POLICY EVALUATION:

METHODS AND APPROACHES

Edited by

ANNE REVILLARD

Policy Evaluation: Methods and Approaches

Edited by Anne Revillard

QUEBEC CITY: ÉDITIONS SCIENCE ET BIEN COMMUN

Title: Policy Evaluation: Methods and Approaches

Edited by Anne Revillard

Cover design: Kate McDonnell

Éditions science et bien commun
http://editionscienceetbiencommun.org
3-855 avenue Moncton
Québec (Québec) G1S 2Y4

Contact: info@editionscienceetbiencommun.org

Contents

Partie II. QUALITATIVE METHODS

Partie III. MIXED METHODS AND CROSS-CUTTING APPROACHES

This book is dedicated to the memory of Professor Pierre Pluye, who contributed so much to the reflection on mixed methods.

Introduction

This publication follows a first book (*Évaluation : fondements, controverses, perspectives*) published at the end of 2021 by Editions Science et Bien Commun (ESBC) with the support of the Laboratory for interdisciplinary evaluation of public policies (LIEPP), compiling a series of excerpts from fundamental and contemporary texts in evaluation (Delahais et al. 2021). Although part of this book is dedicated to the diversity of paradigmatic approaches, we chose not to go into a detailed presentation of methods on the grounds that this would at least merit a book of its own. This is the purpose of this volume. This publication is part of LIEPP's collective project in two ways: through the articulation between research and evaluation, and through the dialogue between quantitative and qualitative methods.

Methods between research and evaluation

Most definitions of programme evaluation[1] articulate three dimensions, described by Alkin and Christie as the three branches of the "evaluation theory tree" (Alkin and Christie 2012). These are the mobilisation of research *methods* (evaluation is based on systematic empirical investigation), the role of *values* in providing criteria for judging the intervention under study, and the focus on *the usefulness* of the evaluation.

[1] For example, Michael Patton's definition of evaluation as "the systematic collection of information about the activities, characteristics, and outcomes of programs to make judgements about the program, improve program effectiveness, and/or inform decisions about future programming" (Patton 1997, 23)

The use of systematic methods of empirical investigation is therefore one of the foundations of evaluation practice. This is how evaluation in the sense of evaluative research differs from the mere subjective judgement that the term 'evaluation' in its common sense may otherwise denote (Suchman 1967). Evaluation is first and foremost an applied research practice, and as such, it has borrowed a whole series of investigative techniques, both quantitative and qualitative, initially developed in basic research (e.g. questionnaires, quantitative analyses on databases, experimental methods, semi-structured interviews, observations, case studies, etc.). Beyond the techniques, the borrowing also concerns the methods of analysis and the conception of research designs. Despite this strong methodological link, evaluation does not boil down to a research practice (Wanzer 2021).This is suggested by the other two dimensions identified earlier (the concern for values and utility). In fact, the development of programme evaluation has given rise to a plurality of practices by a variety of public and private actors (public administration, consultants, NGOs, etc.), practices within which methodological issues are not necessarily central and where methodological rigour greatly varies.

At the same time, the practice of evaluation has remained weakly and very unevenly institutionalised in the university (Cox 1990), where it suffers in particular from a frequent devaluation of applied research practices, suspicions of complacency towards commissioners, and difficulties linked to its interdisciplinary nature (see *below*) (Jacob 2008). Thus, although it has developed its journals and professional conferences, evaluation is still the subject of very few doctoral programmes and dedicated recruitments. Practised to varying degrees by different academic disciplines (public health, economics and development are now particularly involved), and sometimes described as 'transdisciplinary' in terms of its epistemological scope (Scriven 1993), evaluation is still far from being an academic discipline in the institutional sense of the term. From an epistemological point of view, this non- (or weak) disciplinarisation of evaluation is to be welcomed. The fact remains, however, that this leads to weaknesses.

One of the consequences of this situation is a frequent lack of training for researchers in evaluation: particularly concerning the non-methodological dimensions of this practice (questions of values and utility), but also concerning certain approaches more specifically derived from evaluation practice.

Indeed, while evaluation has largely borrowed from social science methods, it has also fostered a number of methodological innovations. For example, the use of experimental methods first took off in the social sciences in the context of evaluation, initially in education in the 1920s and then in social policy, health and other fields from the 1960s onwards (Campbell and Stanley 1963). The link with medicine brought about by the borrowing of the model of the clinical trial (the notions of 'trial' and 'treatment' having thus been transposed to evaluation) then favoured the transfer from the medical sciences to evaluation of another method, systematic literature reviews, which consists in adopting a systematic protocol to search for existing publications on (a) given evaluative question(s) and to draw up a synthesis of their contributions (Hong and Pluye 2018; Belaid and Ridde 2020). Without being the only place where it is deployed, programme evaluation has also made a major contribution to the development and theorising of mixed methods, which consist of articulating qualitative and quantitative techniques in the same research (Baïz and Revillard 2022; Greene, Benjamin and Goodyear 2001; Burch and Heinrich 2016; Mertens 2017). Similarly, because of its central concern with the use of knowledge, evaluation has been a privileged site for the development of participatory research and its theorisation (Brisolara 1998; Cousins and Whitmore 1998; Patton 2018).

While these methods (experimental methods, systematic literature reviews, mixed methods, participatory research) are immediately applicable to fields other than programme evaluation, other methodological approaches and tools have been more specifically developed for this purpose[2]. This is particularly the case of theory-based evaluation (Weiss 1997; Rogers and Weiss 2007), encompassing a variety of approaches (realist evaluation, contribution analysis, outcome harvesting, etc.) which will be described below (Pawson and Tilley 1997; Mayne 2012; Wilson-Grau 2018). Apart from a few disciplines in which they are more widespread, such as public health or development (Ridde and Dagenais 2009; Ridde et al. 2020), these approaches are still little known to researchers who have undergone traditional training in research methods, including those who may be involved in evaluation projects.

A dialogue therefore needs to be renewed between evaluation and research: according to the reciprocal dynamic of the initial borrowing of research methods by evaluation, a greater diversity of basic research circles would now benefit from a better knowledge of the specific methods and approaches derived from the practice of evaluation. This is one of the vocations of LIEPP, which promotes a strengthening of exchanges between researchers and evaluation practitioners. Since 2020, LIEPP has been organising a monthly seminar on evaluation methods and approaches (METHEVAL), alternating presentations by researchers and practitioners, and bringing together a diverse audience[3]. This is also one of the motivations behind the book *Evaluation: Foundations, Controversies, Perspectives*, published in 2021, which aimed in particular to make researchers aware of the non-methodological aspects of evaluation (Delahais et al. 2021). This publication completes the process by facilitating the appropriation of approaches developed in evaluation such as theory-based evaluation, realistic evaluation, contribution analysis and outcome harvesting.

[2] As opposed to methods in the sense of methodological tools, approaches are situated in "a kind of in-between between theory and practice" (Delahais 2022), by embodying certain paradigms. In evaluation, some may be very methodologically oriented, but others may be more concerned with values, with the use of results, or with social justice (ibid.).

[3] The programme and resources from previous sessions of this seminar are available online: https://www.sciencespo.fr/liepp/fr/content/cycle-de-seminaires-methodes-et-approches-en-evaluation-metheval.html

Conversely, LIEPP believes that evaluation would benefit from being more open to methodological tools more frequently used in basic research and with which it tends to be less familiar, particularly because of the targeting of questions at the scale of the intervention. In fact, evaluation classically takes as its object an intervention or a programme, usually on a local, regional or national scale, and within a sufficiently targeted questioning perimeter to allow conclusions to be drawn regarding the consequences of the intervention under study. By talking about policy evaluation rather than programme evaluation in the strict sense, our aim is to include the possibility of reflection on a more macro scale in both the geographical and temporal sense, by integrating reflections on the historicity of public policies, on the arrangement of different interventions in a broader policy context (a welfare regime, for example), and by relying more systematically on international comparative approaches. Evaluation, in other words, must be connected to policy analysis – an ambition already stated in the 1990s by the promoters of an "*évaluation à la française*" (Duran, Monnier, and Smith 1995; Duran, Erhel, and Gautié 2018). This is made possible, for example, by comparative historical analysis and macro-level comparisons presented in this book. Another important implication of programme evaluation is that the focus is on the intervention under study. By shifting the focus, many basic research practices can provide very useful insights in a more prospective way, helping to understand the social problems targeted by the interventions. All the thematic research conducted in the social sciences provides very useful insights for evaluation in this respect (Rossi, Lipsey, and Freeman 2004). Among the methods presented in this book, experimental approaches such as laboratory experimentation or testing, which are not necessarily focused on interventions as such, help to illustrate this more prospective contribution of research to evaluation.

A dialogue between qualitative and quantitative approaches

By borrowing its methods from the social sciences, policy evaluation has also inherited the associated methodological and epistemological controversies. Although there are many calls for reconciliation, although evaluation is more likely to emphasise its methodological pragmatism (the evaluative question guides the choice of methods), and although it has been a driving force in the development of mixed methods, in practice, in evaluation as in research, the dialogue between quantitative and qualitative traditions (especially in their epistemological dimension) is not always simple.

Articulating different disciplinary and methodological approaches to evaluate public policies is the founding ambition of LIEPP. The difficulties of this dialogue, particularly on an epistemological level (opposition between positivism and constructivism), were identified at the creation of the laboratory (Wasmer and Musselin 2013). Over the years, LIEPP has worked to overcome these obstacles by organizing a more systematic dialogue between different methods and disciplines in order to enrich evaluation: through the development of six research groups co-led by researchers from different disciplines, through projects carried out by interdisciplinary teams, but also through the regular discussion of projects from one discipline or family of methods by specialists from other disciplines or methods. It is also through these exchanges that the need for didactic material to facilitate the understanding of quantitative methods by specialists in qualitative methods, and vice-versa, has emerged. This mutual understanding is becoming increasingly difficult in a context of growing technicisation of methods. This book responds to this need, drawing heavily on the group of researchers open to interdisciplinarity and to the dialogue between methods that has been built up at LIEPP over the years: among the 25 authors of this book, nine are affiliated to LIEPP and eight others have had the opportunity to present their research at seminars organised by LIEPP.

This book has therefore been conceived as a means of encouraging a dialogue between methods, both within LIEPP and beyond. The aim is not necessarily to promote the development of mixed-methods research, although the strengths of such approaches are described (Part III). It is first of all to promote mutual understanding between the different methodological approaches, to ensure that practitioners of qualitative methods understand the complementary contribution of quantitative methods, their scope and their limits, and vice versa. In doing so, the approach also aims to foster greater reflexivity in each methodological practice, through a greater awareness of what one method is best suited for and the issues for which other methods are more relevant. While avoiding excessive technicality, the aim is to get to the heart of how each method works in order to understand concretely what it allows and what it does not allow. We are betting that this practical approach will help to overcome certain obstacles to dialogue between methods linked to major epistemological oppositions (positivism versus constructivism, for example) which are not necessarily central in everyday research practice. For students and non-academic audiences (particularly among policymakers or NGOs who may have recourse to programme evaluations), the aim is also to promote a more global understanding of the contributions and limitations of the various methods.

Far from claiming to be exhaustive, the book aims to present some examples of three main families of methods or approaches: quantitative methods, qualitative methods, and mixed methods and cross-sectional approaches in evaluation[4]. In what follows, we present the general

[4] In doing so, it complements other methodological resources available in handbooks (Ridde and Dagenais 2009; Ridde et al. 2020; Newcomer, Hatry, and Wholey 2015; Mathison 2005; Weiss 1998; Patton 2015) or online: for example the Methods excellence network (https://www.methodsnet.org/), or in the field

organisation of the book and the different chapters, integrating them into a more global reflection on the distinction between quantitative and qualitative approaches.

At a very general level, quantitative and qualitative methods are distinguished by the density and breadth of the type of information they produce: whereas quantitative methods can produce limited information on a large number of cases, qualitative methods provide denser, contextualised information on a limited number of cases. But beyond these descriptive characteristics, the two families of methods also tend to differ in their conception of causality. This is a central issue for policy evaluation which, without being restricted to this question[5], was founded on investigating the impact of public interventions: to what extent can a given change observed be attributed to the effect of a given intervention? – In other words, a causal question (can a cause-and-effect relationship be established between the intervention and the observed change?). To understand the complementary contributions of quantitative and qualitative methods for evaluation, it is therefore important to understand the different ways in which they tend to address this central question of causality.

of evaluation, the resources compiled by the OECD (https://www.oecd.org/fr/cad/evaluation/keydocuments.htm), the UN's Evalpartners network (https://evalpartners.org/), or in France the methodological guides of the Institut des politiques publiques (https://www.ipp.eu/publications/guides-methodologiques-ipp/) and the Société coopérative et participative (SCOP) Quadrant Conseil (https://www.quadrant-conseil.fr/ressources/evaluation-impact.php#/).

