 10). The contexts and outcomes were relatively easy to assess, while evidence for the causal mechanisms, or the reasoning behind actors' behaviour and choices, was gathered mostly from interviews. Below are some examples of observations that were targeted and assessed by the evaluation team. All observations have been deemed to have weakening power for the associated mechanisms (had they not been made) and, at the same time, three were quite specific to the mechanisms, too (hence had strengthening power, too) so were considered doubly decisive (while the remaining two merely hoop tests).

While confidence on the mechanism is important, the overall assessment for the CMO configuration depends on confidence about the other two parts as well: context and outcome. If these parts of the theory are assessed separately, it's important that overall confidence in the CMO configuration does not exceed the lowest level of confidence achieved for the single parts (Befani, D'Errico, Booker, & Giuliani, 2016; D'Errico, Befani, Booker, & Giuliani, 2017).

Table 10: CMO configurations for the Energy Capacity Market evaluation

Broader theory	Mechanism	Observations	Type of Test
The second TA contributes to more and/or more competitive flexible capacity for the capacity market in 2018-19 and subsequent years	Our experience of participating in the second TA means the capacity market seems less risky	The participant says in interview that they now have more confidence in being able to meet CM rules and regulations/be competitive in other CM auctions as a result of their participation in the second TA (e.g., because they developed skills/strategies/learning)	DD
	In order to participate in the second TA, we invested in capacity or the ability to provide capacity which will make us better positioned to participate in the main CM	Second TA participant saying in interview that they or their clients have developed or invested in assets (e.g., controls/metering) for the second TA that reduce costs of participation in future CM	HT
	(for new entrants) In order to participate in the second TA we have built a customer base and so now we want to continue with the CM	The participant saying in interview that they have developed markets (e.g. building a client base, entering the UK market) for the second TA that they plan to use in one or more main CM auctions	HT
The second TA made no difference to the capacity available to	(for existing aggregators) We have always intended to participate in the	Existing aggregators and direct participants state in the interview that they would have invested in, or maintained, capacity for	DD

Broader theory	Mechanism	Observations	Type of Test
the CM in 2018/19 and subsequent years and therefore is not additional	CM and the TA did not help us to grow our flexibility business.	future CM auctions regardless of the TA.	
	(for new entrants) We are a new entrant to flexibility in the CM but would have started participating with flexible capacity in the CM at the same level anyway, because of other changes, not the TA	TA participants new to flexibility in the CM state in the interview that they would have invested in, or maintained, the same level of flexible capacity for future CM auctions regardless of the TA	DD

3.2.5 Organising multiple observations and working with evidence packages

To some extent, the considerations made in the above section might sound unfamiliar because they refer to the assessment of single observations, as if we collected only one piece of evidence for each theory; and as if, every time we made a new observation, we forgot the previous one. However, in any given evaluation, it's very unlikely we will observe only one piece of evidence for each theory: normally, our data collection and desk reviews produce several individual interviews where key informants make several statements; include documents with extensive content, perhaps focus groups (as, indeed, the real-life examples considered so far show). Similarly, in medical diagnosis, the physician might consider a combination of symptoms, blood tests, and diagnostic images like x-rays, MR scans, etc. That's

why it's important to address evidence packages, or the combined analysis of multiple observations.

In this section we present multiple ways of approaching the assessment of evidence packages. We draw on these insights, for example, when we need to assemble the results of a small survey or a set of interviews, or in general evidence stemming from different individuals or organisations. If we have enough resources to carry out formal Bayesian Updating, we have four options:

1. If we manage to define some observations as independent from each other, we can multiply their single piece-estimates of Sensitivity and Type I error to obtain the Bayes formula values for the whole package;
2. If we can't draw independence boundaries among observations, we can:
    a. consider a group of (inter-dependent) interviews, documents, etc. as one single piece of evidence;
    b. apply the above multiplication procedure but with discount coefficients that take inter-dependence into account in order to reduce the probative value;
    c. calculate the conditional probabilities of observing each single piece of evidence after having observed others (under the two hypotheses of the theory being true and not being true).

If probability estimates are not available or can't be obtained, we can create rubrics defining different degrees of empirical support for the theory (option five – illustrated in section 3.3.3.4).

3.2.5.1 The notion of stochastic independence

Option #1 and option #5 require stochastic independence. One observation is stochastically independent from another if observing it does not alter the probability of observing the other: the

conditional and unconditional probabilities are the same, where the conditioning is intended to apply to the observation, and not (only) on whether the theory is true or not. In medical diagnosis, we can argue that scan machines and lab processes do not interact; findings from blood tests are independent from findings from imaging tests once the diagnosis is known. Another example of independent events is when we interview informants whose opinions have not been influenced by each other.

In general, if observations are independent under the theory, it does not necessarily imply that they're independent under alternatives to the theory. For example, if employees of an organisation are interested in showing that the theory is true but (unfortunately for them) it isn't, they might have more incentives to present an agreed, positive view during interviews, compared to a situation where the theory is actually true, and they are more relaxed about agreeing on what to say during evaluation related interviews.

How do we assess stochastic independence among a series of empirical observations? We consider the probabilities of observing the different pieces of evidence under different assumptions concerning each theory (true or not true) and assuming we have made one or more of the other observations. We won't necessarily need to assemble pieces of evidence for every theory: for example, for the first theory (component) of our Uganda wildlife evaluation (Table 9), we have a doubly decisive test and most other observations are irrelevant to the theory. It's a similar situation for the fifth theory: there is only one key piece of evidence.

Below we illustrate how we carried out this task for the Uganda wildlife examples and for the forestry policy evaluation.

3.2.5.2 Assessing observations' independence in the Uganda wildlife evaluation

For the second claim (PCLG used above findings to request to UWA that the shared fee be raised from $5 to $10) we have two smoking guns (Table 9); if we observe either our confidence becomes quite high, but what happens if we observe both? Is our confidence the same (as in, the second observation doesn't increase it further), or is it even higher? To answer this question, we need to ask if the two smoking gun observations are stochastically independent; or, in other words, if making one of the two alters the chances of making the other, conditioned on the theory being true or not. If PCLG used the findings to make that specific request, and we see the letter, what are the chances of observing the PCLG member making the same request during the meeting? The second event is not necessarily implied from the first: it's possible that only the formal request could have been made, without it being picked up during the meeting. At the same time, if we know about the letter and the theory is true, we wouldn't be surprised if the same group (and especially if it's the same person) reiterates the request at the board meeting. So, we cannot argue for complete independence but at the same time the two events are quite distinct and do not necessarily imply each other. If the theory isn't true, we would be surprised by the first observation and even more by the second; we would still not expect the second even though we have observed the first, knowing that the theory isn't true; it's easier to argue for independence under the assumption that the theory isn't true.

For the broader claim "PCLG contributed to the decision taken by UWA, by influencing the content and perhaps the timing", let's try to consider the tests sequentially and see what insights we can draw. If the theory is true and we have observed the report, does this affect the chances of the findings being discussed by the relevant stakeholders? Probably slightly as we now know the report exists, although its mere existence does not imply that it will be discussed. If the report exists and it's discussed, does this affect the chances of

observing the formal letter? Probably not in a substantial way unless the conversations were somehow anticipating the letter submission. If we see all of this plus the letter, do we predict that the UWA will pick up the request during the board meeting? Not necessarily, as this signals a commitment that the other tests were not necessarily indicating. The discovery of the alternative report is perhaps even more clearly independent, as is written evidence of what the UWA board had in mind before PCLG started its lobbying activities on the issue. We can make a cautious case for considering these tests independent if the theory is true, and perhaps use a mitigating factor when independence is not quite as clear cut as in other cases (see section 3.3.3.3 for calculation details).

As for predictions if the theory isn't true and we have observed the report, what can we say for the conversations acknowledging the research results? We might find that odd knowing the theory isn't true, but the research report might not affect our assessments by much. Observing the letter and the proposal in the board might puzzle us but the chances of observing those would be more fundamentally affected by the theory not being true than by the other observations, as we might simply assume that they're related to some process that has not resulted in influence. The two other tests can also be considered largely independent and fundamentally influenced by the condition of the theory not being true rather than by other observations.

3.2.5.2 Assessing observations' independence in the forestry policy example

For this example, we have reported three major claims (section 3.2.3), for each of which we have four relevant pieces of evidence. The strongest argument for mutual independence of all the pieces of evidence in relation to each other can be made for the third claim:

"the Cattle Agreements pressured slaughterhouses not to buy cattle from deforesters and substantially reduced the deforesting behaviour of large holders when they came into effect".

The evidence we have for the claim is that: 1) all major relevant slaughterhouses signed the Agreement; 2) several ranchers expressed the almost impossibility of continuing as before after the Cattle Embargo (the only policy to draw such judgement); 3) slaughterhouses invested money (more than 1M USD) to create information infrastructure to comply with the information requirements of the new law; and 4) Two municipal meetings taking place in one month signalled a mood shift with regard to deforestation culture.

One can argue that the first two claims are independent because they concern different groups who would not necessarily cooperate or present a united front; the third is independent from the first because signature of the agreement can be a formality and doesn't imply that companies make actual financial investment to comply with the law in practice; and for the second and third we can make the same argument as for the first and second (different groups). We can make it when we compare the fourth with the first and third, too; for the second and fourth we can argue that the groups involved in the two pieces of evidence did not know each other.

However, the issue is not so simple because the assessments need to be made under the two assumptions that the theory is true and that it isn't; and they might differ. If we know that the theory is true, we might not be surprised to find further theory-friendly evidence after one or two pieces of strengthening observations; so the multiplication reduces the sensitivity more that it should. Conversely, under the assumption that the theory isn't true, it would be more puzzling to find multiple pieces of theory-friendly evidence than without the assumption; so the multiplication reduces the Type I error less than it should. What we can say is that, by using the multiplication, we do not overestimate the probative value of the

evidence unless the additional evidence is expected under the assumption that theory isn't true. In all other cases we might actually be underestimating it.

We should perhaps mention the order or sequence with which the evidence is observed. If the theory is true, we would expect slaughterhouses to sign the agreement because it would be costly not to (say 95%), and once we have observed that, we would have a lower expectation of them actually making investments to respect it (say 80%). If we observe the investment first, our expectation of it initially could be maybe 77% or lower; and once we have seen this, we would be practically certain that they've signed the agreement, too (99%). The compounded probability is 76% in both cases. Even when it's not exactly the same number expressed in %, it should be roughly the same, and the sequence shouldn't make a substantial difference to the overall composite probability, no matter where we start from. Table 11 illustrates the arguments in favour of independence between the various pieces of evidence considered for this theory component.

Table 11: Stochastic independence arguments for various combinations of five observations against one theory

	All major relevant slaughter-houses signed the Agreement	Several ranchers expressed the almost impossibility of continuing as before after the Cattle Embargo	Slaughter-houses invested in information infrastructure to comply with the information requirements of the new law	Two municipal meetings taking place in one month signalled a mood shift with regard to deforestation culture
All major relevant slaughter-houses signed the Agreement		If T true Different groups who would normally not be expected to cooperate or present a united front	If T true Signature of the agreement can be a formality and doesn't imply that companies make actual financial investment to comply with the law in practice	If T true Different groups who would normally not be expected to cooperate or present a united front
Several ranchers expressed the almost impossibility of continuing as before after the Cattle Embargo	If T not true Different groups who would normally not be expected to cooperate or present a united front		If T true Different groups who would normally not be expected to cooperate or present a united front	If T true The groups involved in did not know each other or collaborate

Slaughter-houses invested in information infrastructure to comply with the information requirements of the new law	If T not true They could be using the infrastructure for undesirable purposes, but perhaps not very surprising that they would sign the agreement too	If T not true Different groups who would normally not be expected to cooperate or present a united front		If T true Different groups who would normally not be expected to cooperate or present a united front
Two municipal meetings taking place in one month signalled a mood shift with regard to deforestation culture	If T not true Different groups who would normally not be expected to cooperate or present a united front	If T not true The groups involved in did not know each other or collaborate	If T not true Different groups who would normally not be expected to cooperate or present a united front	

3.3 Step Three: Estimating the Bayes formula values and updating confidence

When we complete the previous phase, we are at a stage where we have organised the observations against the various theories we're testing and have established qualitative levels of confidence for Sensitivity and Type I error; in addition, we have assessed the extent to which we can consider the relevant observations stochastically independent.

We can now proceed to formal confidence updating by estimating numerical values of Sensitivity and Type I error, inputting them into the Bayes formula, and calculating the posterior, post-observation level of confidence[31]. This can be undertaken for different theories and statements[32], which allows us to see which ones are most strongly supported empirically. In Table 2, for example, we would choose or prioritise Theory Two because the posterior confidence value associated to it is the highest.

In reality, we almost always handle multiple observations and Table 12 shows the posterior confidence values from two hypothetical examples (the first two theories of Table 7). If we look at the first theory, it's supported by O_1 and weakened by O_n; while the second theory is supported by O_1 and weakened by O_2. As for pieces of evidence, O_1 supports both theories; O_n mildly supports the second and strongly weakens the first theory.

Table 12: Posteriors showing how different observations affect confidence in different theories (from a prior of 0.5)

Theory	Posterior after observing O_1	Posterior after … observing O_2	Posterior after observing O_n
One (T_1)	$P(T_1\|O_1)$ (0.90)	$P(T_1\|O_2)$ (0.38)	$P(T_1\|O_n)$ (0.09)
Two (T_2)	$P(T_2\|O_1)$ (0.80)	$P(T_2\|O_2)$ (0.12)	$P(T_2\|O_n)$ (0.69)

In the immediate next sections, we illustrate how we updated the prior confidence into the posterior for two of our examples that did not require assembling pieces of evidence. Further below we show what we did when this was not the case.

[31] For the practical calculations, we recommend using this tool (Befani, 2017)

[32] Unlike Fairfield and Charman (2017), we do not require the multiple compared theories to be always and necessarily mutually exclusive. It is not rarely the case in evaluation that multiple explanations of the outcome co-exist. Each theory is compared with its opposite, but for example, in policy influence, there can be multiple sources of influence; and multiple influence processes can co-exist.

3.3.1 Assessing our confidence that the GA Municipality was influenced by the toolkit

In this case we had three theory components that were strongly supported by the evidence. We outline them below with the respective relevant observations and their strength assessments.

1. The draft knowledge product has influenced the interview protocols allegedly used in the research (practical certainty)
    a. Word-by-word matching between draft toolkit and interview protocols: we matched 15 interview protocols of 3 different types with the draft toolkit's templates and found variable but extensive amounts of seemingly copy-pasted text in 11 of them (see table below). The specific reasoning behind the estimates is reported in (Befani & D'Errico, 2020) but they are essentially proportional to the amount of matching text, and they are all very strong smoking guns.

Table 13: Bayes formula values for different groups of interview templates

Template	Prior	Sensitivity	Type I Error	Posterior	Cases used
Business	0.5	0.05-0.10	0.00050	0.991-0.995	1
NG/CS	0.5	0.40-0.50	0.00008	0.998	5
Municipal	0.5	0.20	0.00001	0.999	5

2. The interview protocols have been instrumental to the research that has eventually resulted in the recommendations brief (cautious confidence)
    a. Timeline of access to interview protocols (last accessed around the time field work has been said to begin): for this component, we argued (Befani & D'Errico, 2020) that the sensitivity was around 0.75 and the Type I error around 0.2; which from a prior of 0.5 would return a posterior confidence of 0.79 (cautious confidence). It seems intuitive

that this was the case from the interviews and the context, but we could not find stronger evidence for this part of the claim.

3. The recommendations brief influenced the local government's resilience strategy: in particular, GAM incorporated the advice and recommendations into the resilience strategy (practical certainty).
    a. Word-by-word matching between recommendations brief and resilience strategy: the evaluation team compared the text of the suggested adaptations in the recommendations brief with the text in two versions of the draft strategy and the final strategy. While not present in the version preceding the assessment, both the intermediate and the final version of the strategy include the suggested adaptations, with the entirety of the content (with a minor exception at the end) seemingly being copy-pasted from the recommendations brief. The probability that such an extensive amount of text (650 words) was coincidentally identical if there had been no influence is probably around one on 100,000, which sets the Type I error at 0.00001. Under the hypothesis of influence, the probability of copying such an extensive amount of text is not very high, but still infinitely higher than under the hypothesis of no influence (the Type I error). We proposed to set it (the Sensitivity) at one in one hundred, or 0.01. We cannot think of any other document the authors or IRC could have copied the text from, it looks like original work by all accounts. With these values and the prior set at 0.5, the Bayes formula returns a posterior of 99.9%, meaning practical certainty that GAM's resilience strategy was influenced by the recommendations brief.

3.3.2 Assessing our confidence about vaccination efficacy

In our influenza simulation example (Befani, Elsenbroich, & Badham, 2021), we analyse the implications of the three efficacy levels of protective behaviour in terms of observable empirical evidence. The Agent-Based Model can simulate infection rates under a variety of settings for different efficacy levels of vaccination. Setting the model to represent the three efficacy hypotheses, we obtain three different probability distributions for the final proportion of population ever infected. The expected values are, for efficacy levels of 80%, 90%, and 100%, respectively 0.33, 0.30, and 0.27.

These distributions overlap to some extent; for infections rates higher than 0.33 or lower that 0.26, we know there is only one efficacy level these rates are compatible with. But between 0.26 and 0.33 there is uncertainty: namely between 0.275 and 0.31, infection rates are compatible with all three efficacy levels (or all three theories). The table below illustrates the likelihoods of infection rates falling into five ranges of values.

Table 14: Probabilities of observing ranges of proportions of infected populations by levels of efficacy

	Proportion of population ever infected (evidence)				
Level of efficacy (theory)	<=0.275	>0.275 & <=0.29	>0.29 & <=0.30	>0.30 & <=0.31	>0.31
Ideal 1.0	0.53	0.30	0.13	0.04	0
Improved 0.9	0.09	0.27	0.26	0.24	0.14
Standard 0.8	0	0.02	0.01	0.03	0.94

Source: Befani, Elsenbroich & Badham (2021)

Table 15 builds on Table 14 to calculate Sensitivity and Type I error values of three ranges of empirical observations for each of the efficacy theories. You can see that the middle range is a Straw-In-The-Wind for the middle efficacy theory, while the low range is a smoking gun for ideal efficacy and the high infection range is a doubly decisive for standard efficacy.

In this particular example, we have used a social simulation model to estimate probabilities, and we previously mentioned how we would ideally like to draw on empirical frequencies to conduct probability estimates. However, in most evaluation cases these probabilities will have to be estimated subjectively and with the help of experts and stakeholders (see Section 2.4).
As anticipated at the beginning of Section 3.3, we often work with multiple observations; we thus need to understand which probabilities we need, depending on whether we can consider the evidence as one single piece, as multiple independent pieces, or as multiple, inter-dependent ones.

Table 15: Calculating the Bayesian updating values with the priors all set at 0.33

Level of efficacy	Posteriors after observation of evidence	Sensitivity	Type I Error	Likelihood Ratio	Posterior-Prior
Ideal 1.0 (prior = 0.33)	Infected population <= 0.275, posterior = 0.84	0.53	0.05	10.60	0.51
Improved 0.9 (prior = 0.33)	I.P. 0.275 < p <= 0.31, posterior = 0.58	0.77	0.27	2.85	0.25
Standard 0.8 (prior = 0.33)	I.P. > 0.31, posterior = 0.87	0.94	0.07	13.43	0.54

3.3.3 Which probabilities do we need when we handle evidence packages?

In most evaluations we handle multiple observations; in this section we discuss which probabilities are needed and what assessments and considerations need to be made in such situations. We can formalise the evidence package as "$O_1 \cap O_2 \cap \ldots \cap O_n$"; or the combination of empirical observations O_1, O_2, … O_n (Table 16). We can then estimate the Sensitivity and Type I Error of the package in relation to different theories. In practice this will be done differently depending on the situation, for example on whether the pieces of evidence can be argued to be stochastically independent or not.

Table 16: How prior confidence in different theories is affected by evidence packages[33]

Theory	Prior	Posterior
One (T_1)	$P(T_1)$	$P(T_1 \mid O_1 \cap O_2 \cap \ldots \cap O_n) = 0.36$
Two (T_2)	$P(T_2)$	$P(T_2 \mid O_1 \cap O_2 \cap \ldots \cap O_n) = 0.5625$

The general formula for calculating the probability of combined, multiple events requires calculating the probability of making a given future observation on condition of having already made specific ones. For example, if we have already observed O_1, the probability of also observing O_2 is $P(O_2 \mid O_1)$. The probability of subsequently observing O_3 is $P(O_3 \mid O_2 \cap O_1)$, and so on, until $P(O_n \mid O_{n-1} \cap O_{n-2} \cap \ldots \cap O_1)$ which is the probability of observing O_n after having made the previous n-1 observations. For our purposes we would need to calculate the conditional sensitivity values and the conditional Type I error values (Table 17).