[5] Evaluation also looks at, for example, the relevance, coherence, effectiveness, efficiency or sustainability of interventions. See OECD DAC Network on Development Evaluation (EvalNet) https://www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm

Quantitative methods

Experimental and quasi-experimental quantitative methods are based on a counterfactual view of causality: to prove that A causes B, it must be shown that, all other things being equal, if A is absent, B is absent (Woodward 2003). Applied to the evaluation of policy impact, this logic invites us to prove that an intervention causes a given impact by showing that in the absence of this intervention, all other things being equal, this impact does not occur (Desplatz and Ferracci 2017). The whole difficulty then consists of approximating as best as possible these 'all other things being equal' situations: what would have happened in the absence of the intervention, all other characteristics of the situation being identical? It is this desire to compare situations with and without intervention 'all other things being equal' that gave rise to the development of experimental methods in evaluation (Campbell and Stanley 1963; Rossi, Lipsey, and Freeman 2004).

Most experiments conducted in policy evaluation are field experiments, in the sense that they study the intervention in situation, as it is actually implemented. Randomised controlled trials (RCTs) (see Chapter 1) compare an experimental group (receiving the intervention) with a control group, aiming for equivalence of characteristics between the two groups by randomly assigning participants to one or the other group. This type of approach is particularly well suited to interventions that are otherwise referred to as 'experiments' in public policy (Devaux-Spatarakis 2014). These are interventions that public authorities launch in a limited number of territories or organisations to test their effects[6], thus allowing for the possibility of control groups. When this type of direct experimentation is not possible, evaluators can resort to several quasi-experimental methods, aiming to reconstitute comparison groups from

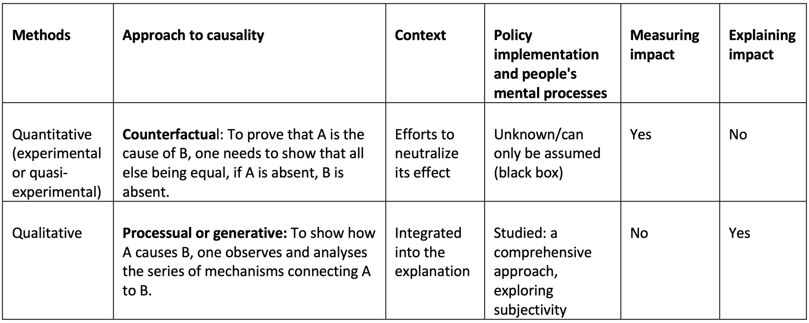
[6] Unfortunately, these government initiatives are far from being systematically and rigorously evaluated.

already existing situations and data (thus without manipulating reality, unlike experimental protocols) (Fougère and Jacquemet 2019). The difference-in-differences method uses a time marker at which one of the two groups studied receives the intervention and the other does not, and measures the impact of the intervention by comparing the results before and after this time (see Chapter 2). Discontinuity regression (see Chapter 3) reconstructs a target group and a control group by comparing the situations on either side of an eligibility threshold set by the policy under study (e.g. eligibility for the intervention at a given age, income threshold, etc.). Finally, matching methods (see Chapter 4) consist of comparing the situations of beneficiaries of an intervention with those of non-beneficiaries with the most similar characteristics.

In addition to these methods, which are based on real-life data, other quantitative impact assessment approaches are based on computer simulations or laboratory experiments. Microsimulation (see Chapter 5), the development of which has been facilitated by improvements in computing power, consists of estimating *ex ante* the expected impact of an intervention by taking into consideration a wide variety of data relating to the targeted individuals and simulating changes in their situation (e.g. ageing, changes in the labour market, fiscal policies, etc.). It also allows for a refined *ex post* analysis of the diversity of effects of a given policy on the targeted individuals. Policy evaluation can also rely on laboratory experiments (see Chapter 6), which make it possible to accurately measure the behaviour of individuals and, in particular, to uncover unconscious biases. Such analyses can, for example, be very useful in helping to design anti-discrimination policies, as part of an *ex ante* evaluation process. It is also in the context of reflection on these policies that testing methods (see Chapter 7) have been developed, making it possible to measure discrimination by sending fictitious applications in response to real offers (for example, job offers). But evaluation also seeks to measure the efficiency of interventions, beyond their impact. This implies comparing the results obtained with the cost of the policy under study and with those of alternative policies, in a cost-efficiency analysis approach (see Chapter 8).

Qualitative methods

While they are also compatible with counterfactual approaches, qualitative methods are more likely to support a generative or processual conception of causality (Maxwell 2004; 2012; Mohr 1999). Following this logic, causality is inferred, not from relations between variables, but from the analysis of the processes through which it operates. While the counterfactual approach establishes whether A causes B, the processual approach shows how (through what series of mechanisms) A causes B, through observing the empirical manifestations of these causal mechanisms that link A and B. In so doing, it goes beyond the behaviourist logic which, in counterfactual approaches, conceives the intervention according to a stimulus-response mechanism, the intervention itself then constituting a form of black box. Qualitative approaches break down the intervention into a series of processes that contribute to producing (or preventing) the desired result: this is the general principle of theory-based evaluations (presented in the third part of this book in Chapter 20 as they are also compatible with quantitative methods). This finer scale analysis is made possible by focusing on a limited number of cases, which are then studied in greater depth using different qualitative techniques. Particular attention is paid to the contexts, as well as to the mental processes and the logic of action of the people involved in the intervention (agents responsible for its implementation, target groups), in a comprehensive approach (Revillard 2018). Unlike quantitative methods, qualitative methods cannot measure the impact of a public policy; they can, however, explain it (and its variations according to context), but also answer other evaluative questions such as the relevance or coherence of interventions. Table 1 summarises these ideal-typical differences between quantitative and qualitative methods: it is important to specify that we are highlighting here the affinities of a given family of methods with a given approach to causality and a given consideration of processes and context, but this is an ideal-typical distinction which is far from exhausting the actual combinations in terms of methods and research designs.



Methods	Approach to causality	Context	Policy implementation and people's mental processes	Measuring impact	Explaining impact
Quantitative (experimental or quasi-experimental)	**Counterfactual**: To prove that A is the cause of B, one needs to show that all else being equal, if A is absent, B is absent.	Efforts to neutralize its effect	Unknown/can only be assumed (black box)	Yes	No
Qualitative	**Processual or generative:** To show how A causes B, one observes and analyses the series of mechanisms connecting A to B.	Integrated into the explanation	Studied: a comprehensive approach, exploring subjectivity	No	Yes

Table 1: Quantitative and qualitative methods, two different approaches to causality

The most emblematic qualitative research technique is probably direct observation or ethnography, coming from anthropology, which consists of directly observing the social situation being studied in the field (see Chapter 9). A particularly engaging method, direct observation is very effective in uncovering all the intermediate policy processes that contribute to producing its effects, as well as in distancing official discourse through the direct observation of interactions. The semi-structured interview (see Chapter 10) is another widely used qualitative research technique, which consists of a verbal interaction solicited by the researcher with a research participant, based on a grid of questions used in a very flexible manner. The interview aims both to gather information and to understand the experience and worldview of the interviewee. This method can also be used in a more collective setting, in the form of focus groups (see Chapter 11) or group interviews (see Chapter 12). As Ana Manzano points out in her chapter on focus groups, the terminologies for these group interview practices vary. Our aim in publishing two chapters on these techniques is not to rigidify the distinction but to provide two complementary views on these frequently used methods.

Although case studies (see Chapter 13) can use a variety of qualitative, quantitative and mixed methods, they are classically part of a qualitative research tradition because of their connection to anthropology. They allow interventions to be studied in context and are particularly suited to the analysis of complex interventions. Several case studies can be combined in the evaluation of the same policy; the way in which they are selected is then decisive. Process tracing (see Chapter 14), which relies mainly but not exclusively on qualitative enquiry techniques, focuses on the course of the intervention in a particular case, seeking to trace how certain actions led to others. The evaluator then acts as a detective looking for the "fingerprints" left by the mechanisms of change. The approach makes it possible to establish under what conditions, how and why an intervention works in a particular case. Finally, comparative historical analysis combines the two fundamental methodological tools of social science, comparison and history, to help explain large-scale social phenomena (see Chapter 15). It is particularly useful for reporting on the definition of public policies.

Mixed methods and cross-cutting approaches in evaluation

The third and final part of the book brings together a series of chapters on the articulation between qualitative and quantitative methods as well as on cross-cutting approaches that are compatible with a diversity of methods. Policy evaluation has played a driving role in the formalisation of the use of mixed methods, leading in particular to the distinction between different strategies for linking qualitative and quantitative methods (sequential exploratory, sequential explanatory or convergent design) (see

Chapter 16). Even when the empirical investigation mobilises only one type of method, it benefits from being based on a systematic mixed methods literature review. While the practice of systematic literature reviews was initially developed to synthesise results from randomised controlled trials, this practice has diversified over the years to include other types of research (Hong and Pluye 2018). The particularity of systematic mixed methods literature reviews is that they include quantitative, qualitative and mixed studies, making it possible to answer a wider range of evaluative questions (see Chapter 17).

Having set out this general framework on mixed methods and reviews, the following chapters present six cross-cutting approaches. The first two, macro-level comparisons and qualitative comparative analysis (QCA), tend to be drawn from basic research practices, while the other four (theory-based evaluation, realist evaluation, contribution analysis, outcome harvesting) are drawn from the field of evaluation. Macro-level comparisons (see Chapter 18) consist of exploiting variations and similarities between large entities of analysis (e.g. states or regions) for explanatory purposes: for example, to explain differences between large social policy models, or the influence of a particular family policy configuration on women's employment rate. Qualitative comparative analysis (QCA) is a mixed method which consists in translating qualitative data into a numerical format in order to systematically analyse which configurations of factors produce a given result (see Chapter 19). Based on an alternative, configurational conception of causality, it is useful for understanding why the same policy may lead to certain changes in some circumstances and not in others.

Developed in response to the limitations of experimental and quasi-experimental approaches to understanding how an intervention produces its impacts, theory-based evaluation consists of opening the 'black box' of public policy by breaking down the different stages of the causal chain linking the intervention to its final results (see Chapter 20). The following chapters fall broadly within this family of evaluation approaches. Realist evaluation (see Chapter 21) conceives of public policies as interventions that produce their effects through mechanisms that are only triggered in specific contexts. By uncovering context-mechanism-outcomes (CMO) configurations, this approach makes it possible to establish for whom, how and under what circumstances an intervention works. Particularly suited to complex interventions, contribution analysis (see Chapter 22) involves the progressive formulation of 'contribution claims' in a process involving policy stakeholders, and then testing these claims systematically using a variety of methods. Outcome harveting (see Chapter 23) starts from a broad understanding of observable changes, and then traces whether and how the intervention may have played a role in producing them. Finally, the last chapter is devoted to an innovative approach to evaluation, based on the concept of cultural safety initially developed in nursing science (see Chapter 24). Cultural safety aims to ensure that the evaluation takes place in a 'safe' manner for stakeholders, and in particular for the minority communities targeted by the intervention under study, i.e. that the evaluation process avoids reproducing mechanisms of domination (aggression, denial of identity, etc.) linked to structural inequalities. To this end, various participatory techniques are used at all stages of the evaluation. This chapter is thus an opportunity to emphasise the importance of participatory dynamics in evaluation, also highlighted in several other contributions.

A didactic and illustrated presentation

To facilitate reading and comparison between methods and approaches, each chapter is organised according to a common outline based on five main questions:

1) What does this method/approach consist of?

2) How is it useful for policy evaluation?

3) An example of the use of this method/approach;

4) What are the criteria for judging the quality of the use of this method/approach?

5) What are the strengths and limitations of this method/approach compared to others?

The book is published directly in two languages (French and English) in order to facilitate its dissemination. The contributions were initially written in one or the other language according to the preference of the authors, then translated and revised (where possible) by them. A bilingual glossary is available below to facilitate the transition from one language to the other.

The examples used cover a wide range of public policy areas, studied in a variety of contexts: pensions in Italy, weather and climate information in Senegal, minimum wage in New Jersey, reception in public services in France, child development in China, the fight against smoking among young people in the United Kingdom, health financing in Burkina Faso, the impact of a summer school on academic success in the United States, soft skills training in Belgium, the development of citizen participation to improve public services in the Dominican Republic, a nutrition project in Bangladesh, universal health coverage in six African countries, etc. The many examples presented in the chapters illustrate the diversity and current vitality of evaluation research practices.