[33] The numerical values in this table are obtained from Table 18 (which in turn can be traced to Table 7 and Table 17), starting from a prior of 0.5.

It is possible but quite cumbersome to calculate these conditional probabilities for each piece of evidence and each theory; as we mention above in option 2a (Section 3.2.5), a relatively acceptable shortcut is to consider the whole package as one single piece of evidence and calculate two single values to feed into the Bayes formula.

If observations are independent[34], however, we are provided with a mathematically formalised shortcut since the conditional probabilities will be equal to the non-conditional ones, which are likely to be already available at this stage. We can thus calculate the probability of observing a combination of observations (or the probabilities of a package) by multiplying the probabilities of observing the single pieces (Table 17). In particular, the Sensitivity of a package will be obtained as the multiplication of Sensitivity values of the single pieces; and the Type I Error of the package will be obtained by multiplying the Type I Error values of the single pieces (Tables 17 and 18). Table 18 uses the formulas of Table 17 with the values of Table 7.

If conditioning on the theory being false (as opposed to being true as above) doesn't change our independence assumptions on the evidence pieces, we can calculate the Type I Error of the package in the same way, by multiplying the Type I Error values of the single package components. The posteriors in the last columns of Table 16 are then obtained from the Sensitivity values and the Type I Error values of Table 18. In general, if observations are independent under the theory, it does not necessarily imply that they're independent under alternatives to the theory: so we might be able to use this method to calculate the sensitivity of the package, but not the Type I Error, or vice versa.

---

[34] Formally, two observations O_1 and O_2 are stochastically independent if $P(O_1|O_2) = P(O_1)$ and $P(O_2|O_1) = P(O_2)$. Therefore $P(O_1 \cap O_2)$, which is $P(O_1)*P(O_2|O_1)$ or $P(O_2)*P(O_1|O_2)$, is equal to $P(O_1)*P(O_2)$.

Table 17: Calculating the Sensitivity and Type I Error of evidence packages

Theory	Sensitivity of package ($O_1 \cap O_2 \cap \ldots \cap O_n$) = $P(O_1 \cap O_2 \cap \ldots \cap O_n \mid T)$ and Type 1 Error of package ($O_1 \cap O_2 \cap \ldots \cap O_n$) = $P(O_1 \cap O_2 \cap \ldots \cap O_n \mid \sim T)$
One (T_1)	**S =** $P(O_1 \mid T_1) * P(O_2 \mid O_1 \cap T_1) * \ldots * P(O_n \mid O_{n-1} \cap \ldots \cap O_1 \cap T_1)$ (if observations are independent = $P(O_1 \mid T_1) * P(O_2 \mid T_1) * \ldots * P(O_n \mid T_1)$) **T1E =** $P(O_1 \mid \sim T_1) * P(O_2 \mid O_1 \cap \sim T_1) * \ldots * P(O_n \mid O_{n-1} \cap \ldots \cap O_1 \cap \sim T_1)$ (if observations are independent = $P(O_1 \mid \sim T_1) * P(O_2 \mid \sim T_1) * \ldots * P(O_n \mid \sim T_1)$)
Two (T_2)	**S =** $P(O_1 \mid T_2) * P(O_2 \mid O_1 \cap T_2) * \ldots * P(O_n \mid O_{n-1} \cap \ldots \cap O_1 \cap T_2)$ (if observations are independent = $P(O_1 \mid T_2) * P(O_2 \mid T_2) * \ldots * P(O_n \mid T_2)$ **T1E =** $P(O_1 \mid \sim T_2) * P(O_2 \mid O_1 \cap \sim T_2) * \ldots * P(O_n \mid O_{n-1} \cap \ldots \cap O_1 \cap \sim T_2)$ (if observations are independent = $P(O_1 \mid \sim T_2) * P(O_2 \mid \sim T_2) * \ldots * P(O_n \mid \sim T_2)$)
…	…
Kay (T_K)	**S =** $P(O_1 \mid T_K) * P(O_2 \mid O_1 \cap T_K) * \ldots * P(O_n \mid O_{n-1} \cap \ldots \cap O_1 \cap T_K)$ (if observations are independent = $P(O_1 \mid T_K) * P(O_2 \mid T_K) * \ldots * P(O_n \mid T_K)$) **T1E =** $P(O_1 \mid \sim T_K) * P(O_2 \mid O_1 \cap \sim T_K) * \ldots * P(O_n \mid O_{n-1} \cap \ldots \cap O_1 \cap \sim T_K)$ (if observations are independent = $P(O_1 \mid \sim T_K) * P(O_2 \mid \sim T_K) * \ldots * P(O_n \mid \sim T_K)$)

Table 18: Calculating the Sensitivity and Type I error of evidence packages when observations are independent

Theory	Sensitivity of package ($O_1 \cap O_2 \cap \ldots \cap O_n$) = $P(O_1 \cap O_2 \cap \ldots \cap O_n \mid T)$ and Type 1 Error of package ($O_1 \cap O_2 \cap \ldots \cap O_n$) = $P(O_1 \cap O_2 \cap \ldots \cap O_n \mid \sim T)$
One (T_1)	(S of O_1 for T1)* (S of O_2 for T_1)*…* (S of O_n for T_1) = = 0.9 * 0.5 * 0.05 = 0.0225
	(T1E of O_1 for T_1)* (T1E of O_2 for T_1)*…* (T1E of O_n for T_1) = = 0.1 * 0.8 * 0.5 = 0.04
Two (T_2)	(S of O_1 for T_2)* (S of O_2 for T_2)*…* (S of O_n for T_2) = = 0.6 * 0.1 * 0.9 = 0.054
	(T1E of O_1 for T_2)* (T1E of O_2 for T_2)*…* (T1E of O_n for T_2) = = 0.15 * 0.7 * 0.4 = 0.042

3.3.3.1 *Combining probabilities to test policy influence theory in Uganda*

For the first component of Table 9, now updated as in Table 19, we have two relevant observations, both quite strongly confirmatory in themselves. If we assume they are independent, we can multiply their sensitivities (0.9 and 0.5) and obtain 0.45 for the sensitivity of the package; and their Type I errors (both lower than 0.01) to obtain a value that is lower than 0.0001. We thus have practical certainty (0.9998) that the theory component is true.

The second component is dealt with in section 3.2.5.2 and section 3.3.3.3; for the third, we can ignore the first observation because it's irrelevant, and if we consider the other four independent (see also section 3.2.5.2), the sensitivity of the package is 0.77*0.50*0.60*0.60 = 0.1386; while the Type I error is 0.28*0.15*0.40*0.05 = 0.00084. With these values input into the Bayes formula from a prior of 0.5, we achieve practical certainty that the theory is true, or 0.9940 confidence.

In the fifth component there is only one relevant observation; for the fourth, under the independence assumption and considering that we did not observe the fourth piece of evidence, we obtain a sensitivity of 0.1*0.4*0.1*0.2 = 0.0008, and a Type I error of 0.5*0.6*0.5*0.5 = 0.075. From a prior of 0.5, the posterior confidence is slashed to 0.0106; in other words, we are reasonably certain the theory component is not true.

3.3.3.2 *Combining probabilities to evaluate the impact of a forestry policy intervention*

In this section we reprise the Cattle Agreement theory component and convert our considerations on the evidence into numerical estimates; then we add our considerations on the observations' stochastic independence and obtain an overall confidence level for

the theory component, which reads "The Cattle Agreements pressured slaughterhouses not to buy cattle from deforesters and substantially reduced the deforesting behaviour of large holders when they came into effect".

We deemed that the first observation "all major slaughterhouses buying cattle from SFX signed the TAC Agreement" is expected with reasonable certainty if the theory is true (0.97, middle point of the range); but it's also quite likely if the theory is not true because they could have signed for political correctness or to avoid the legal consequences (0.80).

The second observation, "several statements from ranchers about how after the Cattle Embargo "there was no way to continue as before", the only policy they commented in this way", is mostly expected if the theory is true, although it's not certain ranchers would be so open in admitting the change: 80%. If the theory isn't true, perhaps they could still make that kind of statement in an attempt to appear politically correct but ranchers have been known not to be afraid of saying they break the law so the Type I error would be 0.4 (more unlikely than not if the theory isn't true).

The third observation, "slaughterhouses started to invest money in the regularization of their supply-chain (more than 1M $USD were invested by companies in that period), initially to implement an information system that was required by the Cattle Agreement" is somehow expected if the theory is true but not very highly, let's say cautious confidence (77%). If the theory isn't true, it doesn't seem very likely that they would make substantial investments to comply with the law; it's likely they experienced a shortage of supply and needed to invest to continue to be able to buy. We are around 20% confident that they would not do this.

The fourth observation is about the meeting happening one month after the other meeting, and the difference in discourse, the sudden openness towards alternatives to deforestation. Two important municipal meetings, one before and one after the intervention, seem

to signal the pivotal role of the Cattle Agreements. It's something we wouldn't necessarily expect to see if the theory were true (perhaps more confident than not, 60%), but if it's not true, such a big difference in only a month's time seems quite unlikely; it could have been a short-lived effect but we're only 20% confident it would still happen.

After arguing that the four observations are stochastically independent (Section 3.2) we can multiply the above values for the sensitivity and type I error and obtain a Sensitivity value of 0.97*0.80*0.77*0.60 = 0.358512 and a Type I error value of 0.80*0.40*0.20*0.20 = 0.0128 for the whole package. From a prior of 0.5, the posterior confidence is raised to 0.9655; or, in qualitative terms, we are reasonably certain that the Cattle Agreements were successful in substantially reducing the deforesting behaviour of medium and large holders.

We assessed the other theory components in a similar way, and we obtained the following levels of posterior confidence for each one of them:

1. BOI PIRATA (and subsequent similar monitoring and enforcement operations over two or three years) changed the behaviour of large and medium holders because it made deforesting riskier and substantially contributed to the drop: **0.9661 confidence (reasonable certainty).**
2. CREDIT RESTRICTIONS effectively provided a major incentive for large holders not to deforest, as credit was successfully restricted (**0.9193 confidence, or high confidence** that the theory component is true).
3. The Cattle Agreements pressured slaughterhouses not to buy cattle from deforesters and substantially reduced the deforesting behaviour of large holders when they came into effect (**0.9655 confidence, or reasonable certainty**).

4. The decrease of beef prices substantially contributed to the drop **(0.9804 confidence or reasonable certainty that the theory is FALSE**).
5. Temperature change substantially contributes to the drop (0.9804 confidence or reasonable certainty that the theory is FALSE).
6. Precipitation substantially contributes to the drop (0.9804 confidence or reasonable certainty that the theory is FALSE).

We can now estimate confidence levels of various combinations of components being true/false by multiplying their probabilities (of being true/false), so for example:

- We have achieved 0.9999 or practical certainty that at least one of the three interventions worked as intended.
    - 1 minus the product of the three inverse probabilities 0.0339*0.0807*0.0345, representing the probability that none of the three worked.
- We have achieved 0.8575 or high confidence that all three interventions worked as intended.
    - The product of the three probabilities 0.9661*0.9193*0.9655 that each of the interventions worked.
- We have achieved 0.9423 or high confidence that at least one of the three interventions worked as intended and none of the alternative explanations hold.
    - The product of the probability at the first bullet point (0.9999) and the probabilities that none of the three alternative explanations hold (0.9804 to the power of 3).

Table 19: Numerical probability assessments in the Uganda wildlife evaluation

	Observation of the Research report and related content	Acknowledgement of research (matching content) in conversations tracked by email or meeting minutes	Letter from PCLG to UWA with specific suggestion	Observation of an alternative research report with similar conclusions	Meeting minutes where PCLG member champions change in the UWA board (and does not mention alternative research as a significant influence)	Written evidence that the UWA board was considering a similar change before the lobbying but had not quite worked out the details nor was aware of how urgent it was
PCLG undertook research with partners and discovered that continued illegal activities in the park were rooted in dissatisfaction with distribution	S highly c (0.9) T practically certain (<0.01) (Doubly Decisive)	S – (0.5) T practically certain (<0.01) (Smoking Gun)	Irrelevant S – (0.5) T – (0.5) (Straw-in-the-Wind)	Irrelevant S – (0.5) T – (0.5) (Straw-in-the-Wind)	Irrelevant S – (0.5) T – (0.5) (Straw-in-the-Wind)	Irrelevant S – (0.5) T – (0.5) (Straw-in-the-Wind)

	Observation of the Research report and related content	Acknowledgement of research (matching content) in conversations tracked by email or meeting minutes	Letter from PCLG to UWA with specific suggestion	Observation of an alternative research report with similar conclusions	Meeting minutes where PCLG member champions change in the UWA board (and does not mention alternative research as a significant influence)	Written evidence that the UWA board was considering a similar change before the lobbying but had not quite worked out the details nor was aware of how urgent it was
of conservation resources						
PCLG used above findings to request to UWA that the shared fee be raised from $5 to $10	S reasonable certainty (0.95) T – (0.5) (Hoop test)	S MCTN pos (0.6) T MCTN neg (0.4) (Straw-in-the-Wind)	S – 0.5 T practical c <0.01 (Smoking Gun)	S cautious c 0.22 T – 0.5 (would weaken the theory if observed)	S MCTN 0.6 T practical c <0.01 (Smoking Gun)	Irrelevant? S – 0.5 T – 0.5 (Straw-in-the-Wind)

	Observation of the Research report and related content	Acknowledgement of research (matching content) in conversations tracked by email or meeting minutes	Letter from PCLG to UWA with specific suggestion	Observation of an alternative research report with similar conclusions	Meeting minutes where PCLG member champions change in the UWA board (and does not mention alternative research as a significant influence)	Written evidence that the UWA board was considering a similar change before the lobbying but had not quite worked out the details nor was aware of how urgent it was
PCLG contributed to the decision taken by UWA, by influencing the content and perhaps the timing	Irrelevant S – 0.5 T – 0.5 (Straw-in-the-Wind)	Strengthening straw? S cautious c P 0.77 T cautious c N 0.28	S – 0.5 T cautious / highly c 0.15 Almost a smoking gun	Not very relevant in itself (not observed) S MCTN N 0.4 T MCTN P 0.6 Straw-in-the-Wind	Did it cite PCLG research? If yes: S MCTN P – 0.6 T reasonable c N 0.05 Smoking Gun	S – 0.5 T – 0.5 (Straw-in-the-Wind)

	Observation of the Research report and related content	Acknowledgement of research (matching content) in conversations tracked by email or meeting minutes	Letter from PCLG to UWA with specific suggestion	Observation of an alternative research report with similar conclusions	Meeting minutes where PCLG member champions change in the UWA board (and does not mention alternative research as a significant influence)	Written evidence that the UWA board was considering a similar change before the lobbying but had not quite worked out the details nor was aware of how urgent it was
Research used was produced by someone else… this is tricky… it was clear they used their research, but what if PCLG had copied the research from somewhere else?	Irrelevant S – 0.5 T – 0.5 (Straw-in-the-Wind)	Absence is almost a Hoop test S high c N 0.1 T – 0.5	Not very relevant S MCTN N 0.4 T MCTN P 0.6 (Straw-in-the-Wind)	Almost Hoop test S high c P 0.9 T – 0.5 (We did not observe this, so the values are inverted into 0.1 and 0.5)	If PCLG hadn't copied the research from someone else: S 0.2 T 0.5 Observation strongly weakens the theory	S – 0.5 T – 0.5 (Straw-in-the-Wind)

	Observation of the Research report and related content	Acknowledgement of research (matching content) in conversations tracked by email or meeting minutes	Letter from PCLG to UWA with specific suggestion	Observation of an alternative research report with similar conclusions	Meeting minutes where PCLG member champions change in the UWA board (and does not mention alternative research as a significant influence)	Written evidence that the UWA board was considering a similar change before the lobbying but had not quite worked out the details nor was aware of how urgent it was
UWA was considering a similar decision before PCLG completed the research and started lobbying but they had not quite worked out the details and didn't know how	Irrelevant S – 0.5 T – 0.5 (Straw-in-the-Wind)	Irrelevant S – 0.5 T – 0.5 (Straw-in-the-Wind)	Irrelevant S – 0.5 T – 0.5 (Straw-in-the-Wind)	Irrelevant S – 0.5 T – 0.5 (Straw-in-the-Wind)	Irrelevant S – 0.5 T – 0.5 (Straw-in-the-Wind)	S – 0.5 T – high c N 0.1 Smoking Gun

	Observation of the Research report and related content	Acknowledgement of research (matching content) in conversations tracked by email or meeting minutes	Letter from PCLG to UWA with specific suggestion	Observation of an alternative research report with similar conclusions	Meeting minutes where PCLG member champions change in the UWA board (and does not mention alternative research as a significant influence)	Written evidence that the UWA board was considering a similar change before the lobbying but had not quite worked out the details nor was aware of how urgent it was
urgent it really was						

3.3.3.3 Handling degrees of stochastic inter-dependence

As we have briefly mentioned above, in this section we propose a third way of treating groups of inter-dependent observations: using coefficients to "discount" the package's probative value in a way that is proportional to the degree of inter-dependence. The degree of inter-dependence sits along a continuum, the extremes of which are full dependence and full inter-dependence. These extremes have corresponding numerical values: under full inter-dependence, each value is fully predictable from any other value; in other words, we only need to know one value (of a random package component) to know the values of every other component and of the whole package. The probative value added by each additional observation is null. Numerically, this is akin to a situation where we have independent pieces of additional evidence, but their sensitivities are the same as their Type I errors. When applying the above method, the values for the package will be the same as the value of the only informative observation. In other words, the additional observations do not add any information that alters our level of confidence that the theory is true or false.

In order to take degrees of interdependence between observations into account, we recreate a situation where the values of Sensitivity and Type I error for these observations "regress" towards each other with a speed that is proportional to their degree of inter-dependence. We calculate the distance between the two values and divide it by four (or in half twice); then we move the values closer to each other by a fourth each, so that the new distance between them is halved. We then apply the simple multiplication method that we use for independent observations to the new, lower-probative value estimates. The result is something in-between the high-strength/high-change estimates that we make under independence assumptions and the complete absence of change that occurs under the hypothesis of full inter-dependence.

In the second claim of our Uganda wildlife evaluation, (PCLG used above findings to request to UWA that the shared fee be raised from $5 to $10) there are two smoking guns and – while observing either increases our confidence considerably – we need to ask ourselves, what happens if we observe both? Is our confidence the same (as in, the second observation doesn't increase it further), or is it even higher? Our assessment above determined that we cannot argue for complete independence but at the same time the two events are quite distinct and do not necessarily imply each other. In other words, this is a typical situation where there is some degree of inter-dependence that is somewhere in the middle between the two extremes.

As such, this situation qualifies for the application of the above-mentioned procedure. The two observations have sensitivities of 0.5 and 0.6 and Type I Errors of 0.01 in both cases. If we used the simple multiplication under the assumption of full independence, the values for the package would be 0.3 and 0.0001, and a prior of 0.5 would be increased to 0.9997. If we observed the 0.6-0.01 piece of evidence after the first one, and wanted to apply the discount coefficients, it would become a weaker 0.45-0.16. Applying the simple multiplication to the first 0.5-0.01 test and the second one with modified values, we would have package S-T1E values of 0.225-0.0016, which return a posterior of 0.9929 from a prior of 0.5. Iterating the method twice and further discounting the second piece of evidence (to 0.38-0.23) would yield package values of 0.19-0.0023, for a slightly lower posterior of 0.9880, which is lower than the last value above, but still quite higher than the posterior we would get with just the first test (0.9804).

While this example is aimed at explaining and clarifying the procedure, the latter is most relevant when we are presented with ambiguous evidence: for example, with two observations presenting 0.30-0.40 and 0.60-0.20 (values of S and T1E); if they're independent, the package values are 0.18-0.08 and the posterior is 0.69, which indicates some confidence that the theory is true. However, if the second test has some degree of dependence to the

first, the situation is different: with one iteration, the "reduced probative-value" second test leads to package values of 0.15-0.12 and a posterior of 0.55, which is very close to middle ignorance; with two iterations nearing full inter-dependence, the package values become 0.135-0.14 and the posterior drops to 0.49, approximating perfect ignorance very closely.

We now address a situation where, for various reasons (lack of specialists, epistemic preference, etc), we might not be willing or able to estimate numerical values for the Bayes formula and can only draw on qualitative PT-like tests. We can still make sense of evidence packages as proposed in the next section.