Far from claiming to be exhaustive, this publication is an initial summary of some of the most widely used methods. The collection is intended to be enriched by means of publications over time in the open access collection of LIEPP methods briefs[7].

[7] LIEPP Methods briefs : https://www.sciencespo.fr/liepp/en/publications.html#LIEPP%20methods%20briefs

Cited references

Alkin, Marvin. and Christina Christie. 2012. ‘An Evaluation Theory Tree’. In *Evaluation Roots*, edited by Marvin Alkin and Christina Christie. London: Sage. https://doi.org/10.4135/9781412984157.n2.

Baïz, Adam. and Anne Revillard. 2022. *Comment Articuler Les Méthodes Qualitatives et Quantitatives Pour Évaluer L'impact Des Politiques Publiques?* Paris: France Stratégie. https://www.strategie.gouv.fr/publications/articuler-methodes-qualitatives-quantitatives-evaluer-limpact-politiques-publiques.

Belaid, Loubna. and Valéry Ridde. 2020. ‘Une Cartographie de Quelques Méthodes de Revues Systématiques’. *Working Paper* CEPED N°44.

Brisolara, Sharon. 1998. ‘The History of Participatory Evaluation and Current Debates in the Field’. *New Directions for Evaluation*, (80): 25–41. https://doi.org/10.1002/ev.1115.

Burch, Patricia. and Carolyn J. Heinrich. 2016. *Mixed Methods for Policy Research and Program Evaluation*. Los Angeles: Sage.

Campbell, Donald T.. and Julian C. Stanley. 1963. ‘Experimental and Quasi-Experimental Designs for Research’. In *Handbook of Research on Teaching*. Houghton Mifflin Company.

Cousins, J. Bradley. and Elizabeth Whitmore. 1998. ‘Framing Participatory Evaluation’. *New Directions for Evaluation*, (80): 5–23.

Cox, Gary. 1990. ‘On the Demise of Academic Evaluation’. *Evaluation and Program Planning*, 13(4): 415–19.

Delahais, Thomas. 2022. ‘Le Choix Des Approches Évaluatives’. In *L'évaluation En Contexte de Développement: Enjeux, Approches et Pratiques*, edited by Linda Rey, Jean Serge Quesnel, and Vénétia Sauvain, 155–80. Montréal: JFP/ENAP.

Delahais, Thomas, Devaux-Spatarakis, Agathe, Revillard, Anne, and Ridde, Valéry. eds. 2021. *Evaluation: Fondements, Controverses, Perspectives*. Québec: Éditions science et bien commun. https://scienceetbiencommun.pressbooks.pub/evaluationanthologie/.

Desplatz, Rozenn. and Marc Ferracci. 2017. *Comment Évaluer l'impact Des Politiques Publiques? Un Guide à l'usage Des Décideurs et Praticiens*. Paris: France Stratégie.

Devaux-Spatarakis, Agathe. 2014. 'L'expérimentation "telle Qu'elle Se Fait": Leçons de Trois Expérimentations Par Assignation Aléatoire'. *Formation Emploi*, (126): 17–38.

Duran, Patrice. Erhel, Christine. and Gautié, Jérôme. 2018. 'L'évaluation des politiques publiques'. *Idées Économiques et Sociales*, 193(3): 4–5.

Duran, Patrice. and Monnier, Eric. and Smith, Andy. 1995. 'Evaluation à La Française: Towards a New Relationship between Social Science and Public Action'. *Evaluation*, 1(1): 45–63. https://doi.org/10.1177/135638909500100104.

Fougère, Denis. and Jacquemet, Nicolas. 2019. 'Causal Inference and Impact Evaluation'. *Economie et Statistique*, (510-511–512): 181–200. https://doi.org/10.24187/ecostat.2019.510t.1996.

Greene, Jennifer C.. and Lehn, Benjamin. and Leslie Goodyear. 2001. 'The Merits of Mixing Methods in Evaluation'. *Evaluation*, 7(1): 25–44.

Hong, Quan Nha. and Pierre Pluye. 2018. 'Systematic Reviews: A Brief Historical Overview'. *Education for Information*, (34): 261–76.

Jacob, Steve. 2008. 'Cross-Disciplinarization a New Talisman for Evaluation?' *American Journal of Evaluation* 19(2): 175–94. https://doi.org/10.1177/1098214008316655.

Mathison, Sandra. 2005. *Encyclopedia of Evaluation*. London: Sage.

Maxwell, Joseph A. 2004. ‘Using Qualitative Methods for Causal Explanation’. *Field Methods*, 16(3): 243–64. https://doi.org/10.1177/1525822X04266831.

———. 2012. A *Realist Approach for Qualitative Research*. London: Sage.

Mayne, John. 2012. ‘Contribution Analysis: Coming of Age?’ *Evaluation*, 18(3): 270–80. https://doi.org/10.1177/1356389012451663.

Mertens, Donna M. 2017. *Mixed Methods Design in Evaluation*. Thousand Oaks: Sage. Evaluation in Practice Series. https://doi.org/10.4135/9781506330631.

Mohr, Lawrence B. 1999. ‘The Qualitative Method of Impact Analysis’. *American Journal of Evaluation*, 20(1): 69–84. https://doi.org/10.1177/109821409902000106.

Newcomer, Kathryn E.. and Harry P. Hatry. and Joseph S. Wholey. 2015. *Handbook of Practical Program Evaluation*. Hoboken: Wiley.

Patton, Michael Q. 1997. *Utilization-Focused Evaluation*. 3rd ed. Thousand Oaks: Sage.

———. 2015. *Qualitative Research and Evaluation Methods: Integrating Theory and Practice*. London: Sage.

———. 2018. *Utilization-Focused Evaluation*. London: Sage.

Pawson, Ray, and Nicholas Tilley. 1997. *Realistic Evaluation*. London: Sage.

Revillard, Anne. 2018. *Quelle place pour les méthodes qualitatives dans l’évaluation des politiques publiques*? Paris: LIEPP Working Paper n°81.

Ridde, Valéry. and Christian Dagenais. 2009. *Approches et Pratiques En Évaluation de Programme*. Montréal: Presses de l’Université de Montréal.

Ridde, Valéry. and Christian Dagenais (eds). 2020. *Évaluation Des Interventions de Santé Mondiale: Méthodes Avancées*. Québec: Éditions science et bien commun.

Rogers, Patricia J.. and Carol H. Weiss. 2007. 'Theory-Based Evaluation: Reflections Ten Years on: Theory-Based Evaluation: Past, Present, and Future'. *New Directions for Evaluation*, (114): 63–81. https://doi.org/10.1002/ev.225.

Rossi, Peter H.. and Mark W. Lipsey. and Howard E. Freeman. 2004. *Evaluation: A Systematic Approach*. London: Sage.

Scriven, Michael. 1993. 'Hard-Won Lessons in Program Evaluation.' *New Directions for Program Evaluation*, (58): 5–48.

Suchman, Edward A. 1967. *Evaluative Research. Principles and Practice in Public Service and Social Action Programs*. New York: Russell Sage Foundation.

Wanzer, Dana L. 2021. 'What Is Evaluation? Perspectives of How Evaluation Differs (or Not) From Research'. *American Journal of Evaluation*, 42(1): 28–46. https://doi.org/10.1177/1098214020920710.

Wasmer, Etienne. and Christine Musselin. 2013. *Évaluation Des Politiques Publiques: Faut-Il de l'interdisciplinarité?* Paris: LIEPP Methodological discussion paper n°2.

Weiss, Carol H. 1997. 'Theory-Based Evaluation: Past, Present, and Future'. *New Directions for Evaluation* (76): 41–55. https://doi.org/10.1002/ev.1086.

———. 1998. *Evaluation: Methods for Studying Programs and Policies*. Upper Saddle River, NJ: Prentice-Hall.

Wilson-Grau, Ricardo. 2018. *Outcome Harvesting: Principles, Steps, and Evaluation Applications*. IAP.

Woodward, James. 2003. *Making Things Happen a Theory of Causal Explanation*. Oxford Studies in Philosophy of Science. New York: Oxford University Press.

Glossary

English-French glossary

English	***French***	**English**	***French***
Abductive approach	*Démarche abductive*	Ideal-types	*Idéaux-types*
Analytical generalisation	*Généralisation analytique*	Impact	*Impact*
Asymmetrical causality	*Causalité asymétrique*	Impact logic diagram	*Diagramme logique d'impact*
Attitudes	*Attitudes*	Impact path	*Chemin d'impact*
Automatic/non-automatic response	*Réponse automatique/ non-automatique*	Indigenous evaluation	*Évaluation autochtone*
Bayesian reasoning	*Raisonnement bayésien*	Induction	*Induction*
Behaviours	*Comportements*	Internal/ external validity	*Validité interne/ externe*
Case study	*Étude de cas*	Interpretivism	*Interprétativiste*
Causal chain	*Chaîne causale*	Interval amplitude	*Amplitude d'intervalle*
Causal complexity	*Complexité causale*	Intervention logic	*Logique d'intervention*
Causal pathways	*Chemins d'impact*	Interview	*Entretien*
Causal principles	*Principes causaux*	Laboratory experimentation	*Expérimentation en laboratoire*
Combinations of conditions	*Combinaisons de conditions*	Literature review	*Revue de la littérature*
Comparison	*Comparaison*	Longitudinal dimension of data	*Dimension longitudinale des données*
Complex interventions	*Interventions complexes*	Macro-social units	*Unités macrosociales*
Confidence interval	*Intervalle de confiance*	Middle range theory	*Théorie de moyenne portée*
Configurations	*Configurations*	Mixed method	*Méthode mixte*
Conjunctural causation	*Causalité conjoncturelle*	Mixed methods literature review	*Revue mixte de la littérature*

Constructivism	*Constructivisme*	Modeling	*Modélisation*
Contamination	*Contamination*	Monotonicity	*Monotonicité*
Context-mechanism-outcome (CMO) configuration	*Configuration contexte -mécanisme-effet* (CME)	Narrative approach	*Approche narrative*
Contributing claims	*Hypothèses de contribution*	Non-conscious behavioural bias	*Biais de comportement non conscients*
Contribution pathways	*Chemins de contribution*	Observable changes	*Changements observables*
Convergent design	*Devis/design convergent*	Observation window	*Fenêtre d'observation*
Cost-effectiveness,	*Coût/efficacité*	Outcome harvesting	*Récolte d'incidences*
Critical juncture	*Point d'inflexion*	Outcome statements	*Enoncé d'incidences*
Critical realism	*Réalisme critique*	Parallel trends	*Tendances parallèles*
Culturally sensitive evaluation	*Evaluation attentive aux différences culturelles*	Path dependency	*Dépendance au sentier emprunté*
Decoloniality	*Décolonialité*	Policy implementation	*Mise en œuvre des politiques publiques*
Design	*Devis/design*	Process theory of change (ptoc)	*Théorie du changement relative aux processus* (TCP)
Difference-in-differences	*Doubles/triples différences*	Process tracing	*Traçage de processus*
Direct observation	*Observation directe*	Propensity score	*Score de propension*
Effectiveness	*Efficacité*	Qualitative method	*Méthode qualitative*
Efficiency	*Efficience*	Quantitative method	*Méthode quantitative*

Eligibility threshold	*Seuil d’éligibilité*	Quasi-experimental methods	*Méthodes quasi-expérimentales*
Empirical triangulation	*Triangulation empirique*	Random assignment	*Affectation aléatoire*
Entropy balancing	*Equilibrage par entropie*	Realist evaluation	*Evaluation réaliste*
Equifinality	*Equifinalité*	Semi-structured interview	*Entretien semi-directif*
Ethnography	*Ethnographie*	Sequential explanatory design	*Devis/design séquentiel explicatif*
Evidence	*Preuves*	Sequential exploratory design	*Devis/design séquentiel exploratoire*
Evidence repository	*Archives ouvertes compliant les résultats d’évaluations déjà réalisées*	Similarities	*Similitudes*
Ex post evaluation	*Evaluation ex post*	Single/multiple cases	*Cas unique/multiples*
Experimental method	*Méthode expérimentale*	Standard deviation	*Ecart-type*
Experimental/treatment and control groups	*Groupes expérimentaux/de traitement et de contrôle*	Static/dynamic microsimulation	*Micro-simulation statique/dynamique*
Fingerprints	*Empreintes digitales*	Strict/fuzzy regression discontinuity	*Régression sur discontinuité stricte/floue*
Flow chart	*Logigramme*	Synthetic/artificial control group	*Groupe de contrôle synthétique/artificiel*
Focus group	*Focus group*	Systematic identification of cross patterns	*Identification systématique de schémas croisés*
Forcing variable	*Variable de forçage*	Systematic mixed review	*Revue systématique mixte*