3.3.4 Assembling observations in the absence of Bayes formula estimates

So far, section 3.3 has almost entirely dealt with the numerical values required for the updating. This section will cover the situations where we're dealing with multiple, arguably independent pieces of evidence but for various reasons we are unable to produce estimates for Sensitivity and Type I Error, neither qualitative nor qualitative. We can still combine the observations in a sensible way, provided we at least assign it Process Tracing categories like Smoking Gun (SG), Hoop test (HT), Doubly Decisive (DD), and Straw-in-the-Wind (SW).[35]

Building on pioneering work by (BEIS, 2018; Anderson, Ahmed, Befani, & Michaelis, 2020), we propose a way to establish the level of empirical support for a theory on the basis of multiple independent observations, of which we know the PT test category (see Table 20) but not the Bayes formula values. The system includes

[35] For our purposes here, observing a DD is equal to observing (passing) a SG, while not observing a DD is equal to failing (not observing) a HT. At the same time, observing or not observing a SW does not provide conclusive information and thus cannot replace neither a HT nor a SG.

five levels of support for a particular theory T: strong support, possible support, contradictory support, possible support for the opposite theory, and strong support for the opposite theory. We describe these levels in the rest of the section.

"Strong support for the opposite theory" (last row of Table 20) is associated to a situation where the theory is weakened by the failure of at least one Hoop test, and at the same time it is not strengthened because no Smoking Gun is passed (with at least one having been tried and having failed). In other words, when we have evidence of absence (the HT fail) without having evidence of presence (no SG); plus some absence of evidence (the failed SG). If we consider the matching text a SG and the broad content alignment a HT, this situation ("strong support for absence of influence") is equivalent to failing to observe broad content alignment, and also failing to observe matching text, with no other SGs passed. When interviewing key informants, if we might expect support for a theory the informants have a stake in supporting (which would be a HT); and lack of support from informants who have a stake in opposing it (which would be a SM if they showed indeed support). This situation would be akin to the supposedly friendly (to the theory) informants showing a lack of support for the theory, just like the supposedly hostile informants.

Conversely, we witness "strong support for the theory" (first row of Table 20) when at least one SG is passed, giving us "evidence of presence"; and no HTs fail, preventing us from finding evidence of absence. Since we're supposed to have tried some Hoops, at least one would have passed, hence we also have presence of evidence. In our examples, this is like when we observe matching text (a SG), and we also observe broad content alignment (a HT). Or when the supposedly hostile informants support the theory, just like the supposedly friendly ones.

The above two cases are somewhat symmetrical; two of the in-between scenarios are also symmetrical while the third is qualitatively different. In the "possible support for theory" case (second row of

Table 20), we observe some evidence (presence of evidence); but this is because some HTs are passed; not because any SGs are: the latter might not even be identified. The theory can be true: we haven't been able to weaken it but, unfortunately, we aren't quite able to strengthen it, either. This is like observing broad content alignment (HT) but failing to observe matching text (SG), or not even seeking to observe the latter. In our interviews, our supposedly friendly informants would support the theory, while the supposedly hostile ones would either not support it or would not provide any response (perhaps because they haven't been asked).

The case of "possible support for the opposite of the theory" (fourth row of Table 20) is similar: we can't neither strengthen nor weaken the theory (we don't have evidence of presence nor evidence of absence); but instead of passing HTs and having some weak evidence for the theory, HTs might not even be identified. SGs instead (at least one) are identified and fail, making us worry about absence of evidence. This is like seeking to observe matching text and failing, while not seeking to observe broad content alignment; and like the supposedly hostile informants not supporting the theory, while the supposedly friendly ones either supporting or not providing a response (whether asked or not).

Finally, the contradictory case (third row of Table 20) is when at least one SG passes and at least one HT fails: producing the logical contradiction of witnessing evidence of presence and evidence of absence at the same time. This is like observing matching text but failing to observe broad content alignment. The theory and its opposite are seemingly equally supported. In our other example, we would witness two surprises, in the supposedly hostile informants supporting the theory, and the supposedly friendly ones not supporting it.

Table 20: Assembling pieces of evidence when we lack Bayesian estimates

	Strengthening tests (Smoking Gun)		Weakening tests (Hoop test)	
	Passed: evidence of presence	Failed: absence of evidence	Passed: presence of evidence	Failed: evidence of absence
Strong Support for T	Y	-	Y	N
Possible Support for T	N	-	Y	N
Contradiction	Y	-	-	Y
Possible Support for ~T	N	Y	-	N
Strong Support for ~T	N	Y	-	Y

The more uncertain situations are the ones that would benefit the most from formal Bayesian Updating, especially the contradictory case. As an example, let's assume we have two observations, with Sensitivity values of 0.5 and 0.95, and Type I Error values of 0.05 and 0.5. If we observe the first (SG) but not the second[36] (HT), the two observations neutralise each other[37] and the posterior is identical to the prior (0.5). But with slightly different values that wouldn't change our qualitative assessment of "contradictory case", for example sensitivities of 0.3 and 0.97, and Type I errors of 0.1 and 0.4, making the first observation and not the second[38] would yield a posterior based on the combined package of 0.13. Qualitatively, the two pieces of evidence would still be a Smoking Gun and a

[36] or vice versa

[37] Taken one by one, the first raises the posterior to 0.91 and the unobserved second drops it to 0.09.

[38] making the first observation raises the prior to 0.75 and failing to make the second decreases it to 0.05.

Hoop test, but while 0.5 in the first case conveys the uncertainty that we would normally associate with contradictions, the 0.13 of the second case is a pretty convincing disconfirmation of the theory.

We now exemplify this assessment system using the Energy Policy evaluation. Note that the system used in the actual evaluation is different, but since it inspired this one it seems fitting to see how our proposed system would look like if it had been used in that evaluation.

3.3.4.1 *Combining observations without Bayesian estimates to assess the impact of energy policy*

Table 21 provides an example of how this method could be applied in the energy market evaluation. A similar but different method was applied in the actual evaluation (which inspired this one), so the evidence assessments below are fictitious and are presented for pure illustration purposes. The third and fourth column indicate how many tests of each kind are passed or failed against the corresponding mechanism in column two; with the fifth column drawing conclusions from the situation as suggested by Table 20.

Table 21: Qualitative assessment system for multiple observations in the energy policy evaluation

Broader theory	Mechanism	Strengthening Test	Weakening Test	Verdict
The second TA contributes to more and/or more competitive flexible capacity for the capacity market in 2018-19 and subsequent years	Our experience of participating in the second TA means the capacity market seems less risky	1 passed, 2 failed	2 passed, 0 failed	Strong Support for T
	In order to participate in the second TA, we invested in capacity or the ability to provide capacity which will make us better positioned to participate in the main CM	0 passed, 0/1 failed	2 passed, 0 failed	Possible Support for T
	(for new entrants) In order to participate in the second TA we have built a customer base and so now we want to continue with the CM	1 passed, 1 failed	0 passed, 1 failed	Contradiction

Broader theory	Mechanism	Strengthening Test	Weakening Test	Verdict
The second TA made no difference to the capacity available to the CM in 2018/19 and subsequent years and therefore is not additional	(for new entrants) We are a new entrant to flexibility in the CM but would have started participating with flexible capacity in the CM at the same level anyway, because of other changes, not the TA	0 passed, 2 failed	0/1 passed, 0 failed	Possible Support for ~T
	(for existing aggregators) We have always intended to participate in the CM and the TA did not help us to grow our flexibility business.	0 passed, 2 failed	2 failed	Strong Support for ~T

4. Concluding remarks

The goal of this report is to show that taking a diagnostic approach to Theory Based Evaluation (in particular using Bayesian Updating and the Confusion Matrix) improves the transparency of the evaluation process and ultimately the credibility and reliability of the findings. This is achieved by forging a close connection between theory and empirical data mediated by the Bayes formula and the Confusion Matrix (although as a stakeholder you don't need to understand the technicalities to engage with the method).

We hope the examples provided have helped bring to life the method's otherwise "dry" technicalities, although we chose to be very transparent on the technical aspects to appeal to a wide range of readers, including specialists, quantitative evaluators, qualitative evaluators, non-specialists, and commissioners (for which we drafted a dedicated annex).

Furthermore, we hope to have clarified how the method helps track the reasoning taking place in the researcher's mind when they make a claim on the basis of empirical evidence. The more detail we can have about this process, the easier it is to track, reproduce, and challenge. Our suggestion is to categorise evidence according to its strength and direction for a particular theory or theory component – that is, according to its power to strengthen the theory, weaken it, or both.

Stakeholders can be engaged and their implicit/unthought beliefs about the value of evidence can emerge through an elicitation process that can take various forms: it can be formal or informal, more or less structured, more or less individual vs. group based.

The closest well-known method to diagnostic theory-based evaluation is Process Tracing – and the reader might have wondered why we can't just apply Process Tracing and instead have to deal with formulas? While it has recently been recognised that PT tests might not be clear cut boxes and there can be fuzziness in those smoking

guns and hoop tests (Beach & Pedersen, 2019), the currently formulated, non-explicitly Bayesian version of Process Tracing falls short of providing tools to deal with the fuzziness. We believe and we hope to have convincingly shown that a method grounded on informal Bayesian probability can only deal with the uncertainty of its assessments by fully and formally embracing its probabilistic roots. The statement "this is a smoking gun" is not always clear in that it can simply mean that the evidence is confirmatory, or that it is also conclusive and refer to its probative value. Moreover, the construct "a strong smoking gun", while implying that the probative value is high, does not convey the magnitude very precisely: how high is that value? Once the method's learning curve plateaus and we start to handle numerical formalisation more easily, it should be self-evident how much easier it is to communicate and challenge explicit assessments on the value of empirical evidence.

Reliability (or robustness) is linked to transparency: once we are able to replicate the process, how likely is it that we'll obtain the same findings? We might find ourselves in a situation where stakeholders disagree on interpreting the same empirical data; but instead of gravitating towards different findings, the method allows stakeholders to make the reasoning behind their assessment explicit, increasing the chances of mutual understanding and agreement. If there is agreement on the reasoning but disagreement on the probabilities, numerical ranges associated with qualitative levels of confidence can be used. If there is still disagreement, a formal, workshop-based process inspired by the SHELF method for elicitation of expert probability judgements can be adapted for evaluation purposes. This is particularly needed when evidence is mixed, confusing or contradictory, and the risk of bias skewing judgement is the highest.

Obtaining reliable findings that won't change much over time or over replications also means testing how sensitive our confidence is to possible new discoveries and answering the question "how much evidence is enough? When can we stop seeking and collecting?".

More specifically, this question[39] can become "how strong a piece of evidence from the opposite direction do we need in order to substantially change our assessment"? The "cast the net widely" suggestion (Bennett, Checkel, & (eds), 2014), encouraging us to consider the widest possible range of theories as well as observations, rather than cherry picking our favourite ones, can now take the form of a structured matrix, where each cell is associated with a theory and with an observation with a given strengthening and/or weakening power. Seeing only strengthening pieces in the package should create suspicions and make quality assurers enquire about what is possibly missing. Similarly, low estimates of Sensitivity and Type I error should raise suspicions of Analysis and Memory Confirmation Biases, that make us overestimate the observations' strengthening powers and underestimate its weakening abilities.

Finally, the following factors make findings credible: a) transparency, or the reader being able to follow the process whereby findings have been obtained and potentially replicating it; b) believing that the process is correct and leads to as unbiased as possible estimates; and c) high levels of confidence in the statements that make up the findings' content. We have covered the first two; and argued that diagnostic TBE allows for an easier identification of evidence strength and updating direction; we have also seen in the report how it protects against conservatism bias and confirmation bias. But ultimately, the benefits of declaring an explicit level of confidence and making the process to obtain it transparent are that high levels of confidence are more credible than they would be if this process were less transparent. In addition, the benefits of formalisation extend to evidence packages: the method also reduces uncertainty in the trickiest situations when we have mixed or contradictory evidence, with some observations strengthening and others

[39] For a more comprehensive discussion of this question that is not limited to probability estimates, see Bennett and Checkel 2014. See also Beach and Pedersen 2019 (page 230).

weakening the theory; or when multiple weak pieces of seemingly inconclusive evidence seem to lean towards a similar direction (mostly strengthening or mostly weakening).

In the first case, we might know that one observation weakens, and another strengthens, but we usually cannot tell if the former weakens more strongly than the latter strengthens. This becomes a lot more complicated with more than two observations. Let's consider the case of Theory One in Table 12: O_1 increases our confidence (0.90), but O_2 (0.38), and most of all O_n (0.09), decrease it. Without Bayesian Updating our verdict would simply be "mixed evidence" or "contradictory evidence"; but most of the times – because of conservatism – it is scientifically demonstrated that the evidence is more informative than we think, and the actual posterior after observing the package is 0.36, which is lower than we were probably expecting (Table 16). If we use the currently available tools to automatically check the direction and strength of the evidence package, we often discover that it is more conclusive than we expected (unless our beliefs were skewed by confirmation bias). In other words, humans' idea of uncertainty or mixed evidence tends to be anchored more closely to the 0.5 middle point than it should be (section 1.2.1).

As for the second situation where observations are all of the same kind (say, all strengthening), but are individually weak, it's difficult to accurately assess how strong the package is unless we use the formula (and, for independent observations, the automatic combination function of the tool). Due to conservatism, humans tend to underestimate the power of evidence to change their priors, and they will usually be surprised at how strongly conclusive a package comprised of 3 or 4 weak straws of the same kind can be. And even those humans less subject to this bias won't know how many straws of the same kind they need (say, similar responses in independent interviews) to reach a certain level of overall confidence, unless they use the formula. In general, it's quite difficult to predict how much

adding one piece of evidence of a certain kind and strength, affects the kind and strength of an existing package. It can easily be seen by playing with the updating tool (Befani, 2017).

Finally, the method can protect us from confirmation bias because it draws our attention away from our preferred theory and forces us to think, not only of alternatives, but of the opposite of the theory: we are forced to consider a situation where the theory is not true several times during the process and reflect on the (empirical) implications of such ontological reality. Because the Type I error estimates are required and require us to immerse ourselves in a world that is the opposite state of our preferred reality.

Against the above-mentioned benefits, the only substantial cost or challenge we see is the learning curve; it is not an immediately intuitive method because most people even with an advanced university degree have likely not covered the concepts illustrated in the Confusion Matrix or the Bayes formula in their studies; and while every human daily updates possibly hundreds of beliefs, they never do it formally (which is why our beliefs are systematically biased). But the learning curve can be navigated with dissemination, training, and exchange of lessons learned from testing and application; and immense progress has already been achieved in the last 5 or 10 years. As with every other process of learning or cultural change, it might take some time but it's possible to get there. We believe that some initial patience and an intellectual investment is well worth the future benefits that mastering this method will yield.

References

AA.VV. (2017). Theory-based impact evaluation. London: International Institute for Environment and Development. Retrieved from http://pubs.iied.org/pdfs/17404IIED.pdf

Anderson, M., Ahmed, T., Befani, B., & Michaelis, C. (2020). Testing the strength of impact evidence - applying contribution tracing within a realist evaluation of Transitional Arrangements for demand-side response. London: Energy Evaluation Europe Conference 29 June-1st July 2020.

Andrew, S., & Halcomb, E. J. (2009). Mixed Methods Research for Nursing and Health Sciences. London: Wiley-Blackwell.

Badham, J. (2019). Agent-Based Modelling for the Self Learner. Retrieved from http://research.criticalconnections.com.au/ABMBook/

Badham, J., & Gilbert, N. (2015). TELL ME Design: Protective Behaviour During an Epidemic. Guildford: CRESS Working Paper. Retrieved from http://cress.soc.surrey.ac.uk/web/publications/working-papers/tell-me-design-protective-behaviour-during-epidemic

Beach, D. (2018). Achieving Methodological Alignment When Combining QCA and Process tracing in Practice. Sociological Methods & Research, 47(1), 64-99. doi:10.1177/0049124117701475

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

Beach, D., & Pedersen, R. (2019). Process Tracing Methods: Foundations and Guidelines, 2nd edition. Ann Arbor: University of Michigan Press.

Befani, B. (2016). Pathways to change: Evaluating development interventions with Qualitative Comparative Analysis (QCA). Stockholm: EBA.

Befani, B. (2017). Bayes Formula Confidence Updater. CECAN. Retrieved from https://www.cecan.ac.uk/wp-content/uploads/2017/03/bayes_formula_confidence_updater.xlsx

Befani, B. (2020). Choosing Appropriate Evaluation Methods – A Tool for Assessment and Selection (Version Two). Guildford: CECAN. Retrieved from https://www.cecan.ac.uk/news/choosing-appropriate-evaluation-methods-a-tool-for-assessment-and-selection-version-two/

Befani, B. (2020). Diagnostic evaluation and Bayesian Updating: Practical solutions to common problems. Evaluation, 26(4), 499-515. doi:10.1177/1356389020958213

Befani, B. (2020). Quality of Quality: a diagnostic approach to qualitative evaluation. Evaluation, 26(3), 333-349. doi:10.1177/1356389019898223

Befani, B., & D'Errico, S. (2020). Letting evidence speak for itself: Measuring confidence in mechanisms. New Directions for Evaluation, 2020(167), 27– 43.

Befani, B., & D'Errico, S. (2020). Letting evidence speak for itself: Measuring confidence in mechanisms. New Directions for Evaluation, 2020(167), 27– 43.

Befani, B., & Mayne, J. (2014). Process Tracing and Contribution Analysis: A Combined Approach to Generative Causal Inference for Impact Evaluation. (B. Befani, C. Barnett, & E. Stern, Eds.) IDS Bulletin, 45(6), 17-36.

Befani, B., & Stedman-Bryce, G. (2017). Process Tracing and Bayesian updating for impact evaluation. Evaluation. Retrieved from http://evi.sagepub.com/content/early/2016/06/24/1356389016654584.abstract

Befani, B., D'Errico, S., Booker, F., & Giuliani, A. (2016). Clearing the fog: new tools for improving the credibility of impact claims. IIED Briefing. London: International Institute for Environment and Development. Retrieved from http://pubs.iied.org/17359IIED.html

Befani, B., Elsenbroich, C., & Badham, J. (2021). Diagnostic evaluation with simulated probabilities. Evaluation, 27(1), 102-115. doi:10.1177/1356389020980476

Befani, B., Ramalingam, B., & Stern, E. (2015). Introduction – Towards Systemic Approaches to Evaluation and Impact. (B. Befani, B. Ramalingam, & E. Stern, Eds.) IDS Bulletin, 46(1), 1-6.

Befani, B., Rees, C., Varga, L., & Hills, D. (2016). Testing Contribution Claims with Bayesian Updating. EPPN 2. CECAN. Retrieved from https://www.cecan.ac.uk/sites/default/files/2018-01/BARBARA%20v2.5.pdf

BEIS. (2018). Evaluation of the Transitional Arrangements for Demand Side Response. London: BEIS.

Bennett, A. (2008). Process Tracing: a Bayesian Perspective. In J. Box-Steffensmeier, H. Brady, & D. Collier, The Oxford Handbook of Political Methodology. OUP.

Bennett, A. (2010). Process Tracing and Causal Inference. In H. Brady, & D. Collier, Rethinking Social Inquiry. Rowman and Littlefield.

Bennett, A. (2014). Disciplining our conjectures. In A. Bennett, & J. Checkel, Process Tracing: From Metaphor to Analytic Tool. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139858472.015

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

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

Bhaskar, R. (2009). Scientific Realism and Human Emancipation. Routledge.

Brandao, F., & Befani, B. (2021). The effectiveness of six initiatives to halt deforestation: a process-tracing approach in an Amazonian frontier. Working Paper.

Bryman, A. (2012). Social Research Methods (4th ed.). Oxford: Oxford University Press.

Cartwright, N. (2020). Using middle-level theory to improve programme and evaluation design. Oxford: CEDIL Methods Brief. Retrieved from https://cedilprogramme.org/blog/using-mid-level-theory-to-improve-programme-and-evaluation-design/

Cartwright, N., & Hardie, J. (2012). Evidence-Based Policy: A Practical Guide to Doing It Better. Oxford University Press.

Cartwright, N., Charlton, L., Juden, M., Munslow, T., & Williams, R. (2020). Making predictions of programme success more reliable. Oxford: CEDIL Methods Working Paper.

Chen, H.-T. (1990). Theory-Driven Evaluations. Sage.

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the Evidence Hierarchy. Topoi, 33, 339–360. Retrieved from https://doi.org/10.1007/s11245-013-9220-9

Collier, D. (2011). Understanding Process Tracing. Political Science and Politics, 44(4), 823-830.

Cooke, R. M. (1991). Experts in Uncertainty: Opinion and Subjective Probability in Science. New York, Oxford: Oxford University Press.

D'Errico, S., Befani, B., Booker, F., & Giuliani, A. (2017). Influencing policy change in Uganda: An impact evaluation of the Uganda Poverty and Conservation Learning Group's work. London: IIED. Retrieved from https://pubs.iied.org/g04157

DFID. (2014). How to note: Assessing the strength of evidence. London: Department for International Development. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/291982/HTN-strength-evidence-march2014.pdf

Edwards, W. (1982). Conservatism in Human Information Processing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment Under Uncertainty: Heuristics and Biases (pp. 359-369). Cambridge University Press.