Fuzzy cognitive mapping	*Cartographie cognitive floue*	Theory of change	*Théorie du changement*
Group interview	*Entretien de groupe*	Theory-based evaluation	*Evaluation basée sur la théorie*

French-English glossary

French	***English***	**French**	***English***
Affectation aléatoire	*Random assignment*	Evaluation réaliste	*Realist evaluation*
Amplitude d'intervalle	*Interval amplitude*	Expérimentation en laboratoire	*Laboratory experimentation*
Approche narrative	*Narrative approach*	Fenêtre d'observation	*Observation window*
Archives ouvertes compliant les résultats d'évaluations déjà réalisées	*Evidence repository*	Focus group	*Focus group*
Attitudes	*Attitudes*	Généralisation analytique	*Analytical generalisation*
Biais de comportement non conscients	*Non-conscious behavioural bias*	Groupe de contrôle synthetique/ artificiel	*Synthetic/ artificial control group*
Cartographie cognitive floue	*Fuzzy cognitive mapping*	Groupes expérimentaux/ de traitement et de contrôle	*Experimental/ treatment and control groups*
Cas unique/multiples	*Single/multiple cases*	Hypothèses de contribution	*Contributing claims*
Causalité asymétrique	*Asymmetrical causality*	Idéaux-types	*Ideal-types*
Causalité conjoncturelle	*Conjunctural causation*	Identification systématique de schémas croisés	*Systematic identification of cross patterns*
Chaîne causale	*Causal chain*	Impact	*Impact*
Changements observables	*Observable changes*	Induction	*Induction*
Chemin d'impact	*Impact path*	Interprétativisme	*Interpretivism*
Chemins d'impact	*Causal pathways*	Intervalle de confiance	*Confidence interval*
Chemins de contribution	*Contribution pathways*	Interventions complexes	*Complex interventions*
Combinaisons de conditions	*Combinations of conditions*	Logigramme	*Flow chart*

Comparaison	*Comparison*	Logique d’intervention	*Intervention logic*
Complexité causale	*Causal complexity*	Méthode expérimentale	*Experimental method*
Comportements	*Behaviours*	Méthode mixte	*Mixed method*
Configuration contexte-mécanisme-effet (cme)	*Context-mechanism-outcome (cmo) configuration*	Méthode qualitative	*Qualitative method*
Configurations	*Configurations*	Méthode quantitative	*Quantitative method*
Constructivisme	*Constructivism*	Méthodes quasi-expérimentales	*Quasi-experimental methods*
Contamination	*Contamination*	Microsimulation statique/ dynamique	*Static/dynamic microsimulation*
Coût/efficacité	*Cost-effectiveness*,	Mise en oeuvre des politiques publiques	*Policy implementation*
Décolonialité	*Decoloniality*	Modélisation	*Modeling*
Démarche abductive	*Abductive approach*	Monotonicité	*Monotonicity*
Dépendance au sentier emprunté	*Path dependency*	Observation directe	*Direct observation*
Devis/design	*Design*	Point d’inflexion	*Critical juncture*
Devis/design convergent	*Convergent design*	Preuves	*Evidence*
Devis/design séquentiel explicatif	*Sequential explanatory design*	Principes causaux	*Causal principles*
Devis/design séquentiel exploratoire	*Sequential exploratory design*	Raisonnement bayésien	*Bayesian reasoning*
Diagramme logique d’impact	*Impact logic diagram*	Réalisme critique	*Critical realism*

Dimension longitudinale des données	*Longitudinal dimension of data*	Récolte d'incidences	*Outcome harvesting*
Doubles/triples différences	*Difference-in-differences*	Régression sur discontinuité stricte/floue	*Strict/fuzzy regression discontinuity*
Ecart-type	*Standard deviation*	Réponse automatique/non -automatique	*Automatic/non -automatic response*
Efficacité	*Effectiveness*	Revue de la littérature	*Literature review*
Efficience	*Efficiency*	Revue mixte de la littérature	*Mixed methods literature review*
Empreintes digitales	*Fingerprints*	Revue systématique mixte	*Systematic mixed review*
Enoncé d'incidences	*Outcome statements*	Score de propension	*Propensity score*
Entretien	*Interview*	Seuil d'éligibilité	*Eligibility threshold*
Entretien de groupe	*Group interview*	Similitudes	*Similarities*
Entretien semi-directif	*Semi-structured interview*	Tendances parallèles	*Parallel trends*
Equifinalité	*Equifinality*	Théorie de moyenne portée	*Middle range theory*
Equilibrage par entropie	*Entropy balancing*	Théorie du changement	*Theory of change*
Ethnographie	*Ethnography*	Théorie du changement relative aux processus (tcp)	*Process theory of change (ptoc)*
Etude de cas	*Case study*	Traçage de processus	*Process tracing*
Evaluation attentive aux différences culturelles	*Culturally sensitive evaluation*	Triangulation empirique	*Empirical triangulation*
Evaluation autochtone	*Indigenous evaluation*	Unités macro-sociales	*Macro-social units*

Evaluation basée sur la théorie	*Theory-based evaluation*	Validité interne/ externe	*Internal/ external validity*
Evaluation ex post	Ex post evaluation	Variable de forçage	*Forcing variable*

PARTIE I
QUANTITATIVE METHODS

1. Randomised Controlled Trials

CARLO BARONE

Abstract

Randomised controlled trials (RCTs) aim at measuring the impact of a given intervention by comparing the outcomes of an experimental group (receiving the intervention) and a control group (not receiving it), to which individuals are randomly assigned. It is a useful quantitative method of ex ante evaluation, to test the impact of a program at a stage when it has not yet reached the totality of its target population (making the control group possible).

Keywords: Quantitative methods, experimental method, experimental/ treatment and control groups, random assignment, treatment, contamination

I. What does this method consist of?

Randomised Controlled Trials (RCTs) assess the impact of a policy by comparing two groups: one of them is given access to the policy (experimental group), while the other is temporarily excluded from the policy (control group). The researcher translates the goals of the policy into quantitative outcomes measures and assesses the efficacy of the policy by measuring these outcomes across these two groups. If the experimental group displays better values on these outcome measures, we conclude that the policy is effective. However, this conclusion is valid if, and only if, we can assume that the two groups were perfectly equivalent. This is why the assignment to the two groups must be done randomly: if the sample is sufficiently large, the random assignment ensures that the two groups are, on average, initially equivalent on all characteristics, known or unknown by the researcher, measured or unmeasured in the evaluation study. Hence, any difference in the outcomes observed after the implementation of the policy can be interpreted as an impact of the policy.

When conducting an RCT, the researcher draws a sample of individuals and invites them to participate in the study, explaining that they may be assigned to either the experimental or the control group. Among the participants who have accepted to participate, half of them will be randomly assigned to the treatment and half to the control group. This 50%-50% ratio is the most common one because it results in more precise estimates than unbalanced ratios (e.g., 70%-30%). Before delivering the intervention, we may carry out a baseline measurement of the outcomes. This is not strictly necessary, but it is often done for several reasons, for instance because it allows the researcher to study the impacts of the treatment in a more dynamic way by comparing variations in the outcomes across the two groups.

While the randomisation is a necessary condition to make plausible causal claims when comparing the two groups, it is not a sufficient condition. In particular, the control group must remain excluded from the policy during the entire period of implementation of the policy, that is, we must avoid any form of treatment contamination. This implies, for instance, that individuals of the two groups do not communicate about the treatment objectives and contents. Moreover, when individuals are assigned to the control group, they may react by trying to replace the treatment with a similar treatment. Treatment contamination and replacement can invalidate causal inferences if they happen on a large scale. Hence, the key requirement is that the control group acts 'as usual' and it is important that the researcher designs and presents the study in such a way to ensure that this is the case. Hence, while the randomisation is important, it is no less important to ensure the highest degree of control of these experimental conditions. The term 'randomised controlled trial' thus describes the two key requirements to make solid causal inferences: random assignment and control of the experimental conditions.

II. How is this method useful for policy evaluation?

RCTs are aimed at estimating the causal impacts of policies, that is, at assessing whether policies produce changes in the outcomes reflecting their goals. The main challenge is that, even if a given policy is completely ineffective, these outcomes may change because of other policies, other economic or socio-cultural changes affecting these outcomes. For instance, we may deliver a training programme to unemployed individuals to improve their employability and observe the employment rates of individuals participating in this programme. However, it is unclear whether any observed change in this outcome can be attributed to the policy. For instance, it could be due to the economic cycle as well to any kind of other economic, labour or welfare policy (e.g., fiscal incentives to hire unemployed individuals, changes in eligibility rules for unemployment benefits, etc.). Hence, a simple pre-post comparison would be unable to isolate the genuine causal impact of this policy.

RCTs are not the only type of causal impact evaluation method, for instance regression discontinuity designs are another option. RCTs are a form of ex ante evaluation, that is, they must be carried out before the policy is delivered to the whole population of potential beneficiaries. This is because RCTs suppose that the policy is not delivered to some individuals, who constitute the control group. If the policy has already been generalised, RCTs are unfeasible. We may then resort to other types of causal impact evaluation methods to isolate the genuine causal impact of the policy.

III. An example of application: what messages best favour tax compliance?

Tax compliance, that is the truthful reporting of taxable income and the timely payment of taxes, is essential to finance public services. Researchers partnered with the tax authority in Belgium to test the impact of different messages encouraging tax compliance (De Neve et al, 2019). Between 2014 and 2016, researchers randomly assigned around 2.5 million taxpayers to receive different messages: simplified messages presenting the key information in simpler terms, deterrence messages aimed at making the consequences of non-compliance explicit, and tax morale messages aimed at motivating taxpayers to appreciate the importance of compliance for the provision of public goods. The remaining 4 million taxpayers were assigned to a comparison group where taxpayer communication remained unchanged (this sample size is exceptional, most RCTs are based on a few hundreds or thousands cases). Using administrative data, researchers measured the impact of the intervention on the probability of making a payment or filing their taxes, and the amount of reported income. Simpler communication had the largest effect on tax compliance, inducing people to file and pay their taxes sooner. Adding deterrence messages further enhanced compliance, while tax morale messages were ineffective.

IV. What are the criteria for judging the quality of the mobilisation of this method?

In some contexts, experiments are unfeasible because the risks of treatment contamination or replacement are too high, for instance when treated and controlled, individuals can easily communicate on the contents of an information intervention and are highly motivated to do so. Some policies cannot be tested with an RCT because, by construction, they involve the whole population, therefore we cannot temporarily exclude the control group. For instance, this is the case of several macroeconomic, foreign or defence policies (for instance, a change in the military expenses).

Moreover, while most commonly we assign individuals to the treatment or control group, sometimes we may assign whole families, streets or villages to the treatment or to the control status. For instance, this is the case when a given intervention is more effectively delivered, or can only be delivered, at these supra-individual levels. These types of higher-level randomisations (cluster randomisation) can be necessary or extremely practical, but they demand large sample sizes and thus large budgets.

Finally, we should keep in mind that internal validity (i.e., the strength of causal inferences in the case under study) is only one of the quality criteria in evaluation research. Another important criterion is external validity, that is, the generalisability of conclusions beyond the sample under study. This second criterion, when applied to RCTs, demands that we draw large, random samples of the population under study and that participants do not drop out of the study or that drop out rates are not too high. A third important criterion relates to the validity and reliability of the outcome measures, including the capability to observe the long-term outcomes of a policy, and the coverage of all potential (positive and negative) effects of the policy.

V. What are the strengths and limitations of this method compared to others?

As explained above, the main strength of the RCTs is that they allow assessing the genuine causal impact of a policy before delivering it to the whole population of beneficiaries. In clinical research, RCTs are the standard method to assess the efficacy of any kind of therapy or medicament and they are increasingly used for the evaluation of public policies, more so in educational, labour market, health and housing policies.

The most common applications of this method involve the randomisation between two groups of individuals. However, sometimes we may arrange three or more groups of individuals in order to compare qualitatively different variants of an intervention or different dosages of the intervention. For instance, in a study to promote the use of bike-sharing services, we may compare the control group to a first treatment group that has information about bike-sharing, a second treatment group receiving a monetary incentive and a third group receiving a larger monetary incentive.

RCTS are not always feasible. In particular, policymakers or potential participants may refuse the principle of randomisation. Indeed, some people argue that experiments are 'unethical' because they exclude the individuals of the control group from the benefits of the policy. This critique forgets that the exclusion is temporary, that is, it lasts only for the time needed to demonstrate that policy is effective. This temporary exclusion allows assessing if the policy is effective before generalising it to the whole population. Moreover, the resources available in ex ante evaluation studies allow treating only a small share of the total population, so treating everyone would anyway be impossible: the random assignment instead gives everyone the same chances of being treated.

It is critically important that researchers explain in simple terms why randomisation is ethical and why it is necessary to ensure the reliability of the comparisons between the two groups. Whenever it is possible, the social acceptability of the randomisation can be increased by creating a waiting-list, that is, the control group receives the policy at the end of the study, or a compensatory treatment (a treatment that is different from the one under study and that does not affect the outcome of the study). For instance, in a study providing information on childcare services to pregnant mothers to enhance the recourse to these services, the control group may receive this information at the end of the study or may receive some other type of information for instance on healthy practices during pregnancy. If a waiting list is created, it is not possible to observe long-term outcomes because the control group is no more excluded from the intervention. Waiting lists and compensatory treatments can be used also to reduce the risk that the individuals assigned to the control group drop out of the treatment. It is indeed important that the dropout rates of the two groups are similar in order to preserve their equivalence throughout the study.

Compared to laboratory experiments, RCTs have higher ecological validity, meaning that we are studying people in real life situations and in naturalistic contexts. Hence, the risk that their behaviour is influenced by the awareness of being part of a study is less important. At the same time, relative to laboratory experiments, RCTs allow a lower degree of control on the behaviour of participants. In clinical and psychological experiments, the awareness of being treated is often neutralised by administering placebos to the control group, that is, treatments that are specifically designed to have no effect. In social policies, this practice is less common because we tend to regard the benefits deriving from the awareness of being treated as an integral part of the policy.

Most fundamentally, while RCTs are a reliable tool to assess the causal impacts of policies, they are not in a strong position to investigate the underlying processes. For instance, if an RCTs concludes that a policy is ineffective or less effective than expected, this method is unable to explain what did not work and how we may improve this policy. For this reason, it is extremely useful to integrate RCTs with qualitative techniques of process evaluation. Moreover, the beliefs and perceptions of the policy that beneficiaries and implementers have may be investigated using qualitative or survey interviews.

Some bibliographical references to go further

De Neve, Jan-Emmanuel. and Imbert, Clement. and Spinnewijn, Johannes. and Tsankova, Teodora. and Luts, Maarten. 2019. “How to Improve Tax Compliance? Evidence from Population-wide Experiments in Belgium.” Working paper.

White, Howard. and Sabarwal, Shagun. and De Hoop, Thomas. 2014. *Randomised Controlled Trials*, Notes méthodologiques, Évaluation d’impact no7, Unicef.
https://www.unicef-irc.org/publications/pdf/MB7FR.pdf

Gibson, Michael. and Sautmann, Anja. Last updated April 2021. *Introduction to randomized evaluations*, Abdul Latif Jameel Poverty Action Lab.
https://www.povertyactionlab.org/resource/introduction-randomized-evaluations

Gertler, Paul. and Martinez, Sebastian. and Premand, Patrick. and Rawlings, Laura. and Vermeersch, Christel. 2016. *Impact Evaluation in Practice*, second ed. World Bank Group (Chapters 3 and 4 of this manual)
https://openknowledge.worldbank.org/bitstream/handle/10986/25030/9781464807794.pdf?sequence=2&isAllowed=y

2. Difference-in-differences Method

DENIS FOUGÈRE AND NICOLAS JACQUEMET

Abstract

The difference-in-differences method is a quantitative, quasi-experimental method to assess the impact of an intervention by setting up comparison groups and measuring the change in an outcome between a pre- and a post-intervention period when only one of the two groups has access to the intervention. This method is very useful for ex-post impact evaluation.

Keywords: Quantitative methods, quasi-experimental methods, difference-in-differences, difference-in-difference-in-differences, longitudinal data, parallel trends, entropy balancing, synthetic/artificial control group

I. How is this method useful for policy evaluation?

Although the evaluation of public policies covers a very broad set of issues and tools, which goes well beyond the mere quantification of their effects, the question of the effectiveness of policies implemented in the past is obviously of primary importance, as it constitutes a useful guide for considering their continuation, evolution, generalisation or even abandonment.

Such an evaluation requires a clear definition of the objectives pursued. For example, the effect of raising the minimum retirement age on retirement frequency, the impact of setting up a university grant system on transitions to higher education, or the consequences of introducing financial aid to facilitate access to healthcare on the use of the healthcare system. A natural reflex, which appears (too) often in the public debate, is to compare the situation of people who have benefited from the interventions implemented with that of others who have not. In order to assess the effectiveness of an unemployment insurance reform offering personalised job search assistance, one could thus compare those who benefited from this assistance with those who did not. As the study by Fougère, Kamionka and Prieto (2010, see Figure 3) illustrates, such a comparison shows unambiguously that job search assistance programs lead to a much slower return to employment for those who have benefited from them. Does this mean that services offered are detrimental to the probability of finding a job for unemployed workers?

Of course not. An alternative interpretation is that people who are offered job search assistance are precisely those who have the greatest difficulty in finding a job. When comparing their situation to that of unemployed people who have not received assistance, the implicit assumption is that the return to employment observed in this category can serve as a reference (i.e., a counterfactual) to the situation that would have been experienced by the beneficiaries in the absence of the assistance program. However, the beneficiaries are precisely those whose situation would have been particularly difficult in the absence of the assistance program. To avoid such confusions, the difference-in-differences method consists of defining the comparison group in such a way that the observed difference provides a more convincing estimate of the intervention effect.

II. What does this method consist of?

Suppose we observe changes between two dates in an outcome variable (also called a response variable or dependent variable) in two distinct groups. The first of these groups, called the treatment group, benefits from a given intervention or policy (referred to as the treatment); the second, called the control group, does not. The policy is implemented between the two dates. The measurement of the intervention effect is based exclusively on the variation of the outcome variable between these two dates. This variation differs in the two groups, generally from the moment the treatment comes into effect. It is this inflection in the difference between the two groups that is interpreted here as the average effect of the treatment on the outcome variable.

Why is this procedure called the difference-in-differences method? The first difference is the difference between the average value of the outcome variable in the treatment group at the second date (after implementation of the policy to be evaluated) and the average value of the same variable in the same group at the initial date (before implementation of the policy to be evaluated). From this first difference, we then subtract the analogous difference for the control group. The difference-in-differences method therefore exploits the longitudinal dimension of the data (or pseudo-longitudinal, as the individuals belonging to each of the groups may not remain the same over time) in order to provide an ex-post evaluation of the public policy that has been implemented.

The ability of this method to measure the average effect of the intervention is not based on the hypothesis that the non-beneficiaries can serve as a reference group for the beneficiaries in the absence of the intervention, but only on the fact that in the absence of the intervention, the average evolution of the outcome variable for the individuals in the treated group would have been the same as that observed in the control group (this is called the parallel trends assumption). The validity of this assumption, which cannot be verified, can be supported by the fact that before the policy was implemented, the outcome variable evolved in the same way in both groups (this is called the common pre-trend assumption). In contrast to the previous assumption, this second assumption can be tested using data observed prior to the implementation of the intervention, provided that the pre-intervention observation period is long enough – for example, at least five observations in both groups prior to the implementation of the policy being evaluated (these observations are called leads in the academic literature). The parallel trends assumption is equivalent to assuming that the pre-existing gap between the two groups, which may be explained by the various factors leading to different levels of the outcome variable within these groups, would have remained the same in the absence of the intervention, so that the observed change in this gap can be interpreted as the average effect of the intervention.

This approach is therefore only valid if the intervention leaves the outcome variable in the control group unchanged (this is the so-called SUTVA, i.e., Stable Unit Treatment Value Assumption). Indeed, any indirect effect of the intervention on this group (if, for example, the difficulty of finding a job increases because the acceleration of the return to work in the treatment group increases the tension in the labour market) calls into question the parallel trends assumption. Similarly, the parallel trends assumption could be challenged if the treatment group anticipates a positive effect of the intervention, and subsequently reduces job search intensity — a violation known as the Ashenfelter gap.

Given the many factors that can affect the validity of the approach, recent developments of the-difference-in-differences method aim in particular to refine the constitution of the groups in order to increase their comparability (see Roth et al., 2022, for a detailed description). It is for instance possible to use matching methods which, based on a statistical criterion, associate each person benefiting from the intervention with the person or persons in the control group whose observable characteristics are close – so that the comparison is carried out between statistical nearest neighbors – or the entropy balancing method which permits to equalise the first moments (mean, variance, skewness, etc.) of the distributions of the covariates. A similar approach can be applied to the outcome variable rather than to the distribution of observable characteristics. This is the goal of the synthetic control method, which consists of creating an artificial control group from the observed control group by means of an appropriate system of weights. This synthetic control group is constructed in such a way that the past evolution of the outcome variable within this synthetic group is identical to that of the same variable in the treatment group. For that purpose, we minimise, by reweighting the observations in the control group, the distance between the outcome variable in the treatment group and this variable in the synthetic control group before the intervention. When the number of treated units is very large, it is possible that the synthetic control of a treated unit is not unique. Several recent contributions have proposed solutions to this difficulty. Among these, some suggest the use of matrix completion techniques, others propose sampling-based inferential methods.

One of the most popular extensions which accounts for the existence of unobservable interactions between group and time characteristics that the difference-in-differences method might omit is the difference in difference-in-differences method. This method relies on the observation of two additional groups, a fake treatment group or a fake control group. For example, let us consider a health policy that is implemented in region A to people over 65. In order to evaluate the effects of this policy on the use of health care and on the health status of the persons concerned, it is possible to consider the persons aged 65 to 69 in region A as the treatment group, and to use the status of those aged 60 to 64 in this same region as the control group. A first difference-in-differences applied to these two groups should in principle produce an estimate of the average effect of the intervention on health care use and on the health status of people over 65 in region A. However, this approach can be criticised since it compares populations that are not quite the same in terms of their health status: people aged 68 or 69 are probably in poorer health than those aged 60 or 61, and therefore exposed to higher risks of health deterioration over time. To address this criticism, it is possible to consider the same age groups in a second region, say region B, where the same policy is not implemented, and then calculate a second difference-in-differences (DiD hereafter) estimate in region B. This second DiD estimate in region B can then be subtracted from that calculated in region A. The second DiD estimate applied to the two groups in region B eliminates the differences in health between age groups that naturally prevail in the population as a whole (the assumption of parallel trends is therefore weakened, and here concerns the relative difference between the two categories of population in each of the two regions).

In addition to the quality of the comparison between groups, a second limitation of the difference-in-differences method is that the effect of the intervention is not always identical within different subgroups of beneficiaries, or over time: then the effect of the intervention is said to be heterogeneous. By relying on the evolution of the gap between two groups only, this method only measures an average effect, which is only compatible with very large variations in the intervention effect between different subgroups. In order to study variations in the effect over time, it is useful to have observations of the outcome variable in both groups well beyond the date following the implementation of the intervention (such observations are sometimes called lags). This ensures that the policy being evaluated has significant effects in the medium term, or even in the long term if the statistical follow-up is long enough.

Such heterogeneity in the effects of the intervention also raises important difficulties when its diffusion in the treatment group is gradual. The usual method, which consists of integrating observations into the group of beneficiaries as they become eligible for the intervention, leads to unfounded conclusions (which can go so far as to conclude that an intervention with positive effects for all beneficiaries is ineffective). Recent studies recommend focusing only on observations that correspond to changes in treatment status, which implies to combine multiple difference-in-differences estimates calculated at all the dates at which the set of beneficiaries changes (see de Chaisemartin and d'Haultfoeuille, 2022, for a complete presentation).

III. An example of the use of this method in the field of employment

Like most labour market policies, the introduction of a minimum wage and the setting of its level is a delicate trade-off. When employers have a high bargaining power and can squeeze wages, the minimum wage provides some protection for employees and allows the benefits of production to be distributed more fairly. But the existence of a minimum wage also implies that all jobs that are less profitable than the minimum wage will not be offered on the market because they do not create enough value to cover the cost of wages. The challenge is therefore to set a minimum wage that rebalances wage bargaining without excessively damaging economic efficiency.