EFSA, (. (2014). Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. EFSA Journal, 12(6).

Elster, J. (1998). A plea for mechanisms. In P. Hedstrøm, R. Swedberg, & (eds), Social Mechanisms: An Analytical Approach to Social Theory (pp. 45-73). Cambridge: Cambridge University Press.

Fairfield, T., & Charman, A. (2017). Explicit Bayesian analysis for process tracing: guidelines, opportunities, and caveats. Political Analysis, 25(3), 363-380.

Farrington, D. P., Gottfredson, D. C., Sherman, L. W., & Welsh, B. C. (2002). Maryland Scientific Methods Scale. In D. P. Lawrence W. Sherman (Ed.), Evidence-Based Crime Prevention (pp. 13-21). New York: Routledge.

Friedman, R. (1986). A Close Look at Probative Value. Boston University Law Review, 66, 733-59.

Gertler, P., Martinez, S., Premand, P., Rawlings, L., & Vermeerch, C. (2011). Impact Evaluation in Practice. Washington, D.C.: The World Bank.

Gosling, J. P. (2014). Methods for eliciting expert opinion to inform health technology assessment.

Hummelbrunner, R. (2015). Learning, Systems Concepts and Values in Evaluation: Proposal for an Exploratory Framework to Improve Coherence. (B. Befani, B. Ramalingam, & E. Stern, Eds.) IDS Bulletin, 46(1), 17-29.

Humphreys, M., & Jacobs, A. (2015). Mixing Methods: A Bayesian Approach. American Political Science Review, 109(4), 653-673. doi:10.1017/S0003055415000453

Kahneman, D. (2012). Thinking, Fast and Slow. Penguin.

Kaye, D. (1986). Quantifying Probative Value. Boston University Law Review, 66, 761-766.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Beverly Hills, CA: Sage.

LTS International and the Centre for Development Management. (2017). Enhancing Community Resilience Programme Final Evaluation. London: DFID. Retrieved from http://www.careevaluations.org/wp-content/uploads/The-Enhancing-Community-Resilience-Program.pdf

Mayne, J. (1999). Addressing Attribution Through Contribution Analysis: Using Performance Measures Sensibly. Discussion Paper. Office of the Auditor General of Canada.

Mayne, J. (2008). Contribution Analysis: an approach to exploring cause and effect. ILAC Brief 16. Institutional Learning and Change (ILAC) Initiative (CGIAR).

Oakley, J. E., & O' Hagan, A. (2016). SHELF: the Sheffield Elicitation Framework (version 3.0). Sheffield: School of Mathematics and Statistics, University of Sheffield, UK. Retrieved from (http://tonyohagan.co.uk/shelf

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J., Garthwaite, P. H., Jenkinson, D. J., . . . Rakow, T. (2006). Uncertain Judgements: Eliciting Experts' Probabilities. Wiley.

Patton, M. Q. (2008). Advocacy Impact Evaluation. Journal of Multidisciplinary Evaluation, 5(9).

Pawson, R., & Tilley, N. (1997). Realistic Evaluation. Sage.

Plous, S. (1993). The Psychology of Judgment and Decision Making. McGraw-Hill.

Sage, B., Meaux, A., Osofisan, W., Traynor, M., & Jove, T. (2017). Urban Context Analysis Toolkit. London: IIED. Retrieved from https://pubs.iied.org/10819iied

Sale, J., & Brazil, K. A. (2004). A strategy to identify critical appraisal criteria for primary mixed-method studies. Quality & Quantity, 38(4), 351-365.

Savedoff, W. D., Levine, R., & Birdsall, N. (2006). When will we ever learn? Improving Lives Through Impact Evaluation, Report of the Evaluation Gap Working Group. Washington, D.C.: Center for Global Development (CGD).

Schmitt, J. (2020). Editor's Note. New Directions for Evaluation, 2020(167), 7-10. Retrieved from https://doi.org/10.1002/ev.20425

Schneider, C. Q., & Rohlfing, I. (2013). Combining QCA and Process Tracing in Set-Theoretic Multi-Method Research. Sociological Methods & Research, 42(4), 559-597. doi:10.1177/0049124113481341

Stern, E. (2015). Impact Evaluation: a Guide for Commissioners and Managers. BOND UK.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). Broadening the Range of Designs and Methods for Impact Evaluations. DFID Working Paper 38. UK Department for International Development.

Treasury, H. (2011). The Magenta Book: Guidance for Evaluation. London: UK Government.

Tsang, E. W. (2014). Generalizing from Research Findings: The Merits of Case Studies. International Journal of Management Reviews, 16(4), 369-383. Retrieved from https://doi.org/10.1111/ijmr.12024

Tversky, A., & Kahneman , D. (1974). Judgment under Uncertainty: Heuristics and Biases. Science New Series, 185(4157), 1124-1131.

Vaessen, J., Lemire, S., & Befani, B. (2020). Evaluation of International Development Interventions: An Overview of Approaches and Methods. Washington, DC: The World Bank. Retrieved from https://ieg.worldbankgroup.org/sites/default/files/Data/Evaluation/files/MethodsSourceBook.pdf

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

Wauters, B., & Beach, D. (2018). Process tracing and congruence analysis to support theory-based impact evaluation. Evaluation, 24(3), 284-305.

Weiss, C. (1974). Evaluation. Pearson.

Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. New Directions for Evaluation, 41-55. Retrieved from https://doi.org/10.1002/ev.1086

Weiss, C. H., & Connell, J. P. (1995). Nothing as Practical as Good Theory: Exploring Theory-Based Evaluation for Comprehensive Community Initiatives for Children and Families. In New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts (pp. 65-92). The Aspen Institute.

White, H. (2009). Theory-Based Impact Evaluation: Principles and Practice. International Initiative for Impact Evaluation (3ie). Retrieved from http://www.managingforimpact.org/resource/3ie-working-paper-theory-based-impact-evaluation-principles-and-practice

Williams, B. (2015). Prosaic or Profound? The Adoption of Systems Ideas by Impact Evaluation. (B. Befani, B. Ramalingam, & E. Stern, Eds.) IDS Bulletin, 46(1), 7-16.

Williams, B., & Hummelbrunner, R. (2010). Systems Concepts in Action: a practicioner's toolkit. Stanford University Press.

Appendix: What a commissioner should know

As a commissioner, you might have been pressured to increase the quality of the evaluations you're responsible for; you might have been asked to commission RCTs or quasi-experimental evaluations because they allegedly offer more rigour and come with scientific standards that confer credibility to the findings. At the same time, you might have been struggling with commissioning experimental or quasi-experimental evaluations because the interventions you wanted to evaluate and the conditions in the evaluation process struggled to meet the requirements of these quantitative methods. Besides, you might be unsatisfied with the kind of information experimental or quasi-experimental evaluations provide, because they're focused on whether and to what extent a result has been achieved but they do a poor job of investigating the reasons why interventions work or not; while you acknowledge the duties of accountability, you're also interested in learning.

If the above paragraph somewhat describes your experience, this method might help you because it's tailored to learning and explaining why programmes worked or not but at the same time it retains most advantages of quantitative and experimental methods. The research process is made more transparent and thus subject to scrutiny and challenge, and as such it might be also suitable for more controversial or sensitive evaluations or evaluations that are needed to support or discourage major policy decisions.

In terms of requirements, the method does not necessarily demand a higher allocation of resources than more traditional case studies; and like many other TBE approaches (for example Contribution Analysis or Process Tracing), it is recommended for use in single case studies or in a handful of cases at most. Compared to typical TBE approaches, the difference is that Bayesian or diagnostic TBE requires relatively highly trained consultants who have a good

understanding of probability and the Bayes formula. This knowledge is not required of stakeholders but must be held by the principal methodologist involved in the evaluation design, and ideally by the Quality Assurance reviewer. The principal investigator or at least the methodologist also need to have facilitation skills to obtain the needed information from stakeholders and communication skills to convey the main features of the method to everyone involved in the evaluation.

This last point is quite important because the nature of the evidence that people tend to seek when applying the method is often confidential, for example sensitive documents, meeting minutes, or emails; and stakeholders do not automatically hand out this kind of documentation to any evaluator: a delicate process of trust building is sometimes required. On the bright side, in the author's experience, this process is mostly successful although it might take a bit of time with some stakeholders. Communication of the evaluation's purposes is very important to this regard; it's easier if at least some stakeholders are particularly engaged and motivated, if they see this part of the evaluation as a learning process from which they will benefit; and are informed with high detail on how their sensitive information will be handled.

On the technical side, the method comes with a relatively steep learning curve, at least at the very beginning; and – at the time of writing – it's not necessarily easy to find these skills on the market. But as with any other innovative method, the situation is bound to improve over time, as new knowledge is disseminated, as consultants are trained, and as application experience builds up and allows the exchange of lessons learned within the community.

In terms of usefulness of the results for decision making and which "actions" are taken at the end of a diagnostic evaluation, the situation is not categorically different from any other theory-based evaluation: the advantages of learning and explaining outcomes are preserved; you will be able to understand how and why an intervention worked and perhaps to predict where and under which conditions it will

work in the future. The added value is that those theories and explanations will be more credible and reliable because the thinking process will have been more transparent and thus more capable of being challenged and scrutinised. It will be safer to act on those results compared to traditional TBE because they will be more rigorous and robust.

Finally, it's important to know that the estimation of probabilities or qualitative confidence levels, whether they're fully subjective or backed by empirical or simulation evidence, can be conducted in many different ways. It can draw on the evaluator's own judgement and (explicitly shared) reasoning; or be a common assessment involving a wider evaluation team; or finally be a more formalised undertaking where a group of stakeholders and / or experts are called to estimate probability distributions during specifically organised workshops.

The chosen strategy will depend, again, on the degree of controversy or sensitivity surrounding the topic and the specific evaluation results. While the most important feature of the method is the process that allows the design and analysis steps to be traced, scrutinised, and challenged, and the embedded bias in it to be detected, ultimately the findings' credibility rests on the robustness of the probability or confidence level estimates that are input (with various degrees of processing) into the Bayes formula. It is thus fundamental that this process can ultimately be trusted by an as wide audience as possible and that it appears convincing in the eyes of the readers, stakeholders, and users of the evaluation.