One of the most famous studies of the implementation of the difference-in-differences method is Card and Krueger's (1994) paper on the New Jersey minimum wage increase in April 1992. In this study, Card and Krueger compare the level of employment in the fast-food industry (which is very intensive in low-skilled jobs that are usually paid at the minimum wage level) in New Jersey and Pennsylvania in February 1992 and November 1992. These dates frame an increase in the minimum hourly wage from US$4.25 to US$5.05 in April 1992 in New Jersey, while at the same time the minimum hourly wage remained constant at US$4.25 in Pennsylvania. Observing a change in employment in New Jersey between February and November 1992 by means of a first difference does not allow us to attribute this change to the increase in the minimum wage in that state alone, particularly because other concomitant factors, such as weather or macroeconomic conditions, could also explain this change.

Furthermore, the difference in employment levels between the two states after the minimum wage was raised reflects not only the effect of the minimum wage policy but also the overall differences in the way the industry operates between New Jersey and Pennsylvania.

By including both New Jersey (the treatment group) and Pennsylvania (the control group) fast food restaurants, located on both sides of the state border, Card and Krueger can limit the effects of these two types of factors by a second difference. Under the assumption of parallel trends, the change in fast-food employment in Pennsylvania can be interpreted as the change in fast-food employment in New Jersey that would have occurred if the minimum hourly wage had not increased in that state. Card and Krueger's estimates suggest that the minimum wage increase was not accompanied by a decrease in employment in New Jersey. Specifically, Card and Krueger estimate that the $0.80 increase in the hourly minimum wage in New Jersey resulted in (caused) an increase of 2.75 full-time jobs on average in each fast-food restaurant in this state.

IV. What are the criteria for judging the quality of the mobilisation of this method?

The estimator obtained will be more informative (and the hypothesis of parallel trends more credible) if the control group is similar to the treatment group in terms of observable explanatory characteristics (avoiding over-interpretation of such comparisons, since unobservable heterogeneity may vary considerably between groups without this being detectable). Unless the constitution of the groups follows a procedure that imposes such a condition, it is appropriate to ensure this by comparing the distribution of observable characteristics across subgroups (e.g., in a sample of employees, proportions of women, of different age groups, or the distribution of the levels of education) and then carrying out a set of statistical tests of the absence of significant differences between these subgroups (the procedure is known as a balancing test). A good practice is to condition the statistical analysis on any observable characteristic whose distribution varies across subgroups in order to take into account possible interactions between this characteristic and variations over time.

In order to check the robustness of the results, it is possible to use so-called placebo groups to replicate the analysis on a group of observations that has not been exposed to the intervention being evaluated. A first way to do this is to use a fake treatment group, which can be the same treatment group but observed at least two dates prior to the implementation of the public policy being evaluated, or a third group that is assumed to be unaffected by the policy being implemented. The robustness of the analysis is strengthened if this procedure leads to the conclusion that there is no effect. A second practice is to use another control group, whose observable characteristics are similar to those of the control group. In this case, the estimate of the average treatment effect should be approximately equal to that obtained with the original control group.

While the use of longitudinal data improves the quality of the comparisons that are made, it leads to working with observations that are potentially correlated over time. This drawback has long been neglected in the application of the difference-in-differences method, leading generally to an overestimation of the statistical significance of the estimated treatment effect. It is therefore crucial to take into account the correlation structure of the data in the statistical analysis (see Bertrand et alii, 2004).

V. What are the strengths and limitations of this method compared to others?

The difference-in-differences method is a quasi-experimental method, in the sense that it is primarily used to study changes that occur exogenously and in ways that are not directly related to the evaluation goals, but which produce observations that approximate an experimental situation. Like all quasi-experimental methods, the effects estimated with this method correspond to the effects of the policy on the sub-population that has benefited from this policy (in terms of the causal evaluation model of public policy, it measures the average treatment effect on the treated, hereafter ATT). Since the intervention has been deliberately targeted at particular categories of the population (who are particularly concerned by the intervention implemented, or who are particularly in need of it), this approach does not allow us to measure the average treatment effect (hereafter, ATE), i.e., the effect that it would produce if it were generalised to the whole population, or even the variations in the effect across different treated individuals. Athey and Imbens (2006) propose an alternative approach to the difference-in-differences method which provides an estimate of the entire counterfactual distribution of the outcome variable, and which produces a more refined measurement of variations in the effect of the intervention across different groups of beneficiaries.

Nevertheless, this method measures an average effect in a larger sub-population than most existing quasi-experimental methods. As such, it differs in particular from the regression discontinuity design (see the dedicated separate sheet) and the local average treatment effect (LATE) approach, which both only allow to estimate average treatment effects for some specific sub-populations. This is the case for the subgroup of people (known as compliers) whose access to treatment is solely due to their proximity to an exogenously fixed threshold (e.g., an age or income threshold) in the first case, and those who benefit from it because of an instrumental variable in the second case.

Some bibliographical references to go further

Athey, Susan. and Imbens, Guido. W. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica*, 74(2): 431–97. https://doi.org/10.1111/j.1468-0262.2006.00668.x

Bertrand, Marianne. and Duflo, Esther. and Mullainathan Sendhil. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *Quarterly Journal of Economics*, 119(1): 249–275. https://doi.org/10.1162/003355304772839588

Card, David. and Krueger, Alan B. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review*, 84(4): 772-793. https://www.jstor.org/stable/2118030

De Chaisemartin, Clément. and D'Haultfoeuille, Xavier. 2022. "Difference-in-Differences Estimators of Intertemporal Treatment Effects." NBER Working Paper No. 29873. DOI 10.3386/w29873.

Fougère, Denis. and Kamionka, Thierry. and Prieto, Ana. 2010. "L'efficacité des mesures d'accompagnement sur le retour à l'emploi." *Revue Economique*, 61(3): 599–612. http://dx.doi.org/10.3917/reco.613.0599

Roth, Jonathan, and Sant'Anna, Pedo H. C., and Bilinski Alyssa, and Poe John. 2022. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature", arXiv:2201.01194, https://doi.org/10.48550/arXiv.2201.01194

Resources to implement this method with Stata and R softwares

Cunningham, Scott. 2021. Causal Inference: The Mixtape. Yale University Press: New Haven and London. Available in free access on the website https://mixtape.scunning.com/index.html

Huntington-Klein, Nick. 2022. The Effect: An Introduction to Research Design and Causality, Chapitre 18. Chapman and Hall/CRC Press: Boca Raton, Florida. Available in free access on the website https://theeffectbook.net/ch-DifferenceinDifference.html

3. The Regression Discontinuity Design

DENIS FOUGÈRE AND NICOLAS JACQUEMET

Abstract

The regression discontinuity design is a quasi-experimental quantitative method that assesses the impact of an intervention by comparing observations that are close to an eligibility threshold fixed by the authorities in charge of the policy under study. The existence of such a threshold (for instance, becoming eligible to a policy at a certain age or below a certain income level) generates a treatment group and a control group, in a manner similar to an experimental approach.

Keywords: Quantitative methods, quasi-experimental methods, eligibility threshold, forcing variable, sharp vs fuzzy regression discontinuity design, optimal bandwidth, monotonicity, compliers

I. How is this method useful for policy evaluation?

When one wishes to perform a quantitative evaluation of the effects of a public policy, the main difficulty consists in finding a comparison group (called a control group) whose situation can serve as a reference (i.e., as a counterfactual; see the sheet devoted to the difference-in-differences method) for the beneficiaries of the intervention (the so-called treatment group). The randomised experiment, in which beneficiaries and non-beneficiaries are randomly selected from a given eligible population, is the reference framework for defining a valid control group: by construction, if a large enough sample is available, the distribution of relevant characteristics in the control group (gender, age, education level, etc.) is the same as in the treatment group.

Quasi-experimental methods aim to compensate for the lack of randomised controlled experiments by relying on variations that occur exogenously (usually due to a decision of local or national authorities) and produce observations that approximate an experimental situation. Matching or difference-in-differences estimation methods exploit cases in which the implementation of a public policy produces two groups whose comparison allows us, under certain conditions, to measure its average effect. On the other hand, the regression discontinuity design exploits the existence of an eligibility threshold to conduct a statistical evaluation which is the equivalent of a local randomised experiment in the neighbourhood of the threshold.

II. What does this method consist of?

When the access to a public intervention or policy is conditioned by a threshold set by the authorities in charge of that policy, the intervention produces mechanically two groups, of which only one benefits from the intervention. But these groups are not directly comparable since they differ by construction because of the value of the variable defining the threshold (sometimes called the cutoff). This threshold can be an age condition (for instance, the statutory retirement age), a firm size condition (for instance, a tax reduction policy for firms with less than 20 employees) or a level of resources giving access to a grant scholarship or a tax credit. As these examples show, the assumption that the variable to which the threshold applies (e.g., the age, the firm size), commonly referred to as the forcing variable, would not influence the outcome variable of the intervention, is generally not credible. Retirement goes hand in hand with an increase in age, which in itself has many consequences on health status, consumption habits, social life, etc. Large firms operate within industries that are generally distinct from those in which SMEs operate, and their structure and activity are often very different. Income level obviously has a major impact on many household decisions. In these circumstances, the two groups thus formed do not allow for an evaluation of the effect of the intervention by directly comparing the value of the outcome variable between beneficiaries and non-beneficiaries.

On the other hand, the application of an eligibility threshold produces a sudden discontinuity in the distribution of observations near the threshold: for example, observations with a forcing variable whose value is just below the threshold could benefit from the intervention while their neighbours with a forcing variable whose value is just above the threshold could not. The regression discontinuity design exploits this property by assuming that small variations in the forcing variable around the threshold are the result of pure randomness, similar to a coin toss, which determines the access to the intervention of otherwise identical observations. Near the threshold, the assignment of a person or a firm to the treatment group is thus similar to what happens in a randomised experiment. Under this assumption, when observations are ranked in ascending order according to the value of the forcing variable, any discrepancy in the average value of the outcome variable once the threshold is crossed can be interpreted as a measure of the effect of the intervention.

In its simplest form, the regression discontinuity approach therefore measures the effect of a policy by comparing the average value of the outcome variable in the group of people eligible to the intervention, for example those with an income or an age just below the eligibility threshold, with the average value of that variable in the comparable control group, made up of people with an income or an age just above the threshold. The underlying assumption is that among people with otherwise similar characteristics in terms of qualification, education, or gender, those just below and above the threshold are potentially identical.

This implementation of the method therefore requires defining the interval (called the bandwidth) within which observations are kept for the analysis. This bandwidth choice is based on a trade-off between the quality of the statistical analysis permitted by a larger sample size and the weakening of the hypothesis of similarity that results from a wider interval. Imbens and Kalyanaraman (2012) propose a method to choose the magnitude of the optimal bandwidth.

The regression discontinuity design is said to be sharp when the assignment to the group eligible to r the intervention is mandatory and strictly triggered by the value of the forcing variable. If eligibility is based, for example, on an age criterion, and applied by an authority that has access to an exhaustive census of the population, then the probability of benefiting from the intervention is equal to 1 when the age condition is met; and this probability is equal to 0 otherwise, so that assignment according to the threshold is a certain event. Let us take the example of a training programme for jobseekers aged 25 or over. The principle is then to compare the average value of the outcome variable (e.g., the hiring wage at the time of return to work) for jobseekers who are just above the age threshold, e.g., aged 25 or 26, with the average hiring wage for those aged 23 or 24, who could not benefit from this programme.

The fuzzy regression discontinuity design corresponds in contrast to situations where this threshold is less binding, so that there are observations on both sides of the threshold that are, or are not, beneficiaries of the intervention. In the example of the training programme for jobseekers aged 25 and over introduced above, let us assume that, in a given locality, this training can only be provided to 100 people aged 25 or 26 due to budgetary constraints, and that this training is not compulsory, so that only 80 of these 100 eligible people (i. e., 80%) actually agree to participate in the programme. The local employment agency then offers the remaining 20 places to 100 unemployed people aged 23 or 24; among these 100 persons, only 10 (10%) agree to participate in the programme. Rather than a sudden change in the treatment status, the notion of discontinuity here refers to the ‘jump’ in the probability of benefiting from the intervention when the eligibility threshold (age 25) is crossed. The objective is then to measure the average effect of the intervention by restricting the approach to the variation in the outcome variable that results from this “jump” in the probability of benefiting from the intervention. This procedure is based on a strong assumption, called the monotonicity assumption: this assumption implies that among the unemployed who do not participate in the training programme because their age is below 25, there is a subgroup of individuals who would accept to participate if their age were 25 (or above). In technical terms, these individuals are called the compliers. By construction, the fuzzy regression discontinuity design allows us to estimate the average effect of the intervention for this subgroup only. In addition to the fact that this subgroup can sometimes be very small, it excludes two important groups, namely individuals who are always willing to participate in the programme regardless of the value of the forcing variable (the always takers), and those who do not wish to participate under any circumstances (the never takers).

III. Two examples of the use of this method in education

Variations in housing prices across neighbourhoods reflect the willingness of households to pay for the set of services and amenities (i.e., the benefits delivered by the living environment) to which a house or an apartment gives access. One such amenity is of course the quality of the local school to which children of the residents have access. Attempts to estimate the effect of school quality on housing prices are often unconvincing, as the best schools tend to be located in the best neighbourhoods. Valuations that do not take sufficient account of neighbourhood characteristics therefore tend to overestimate the value of schools located in such areas. To overcome this difficulty, Black (1999) uses a particularly original application of the sharp regression discontinuity design, based on a threshold corresponding to the contours of the Boston school map. The study estimates the value that parents place on the quality of the local public school by comparing prices of dwellings that are located on both sides of the geographic boundaries of a school district. The fact that the average scores of students in schools in different but neighbouring sectors sometimes vary greatly, while the characteristics of dwellings on either side of school divisions change relatively little, allows to identify the relationship between educational outcomes (interpreted as the school quality) and housing prices thanks to the spatial discontinuities. The estimates suggest that a one-point increase in the average school test score leads to a 1.3% to 1.6% increase in the housing prices near the geographical limit of a school district.

The study by Matsudaira (2008) is an example of the implementation of a fuzzy regression discontinuity design, also applied to educational attainment. The study uses an administrative data set from a large school district in the United States. In this district, students advance to the next grade if their grades are above predefined thresholds. Students with grades below these thresholds are required to attend a four- to six-week summer school to avoid repeating a grade. Since the observed characteristics of students near the thresholds are almost identical, the differences in subsequent academic achievement between students just below and just above the thresholds can be attributed to the causal impact of the summer school. The sample is restricted to students enrolled in the third grade of elementary school (at the age of about eight years) and the fifth grade (at the age of about 10 years). Student scores were recorded for math and reading tests in the spring of 2001 and 2002, giving rise to a sample of 338,608 students. However, the regression discontinuity design is fuzzy: the relationship between the end-of-year test scores and summer school attendance is not deterministic. Some students whose scores were below the thresholds did not attend the summer school, while some students whose scores were above the thresholds did. For instance, only 38% of the students in third and fifth grades whose math grades were below the prerequisites at the end of the 2000-2001 school year were enrolled in the 2001 summer school. Estimates from the fuzzy regression discontinuity design method suggest that the scores of 3rd grade compliers increased by 12.8% the following year, while those of 5th grade compliers attending the summer school increased by 24.1%.

IV. What are the criteria for judging the quality of the mobilisation of this method?

For the regression discontinuity technique to mimic a local randomised experiment, it is important that the forcing variable is an exogenous covariate that is beyond the control of the population involved in the intervention. If people or firms can manipulate the value of the threshold, then assignment to the treatment group becomes a choice variable. The classic example is that of a public policy that offers employment subsidies to firms with less than 20 employees. The natural reaction of some firms whose employment level is approaching the threshold is to recruit more temporary workers, in order to increase the firm's labour force without this increase appearing in the tax returns to which they are subject, so as to continue to benefit from employment subsidies. To detect such a manipulation of the threshold, McCrary (2008) proposes a simple statistical test, based on an aggregate reasoning. Firms that actually employ more than 20 employees (e.g., 21 or 22 employees), but whose reported size is less than 20 employees (i.e., 19 or 20), will artificially increase the proportion of firms with less than 20 employees and simultaneously decrease the proportion of firms with 21 or 22 employees. The existence of manipulations in response to the eligibility threshold therefore has a direct consequence on the distribution of firm sizes, which can be checked using a histogram. In theory, this histogram should not show a discontinuity just before and just after the threshold of 20 employees. However, if this were the case, and this can be tested statistically, then the manipulative behavior of some firms could be suspected.

To avoid narrowing the bandwidth around the threshold too much, it is common to add explanatory variables other than the forcing variable, providing control over the variations in the outcome variable that are due to observed covariates. For instance, individual income tends to increase with age, so that widening the bandwidth around the age threshold leads to additional observations for which the outcome variable changes with the level of the individual income. Taking this income effect into account in the statistical analysis undermines such confounding differences between groups. It is important to check that the distributions of covariates other than the forcing variable do not exhibit a discontinuity in the neighbourhood of the threshold considered. If this is the case, it means that the intervention to be evaluated has some effects not only on the outcome variable but also on some of these covariates. Incorporating these covariates into the statistical analysis generally generates a biased estimate of the average effect of the intervention on the outcome variable, since discontinuities in the distributions of these covariates are themselves explained by the implemented intervention.

V. What are the strengths and limitations of this method compared to others?

The main difficulty raised by most quasi-experimental methods is that they are based on strong assumptions, which are often questioned, such as the comparability of the control and treatment groups before the implementation of the intervention. When one wishes to apply the difference-in-differences method, this is for instance the reason why it is necessary to check that the outcome variable has previously followed the same evolution in the two groups and that their observable characteristics are similar. The difficulty is the same when one wishes to use a matching method: it requires to find observations serving as a control group which have similar observable characteristics to those of the treatment group, and which also have a non-zero probability of being eligible to the intervention being evaluated. The regression discontinuity design avoids this difficulty because it is based on a principle of quasi-random assignment for the subpopulation which is close to the exogenous threshold. As in a randomised controlled experiment, the comparability of the two groups is based on a statistical argument: if the sample size is sufficiently large, the distribution of all covariates that are relevant to significantly explain variations in the outcome variable is similar in the two groups.

This assimilation of the regression discontinuity design to a randomised experiment is all the more convincing as the interval within which it is supposed to be applied is narrow, which leads to restricting the measured effect to a very particular subpopulation, characterised by the proximity of its forcing variable to the threshold. The measure provided by this local quasi-randomised experiment is therefore specific to this sub-population. Since the effect of the intervention varies greatly across different sub-groups, the estimated average treatment effect is local and only valid in the neighbourhood of the exogenous threshold (this estimate corresponds to a local average treatment effect, or LATE). An extrapolation of the results obtained for observations far from the threshold (which would define the external validity of the LATE) is of little relevance. This limitation of the method is further amplified in the case of a fuzzy regression discontinuity design, where the local effect is specific to the compliers only. This lack of external validity is problematic since thresholds are often set according to the expected benefit of the intervention for the eligible group. For example, a training programme for long-term unemployed aims to counteract the effects of human capital losses due to increased unemployment spells. Part of the rationale for setting a threshold between long- and short-term unemployment spells is that this human capital loss is minimal when spells are sufficiently short. Estimating the effect of such a programme based on a regression discontinuity design thus amounts to focusing on the specific sub-population (unemployed workers experiencing relatively shorter unemployment spells) for which the program is likely to be the least effective.

The interested reader will find excellent surveys about the regression discontinuity design, for instance, in the article by Lee and Lemieux (2010), and in the textbook by Cattaneo, Idrobo and Titiunik (2019).

Some bibliographical references to go further

Black, Sandra E. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education", *Quarterly Journal of Economics*, 114(2): 577-99. https://doi.org/10.1162/003355399556070

Cattaneo, Matias D.. and Idrobo, Nicolás. and Titiunik, Rocío. 2019. *A Practical Introduction to Regression Discontinuity Designs: Foundations. Elements in Quantitative and Computational Methods for the Social Sciences*. Cambridge University Press. https://doi.org/10.1017/9781108684606

Imbens, Guido. and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator", *Review of Economic Studies*, 79 (3): 933-59. https://doi.org/10.1093/restud/rdr043

Lee, David S.. and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature*, 48(2): 281-355. https://doi.org/10.1257/jel.48.2.281

Matsudaira, Jordan D. 2008. "Mandatory Summer School and Student Achievement." *Journal of Econometrics*, 142(2): 829-50. https://doi.org/10.1016/j.jeconom.2007.05.015

McCrary, Justin. 2008. “Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test”. *Journal of Econometrics*, 142(2): 698-714. https://doi.org/10.1016/j.jeconom.2007.05.005

Resources to implement this method with Stata and R software

Cunningham, Scott. 2021. *Causal Inference: The Mixtape*. Yale University Press: New Haven and London. Available in free access on the website https://mixtape.scunning.com/index.html

Huntington-Klein, Nick. 2022. *The Effect: An Introduction to Research Design and Causality*, Chapter 20. Chapman and Hall/CRC Press: Boca Raton, Florida. Available in free access on the website https://theeffectbook.net/ch-RegressionDiscontinuity.html

4. Matching methods

PAULINE GIVORD

Abstract

Matching is a quantitative method for ex-post evaluation in which, in the absence of direct experimentation, a counterfactual situation is reconstructed by comparing the situations of beneficiaries of an intervention with those of non-beneficiaries with very similar characteristics. This method is particularly useful for evaluating the impact of a programme on a whole population, when sufficiently precise data exist to compare beneficiaries and non-beneficiaries.

Keywords: Quantitative methods, ex post evaluation, causal effect, propensity score, common support

I. What does this method consist of?

Matching methods are among the main quantitative methods for ex-post evaluation, aiming to measure the effect of a public policy tool or programme (e.g. a training programme for jobseekers, or localised aids in certain territories) on the situation of the beneficiaries. As with most quantitative evaluation methods, the aim is to estimate the causal effect of the intervention on the situation of the beneficiaries (for example, a return to employment after training, or the economic activity of the targeted territory). The objective of matching methods is to estimate this causal effect by comparing the situation of beneficiaries of the programme with that of people who have not benefited from it, but whose characteristics are so similar that it would have been possible for them to benefit from it. The observation of these non-beneficiaries is supposed to give an idea of the “counterfactual” situation, that which the beneficiaries would have experienced in the absence of the programme.

The challenge here is to reduce the selection effects that can occur when one wishes to estimate the effect of an intervention. In general, the beneficiaries have not been designated by chance, and they have specific characteristics that can explain by themselves a more or less favourable evolution, even in the absence of the programme being evaluated. For example, the evaluation of a training programme aimed at people furthest from employment cannot be done simply by comparing the chances of return to employment of beneficiaries before and after the training, at the risk of underestimating the effect of the programme for the most disadvantaged public. Nor is it possible to compare the return-to-work rates of trainees with those of the non-trained population as a whole: the latter are too different for their employment situation to be a likely reflection of what the trainees would have experienced in the absence of training.

The principle of matching methods is to restrict the comparison of trainees to comparable non-trainees. Specifically, each beneficiary of the programme being evaluated is matched with one or more “twin” non-beneficiaries, in the sense that they have very similar individual characteristics in all dimensions that may influence both benefiting from the programme and their subsequent situation. In the example of the estimation of the training course impact on the chances of returning to employment, we compare for each trainee the chances of having found a job for instance during the year following the entry into training with the same chances of persons identical or at least closest to this trainee at the date of the entry into training in the dimensions considered important for the return to employment. The average effect of training for trainees is obtained by averaging all these comparisons for each beneficiary.

In principle, one wishes to match on as many dimensions as possible, to avoid the risk of missing an important characteristic, whose non-inclusion in the comparisons would lead to incorrect estimates of the causal effect. However, the more dimensions one wishes to match on, the more difficult it will be to find exactly identical non-beneficiaries for each beneficiary in all these dimensions. In the example of the evaluation of a training programme, it may therefore be relevant to match on age, level of education, length of time unemployed and past experience (e.g. number of previous unemployment episodes), past work experience (e.g. job qualification), type of job sought, possible mobility, which are all variables that may influence both the choice of training and the return to employment (independently of this training). Exact matching on each of these dimensions means that for each vocational trainee one must find a person with exactly the same characteristics in all of these dimensions: the higher the number of variables, the less likely it is to find a perfect "twin", especially if the number of observations is low.

A frequently used response to this limitation is to match not on all these characteristics, but on a summary of them provided by the "propensity score". This corresponds to the probability of being a beneficiary, conditional on the dimensions selected as important for the matching. This means that the estimation is done in two steps. First, the propensity score is estimated, i.e. how the different dimensions predict entry into training, which makes it possible to define an a priori probability of being a beneficiary for each observation, depending on its characteristics. In our example, the probability of entering training will be estimated as a function of age, diploma, etc.... This estimate will be used to calculate for each person, whether or not a trainee, his/her "propensity" to enter training, i.e. the probability predicted as a function of these individual characteristics. The values of the propensity score are generally strictly between zero and one (unless a particular exclusion condition is met, it is rare that a person has no chance of entering training, and conversely it is unlikely that any of the characteristics will automatically result in entry into training). Their distributions overlap between beneficiaries and non-beneficiaries. While those who have a priori a high probability of entering training are more numerous among those who actually enter training, some do not and can be used for comparison. Conversely, some people with an a priori low propensity to enter training may nevertheless choose to train – and it will also be possible to compare them with people who did not train, also having a low propensity to do so. It can be shown that when matching on propensity scores, the important characteristics are on average identical between the beneficiary and non-beneficiary groups.