Previous EBA reports

2021:2, *Målbild och mekanism: Vad säger utvärderingar om svenska biståndsinsatsers måluppfyllelse?* Markus Burman

2021:1, *Data Science Methods in Development Evaluation: Exploring the Potential*, Gustav Engström and Jonas Nóren

2020:07 *Effects of Swedish and International Democracy Aid*, Miguel Niño-Zarazúa, Rachel M. Gisselquist, Ana Horigoshi, Melissa Samarin and Kunal Sen

2020:06 *Sextortion: Corruption and Gender-Based Violence,* Åsa Eldén, Dolores Calvo, Elin Bjarnegård, Silje Lundgren and Sofia Jonsson

2020:05 *In Proper Organization we Trust – Trust in Interorganizational Aid relations,* Susanna Alexius and Janet Vähämäki

2020:04 *Institution Building in Practice: An Evaluation of Swedish Central Authorities' Reform Cooperation in the Western Balkans*, Richard Allen, Giorgio Ferrari, Krenar Loshi, Númi Östlund and Dejana Razić Ilić

2020:03 *Biståndets förvaltningskostnader För stora? Eller kanske för små*?, Daniel Tarschys

2020:02 *Evaluation of the Swedish Climate Change Initiative, 2009–2012,* Jane Burt, John Colvin, Mehjabeen Abidi Habib, Miriam Kugele, Mutizwa Mukute, Jessica Wilson

2020:01 *Mobilising Private Development Finance: Implications for Overall Aid Allocations*, Polly Meeks, Matthew Gouett and Samantha Attridge

2019:09 *Democracy in African Governance: Seeing and Doing it Differently,* Göran Hydén with assistance from Maria Buch Kristensen

2019:08 *Fishing Aid – Mapping and Synthesising Evidence in Support of SDG 14 Fisheries Targets,* Gonçalo Carneiro, Raphaëlle Bisiaux, Mary Frances Davidson, Tumi Tómasson with Jonas Bjärnstedt

2019:07 *Applying a Masculinities Lens to the Gendered Impacts of Social Safety Nets*, Meagan Dooley, Abby Fried, Ruti Levtov, Kate Doyle, Jeni Klugman and Gary Barker

2019:06 *Joint Nordic Organisational Assessment of the Nordic Development Fund (NDF)*, Stephen Spratt, Eilís Lawlor, Kris Prasada Rao and Mira Berger

2019:05 *Impact of Civil Society Anti-Discrimination Initiatives: A Rapid Review*, Rachel Marcus, Dhruva Mathur and Andrew Shepherd

2019:August *Migration and Development: the Role for Development Aid*, Robert E.B. Lucas (joint with the Migration Studies Delegation, Delmi, published as Delmi Research overview 2019:5)

2019:04 *Building on a Foundation Stone: the Long-Term Impacts of a Local Infrastructure and Governance Program in Cambodia*, Ariel BenYishay, Brad Parks, Rachel Trichler, Christian Baehr, Daniel Aboagye and Punwath Prum

2019:03 *Supporting State Building for Democratisation? A Study of 20 years of Swedish Democracy Aid to Cambodia,* Henny Andersen, Karl-Anders Larsson och Joakim Öjendal

2019:02 *Fit for Fragility? An Exploration of Risk Stakeholders and Systems Inside Sida*, Nilima Gulrajani and Linnea Mills

2019:01 *Skandaler, opinioner och anseende: Biståndet i ett medialiserat samhälle*, Maria Grafström och Karolina Windell

2018:10 *Nation Building in a Fractured Country: An Evaluation of Swedish Cooperation in Economic Development with Bosnia and Herzegovina 1995–2018*, Claes Lindahl, Julie Lindahl, Mikael Söderbäck and Tamara Ivankovic

2018:09 *Underfunded Appeals: Understanding the Consequences, Improving the System,* Sophia Swithern

2018:08 *Seeking Balanced Ownership in Changing Development Cooperation Relationships*, Nils Keizer, Stephan Klingebiel, Charlotte Örnemark, Fabian Scholtes

2018:07 *Putting Priority into Practice: Sida's Implementation of its Plan for Gender Integration*, Elin Bjarnegård, Fredrik Uggla

2018:06 *Swedish Aid in the Era of Shrinking Space – the Case of Turkey*, Åsa Eldén, Paul T. Levin

2018:05 *Who Makes the Decision on Swedish Aid Funding? An Overview*, Expertgruppen för Biståndsanalys

2018:04 *Budget Support, Poverty and Corruption: A Review of the Evidence*, Geske Dijkstra

2018:03 *How predictable is Swedish aid? A study of exchange rate volatility,* Númi Östlund

2018:02 *Building Bridges Between International Humanitarian and Development Responses to Forced Migration*, Alexander Kocks, Ruben Wedel, Hanne Roggemann, Helge Roxin (joint with the German Institute for Development Evaluation, DEval)

2018:01 *DFIs and Development Impact: an evaluation of Swedfund,* Stephen Spratt, Peter O'Flynn, Justin Flynn

2017:12 *Livslängd och livskraft: Vad säger utvärderingar om svenska biståndsinsatsers hållbarhet?* Expertgruppen för biståndsanalys

2017:11 *Sweden's Financing of UN Funds and Programmes: Analyzing the Past, Looking to the Future*, Stephen Browne, Nina Connelly, Thomas G. Weiss

2017:10 *Seven Steps to Evidence-Based Anticorruption: A Roadmap*, Alina Mungiu-Pippidi

2017:09 *Geospatial analysis of aid: A new approach to aid evaluation*, Ann-Sofie Isaksson

2017:08 *Research capacity in the new global development agenda*, Måns Fellesson

2017:07 *Research Aid Revisited – a historically grounded analysis of future prospects and policy options*, David Nilsson, Sverker Sörlin

2017:06 *Confronting the Contradiction – An exploration into the dual purpose of accountability and learning in aid evaluation*, Hilde Reinertsen, Kristian Bjørkdahl, Desmond McNeill

2017:05 *Local peacebuilding – challenges and opportunities*, Joakim Öjendal, Hanna Leonardsson, Martin Lundqvist

2017:04 *Enprocentmålet – en kritisk essä*, Lars Anell

2017:03 *Animal health in development – it's role for poverty reduction and human welfare,* Jonathan Rushton, Arvid Uggla, Ulf Magnusson

2017:02 *Do Anti-Discrimination Measures Reduce Poverty Among Marginalised Social Groups?* Rachel Marcus, Anna Mdee, Ella Page

2017:01 *Making Waves: Implications of the irregular migration and refugee situation on Official Development Assistance spending and practices in Europe*, Anna Knoll, Andrew Sherriff

2016:11 *Revitalising the policy for global development,* Per Molander

2016:10 *Swedish Development Cooperation with Tanzania – Has It Helped the Poor?* Mark McGillivray, David Carpenter, Oliver Morrissey, Julie Thaarup

2016:09 *Exploring Donorship – Internal Factors in Swedish Aid to Uganda,* Stein-Erik Kruse

2016:08, *Sustaining a development policy: results and responsibility for the Swedish policy for global development* Måns Fellesson, Lisa Román

2016:07 *Towards an Alternative Development Management Paradigm?* Cathy Shutt

2016:06 *Vem beslutar om svenska biståndsmedel? En översikt,* Expertgruppen för biståndsanalys

2016:05 *Pathways to change: Evaluating development interventions with Qualitative Comparative Analysis (QCA),* Barbara Befani

2016:04 *Swedish responsibility and the United Nations Sustainable Development Goals,* Magdalena Bexell, Kristina Jönsson

2016:03 *Capturing complexity and context: evaluating aid to education,* Joel Samoff, Jane Leer, Michelle Reddy

2016:02 *Education in developing countries what policies and programmes affect learning and time in school?* Amy Damon, Paul Glewwe, Suzanne Wisniewski, Bixuan Sun

2016:01 *Support to regional cooperation and integration in Africa – what works and why?* Fredrik Söderbaum, Therese Brolin

2015:09 *In search of double dividends from climate change interventions evidence from forest conservation and household energy transitions,* G. Köhlin, S.K. Pattanayak, E. Sills, E. Mattsson, M. Ostwald, A. Salas, D. Ternald

2015:08 *Business and human rights in development cooperation – has Sweden incorporated the UN guiding principles?* Rasmus Klocker Larsen, Sandra Atler

2015:07 *Making development work: the quality of government approach,* Bo Rothstein and Marcus Tannenberg

2015:06 *Now open for business: joint development initiatives between the private and public sectors in development cooperation,* Sara Johansson de Silva, Ari Kokko and Hanna Norberg

2015:05 *Has Sweden injected realism into public financial management reforms in partner countries?* Matt Andrews

2015:04 *Youth, entrepreneurship and development*, Kjetil Bjorvatn

2015:03 *Concentration difficulties? An analysis of Swedish aid proliferation*, Rune Jansen Hagen

2015:02 *Utvärdering av svenskt bistånd – en kartläggning*, Expertgruppen för biståndsanalys

2015:01 *Rethinking Civil Society and Support for Democracy*, Richard Youngs

2014:05 *Svenskt statligt internationellt bistånd i Sverige: en översikt*, Expertgruppen för biståndsanalys

2014:04 *The African Development Bank: ready to face the challenges of a changing Africa?* Christopher Humphrey

2014:03 *International party assistance – what do we know about the effects?* Lars Svåsand

2014:02 *Sweden´s development assistance for health – policy options to support the global health 2035 goals,* Gavin Yamey, Helen Saxenian, Robert Hecht, Jesper Sundewall and Dean Jamison

2014:01 *Randomized controlled trials: strengths, weaknesses and policy relevance,* Anders Olofsgård

Denna rapport presenterar en ny metod för teoribaserad utvärdering som kombinerar styrkor från både kvalitativa och kvantitativa metoder. Rapporten innehåller en utförlig diskussion om metodens kunskapsteoretiska principer och hur de förhåller sig till utvärderingspraktiken samt ett ingående avsnitt med praktiska tillämpningar riktat till utvärderare, forskare och konsulter som vill tillämpa metoden i praktiken. Den innehåller också ett kapitel om vad beställare av utvärderare bör känna till om metoden.

This report presents an innovative methodology to conduct theory-based evaluations. It retains several advantages of both qualitative and quantitative methods. This report contains both an in-depth theoretical discussion of the epistemological tenets and a section with practical applications in policy evaluations that is likely to appeal to evaluators, researchers, and consultants who want to apply the method in practice. It also contains a section on what a commissioner of evaluations should know.

EBA
www.eba.se

Expertgruppen för biståndsanalys (EBA) är en statlig kommitté som oberoende analyserar och utvärderar svenskt internationellt bistånd.

The Expert Group for Aid Studies (EBA) is a government committee with a mandate to independently analyse and evaluate Swedish international development aid.