Whether the matching is done on a single dimension (the propensity score), or on several of them, it is difficult to have exactly identical values for the matching: it is therefore done by using the "closest neighbours" of the beneficiaries, i.e. the non-beneficiaries who are closest to the beneficiary according to the dimensions retained (or according to the propensity score). There are then several variants, notably on the number of neighbours retained (it may be preferable to retain several to avoid comparing by misfortune with a non-beneficiary whose behaviour would be atypical) and on the maximum distance allowed between the beneficiary and the comparisons (neighbours who are too far away being by definition less suitable for comparison).

Whichever matching method is used, it is necessary to have individual data to describe the situation and individual characteristics in detail, and a large number of observations to be more confident of finding close neighbours.

II. How is this method useful for policy evaluation?

Matching methods make it possible to estimate ex post the effect of a programme on beneficiaries, on a set of objectively measurable dimensions. For example, they make it possible to answer questions such as: do jobseekers who have chosen to train (at the risk of interrupting a job search) have a higher probability of returning to sustainable employment than jobseekers who do not train? Does this training allow them to expect a higher level of pay? Which jobseekers benefit most from training?

The goal, therefore, is to measure the differences between the situation that was actually experienced by the beneficiaries of a programme and a "counterfactual" situation that would have prevailed in the absence of this programme. In general, these methods are suitable for evaluating the general impact of a programme (compared to a situation where this programme would not exist), but are less suitable for measuring the effect of the different modalities of this programme (in our example, several more or less intensive programmes for training jobseekers).

III. Two examples of application: active employment policies and territorial tax exemptions

Matching methods are very commonly used to evaluate the effects of so-called "active" employment measures (training, job search assistance, etc.), particularly since the methodological study by Heckman, Ichimura and Todd (1997). This method has been used, for example, to study an active employment policy in Sweden (Sianesi, 2004), training programmes in Germany or, more recently, training for job seekers in France (Chabaud et al., 2022).

Another example is the evaluation of the effects of the Zones Franches Urbaines (ZFU), a public policy tool designed to encourage the establishment of companies in disadvantaged urban areas, similar to the Enterprises Zones set up in the United States in the 1980s. Givord, Rathelot and Sillard (2013) look at the effects of these exemptions on the establishment of businesses and the evolution of employment in the targeted neighbourhoods, compared with other neighbourhoods that were initially very close (see also Malgouyres and Py, 2016). These studies suggest a positive effect of the zones on employment and economic activity, but at the expense of the immediately neighbouring zones. Another study also suggested that the effects were not persistent beyond the duration of the exemptions (Givord et al., 2022).

IV. What are the criteria for judging the quality of the mobilisation of this method?

The validity of matching methods depends crucially on how well they can be corrected for selection effects, and therefore on the information available to compare beneficiaries and non-beneficiaries. There must be some assurance that the selection process in the intervention is not based on variables that are not available in the data (e.g. the results of a motivational interview used to enter a training programme, which would aim to measure dimensions that are not very objective and therefore not available to an outside eye). Having individual information on the variable of interest in the past (e.g. the professional trajectory prior to entering the training programme) is generally considered indispensable to avoid capturing selection effects: matching methods are in this case combined with “difference-in-differences” (see separate chapter on difference-in-differences).

Secondly, the method requires the possibility of matching all beneficiaries with non-beneficiaries (this is called “common support”). This last condition means in particular that there is a certain amount of randomness in the fact of benefiting from the programme: if the programme is totally deterministic in terms of observable characteristics (for example, a programme systematically offered to young people without diplomas, which would exclude people above a certain age or income threshold), it will not be possible to match the beneficiaries to non-beneficiaries on these dimensions.

Finally, matching methods provide a statistical estimate, and therefore as such do not allow the "true" effect value to be measured with complete certainty, but only an approximation whose precision, i.e. the degree of confidence with which this estimate can be used, can be quantified. This precision can be measured by means of the standard deviation (the smaller the standard deviation, the greater the confidence that the "true" effect is close to the estimated value) or by means of a confidence interval, which corresponds to the interval of values within which the true effect is found with a given probability: for example, the interval of values within which the true value of the effect is found with a probability of 95% (the smaller the confidence interval, the greater the precision of the estimated value). This measure of precision is used, for example, to check that the effect of the intervention being evaluated is "significant" or "significantly different from zero", i.e. it can be said with some confidence that the programme does indeed have a strictly positive or strictly negative effect.

V. What are the strengths and limitations of this method compared to others?

One of the strengths of matching methods is that they can estimate effects in the "general population", i.e. on the whole population (provided that there are enough observations to be able to find comparisons and that the assignment to the programme is sufficiently random to allow for the availability of beneficiaries on the whole). This can be an advantage over most ex-post quantitative evaluation methods, which only allow an unbiased estimate of a causal effect on 'marginal' populations: for example, people around an eligibility threshold for discontinuity regressions (see separate chapter on discontinuity regressions), or people who are sensitive to the signal given by an instrument.

On the other hand, matching methods may not be sufficient to correct for selection bias. Estimates are very sensitive to the choice of variables used for matching, and it is generally difficult to trust estimators in the absence of past individual measurements of the variable of interest.

Some bibliographical references to go further

Biewen, Martin. and Fitzenberger, Bernd. and Osikominu, Aderonke. and Paul, Marie. 2014. “The Effectiveness of Public-Sponsored Training Revisited: The Importance of Data and Methodological Choices.” *Journal of Labor Economics*, 32: 837-897.

Fitzenberger, Bernd. and Völter, Robert. 2007. “Long-run effects of training programs for the unemployed in East Germany.” *Labour Economics*, 14(4): 730-755.

Givord, Pauline. and Rathelot, Roland. and Sillard, Patrick. 2013. “Place-based tax exemptions and displacement effects: An evaluation of the Zones Franches Urbaines program.” *Regional Science and Urban Economics*, 43(1): 151-163.

Givord, Pauline. and Quantin, Simon. and Trevien, Corentin. 2018. “A long-term evaluation of the first generation of French urban enterprise zones,” Journal of Urban Economics, n°105(C): 149-161.

Heckman, James. and Hidehiko, Ichimura. and Petra, Todd. 1997. “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme.” *Review of Economic Studies*, 64(4): 605-654.

Lechner, Martin. 2002. “Program Heterogeneity And Propensity Score Matching: An Application To The Evaluation of Active Labor Market Policies.” *The Review of Economics and Statistics*, vol. 84, n°2: 205-220.

Malgouyres, Clément. and Py, Loriane. 2016. "Les dispositifs d'exonérations géographiquement ciblées bénéficient-ils aux résidents de ces zones? État des lieux de la littérature américaine et française." *Revue économique*, 67: 581-614.

Sianesi, Barbara. 2004. "An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s." *Review of Economics and Statistics*, 86: 133-155.

5. Microsimulation

MATHIAS ANDRÉ

Abstract

Microsimulation is a quantitative method for estimating the expected impact of an intervention (e.g. the modification of a tax rate) and describing its effects (winners, losers, budgetary envelope, effect on inequality indicators). It is based on taking into account the characteristics of the target population (e.g. age, income, etc.) and modelling public policy effects concerning this population. Because of the diversity of situations that it makes it possible to integrate, this technique provides more precise and complete results than estimates based on average or aggregate reasoning of the representative individual type. Its development has been encouraged by the improvement in computing power and the increase in statistical information (surveys or administrative data). It is an essential tool for ex ante evaluation of the impact of public policies and can also be used for ex post evaluation.

Keywords: Quantitative methods, modeling, static/dynamic microsimulation, demography, socio-fiscal policies, pensions

I. What does this method consist of?

Microsimulation is a method developed by research institutions or government agencies by modelling economic agents, mainly individuals or companies, for the purpose of evaluating public policies. It was developed to address the limitations of macroeconomic analysis, which relies on a single agent representative of the economy (Orcutt, 1957).

The general principle is based on the representation of the economy as a collection of elementary units (e.g. individuals) with specific characteristics (e.g. age, marital status, family size, income). As opposed to modelling on the basis of an average individual, this approach allows for the measurement of variation among individuals and the modeling of changes in their situations.. For example, dynamic methods that model simulate the birth, aging, and death of individuals, which may be useful for assessing pension systems. Other models can study the effect of changes in taxes or social benefits on household disposable income. This is the purpose of socio-fiscal microsimulation. In other words, the microsimulator will use a variety of observed data and simulate changes in situations, unit by unit, either through deterministic relationships (if family benefits increase, households with children see their income increase) or probabilistic relationships (each year, children are born with a certain probability). Microsimulation is said to be dynamic if it incorporates phenomena such as demographic changes (ageing, fertility, mortality) or adjustments in different markets (employment, trade, etc.); it is otherwise said to be static.

Microsimulation relies on computer software to model the range of socio-economic situations. It uses dedicated statistical softwares (e.g. R or Python) to process both the individual data and to write the model itself. The model simulates situations based on observed variables and relies in particular on programming based on the legislation and how the criteria it sets affect individuals' variables (such as the age of retirement or the calculation of income tax). The current legislation is used as a baseline and the proposed reforms are modeled to evaluate their impact. In concrete terms, a microsimulation model is based on three building blocks: the subject matter, the data used and a "calculator", i.e. the core of the code describing the changes or effects of the socio-economic phenomena studied. In the typical framework of static socio-fiscal microsimulation, this allows, for example, for the simulation of the direct effects of policy changes in the form of aggregate total effects (the budget of a tax change, for example), the direct effects on households (such as the number of winners or losers as well as the average gains and losses) and the redistributive effects (as measured by changes in inequality indicators or the description of the populations concerned). More advanced models consider the behaviour of agents in response to simulated policies.

The initial principle of microsimulation methods dates back to the 1960s (Orcutt, 1960), but their development in the 1980s was mainly based on the widespread use of representative survey databases by statistical institutes and greater use of administrative data in quantitative evaluation methods in the 1990s. With the improvement in the power of computer calculations and access to a large variability of individual information, academic or administrative work has become more widespread since the 2000s. In the United Kingdom and the United States, and to a lesser extent in France, these tools have become established in the public debate, particularly in the context of the evaluation of pension systems or social and fiscal proposals, during budgetary debates for example.

Microsimulation techniques are established and recognised methods and deal with a wide variety of subjects: taxation and socio-fiscal transfers, pension systems, health expenditure and the health insurance system, environmental policies, employment and professional trajectories, educational choices, demography, dependency, etc.

II. How is this method useful for policy evaluation?

Microsimulation is commonly used for the ex-ante evaluation of socio-fiscal, educational or environmental reforms. Its results are used in impact assessments of laws or studies published by microsimulation teams. Microsimulation is based on the calculation, the “simulation”, of fictitious situations. The core of the evaluation enabled by microsimulation is based on the comparison of counterfactual situations in the form of ‘with or without reform’. The simulation of new legislation or of developments modified by socio-demographic changes makes it possible to compare two situations. To evaluate a tax change, for example, the model compares individual situations with and without the reform. By difference, it is then possible to estimate the gains and losses and to write down the totals (costs or revenues) and the associated distributions. The microsimulation estimates the population concerned: the better-off, pensioners, single-parent families, etc. The prospective scenarios can be numerous and thus provide both a decision-making aid for the legislator and an ex ante evaluation of public policies.

In France, ministerial departments use models to construct government policies. The Treasury Department uses the Saphir model for monetary social benefits and direct taxes such as income tax. The Direction de la Recherche, des Études, de l'Évaluation et des Statistiques (Drees), the statistical department of the Ministry of Health and Social Affairs, is developing several models, such as Trajectoire for pensions, OMAR for health expenditure, Autonomix for dependency or INES (co-developed with INSEE and Cnaf and freely available since 2016) for socio-fiscal policies. The Ministry of the Environment's Prometheus model, for example, studies the heating and transport expenditure of French households.

It is also a long-standing tradition of the economists of the Paris-Jourdan campus with the Sysiff model developed in the 1970s-1980s at the Delta laboratory (a predecessor of the Paris School of Economics – PSE), the contribution to the EUROMOD model used by Eurostat and various research laboratories in Europe, the tax simulator of Camille Landais, Thomas Piketty and Emmanuel Saez on which the general public book Landais, Piketty, Saez (2011) is based. Currently, the Institute for Public Policy (IPP) is developing the TaxIPP (social tax) or PensIPP (pensions) models. In the United Kingdom, for example, the budget is evaluated by an institute (Institute for Fiscal Studies) prior to debates in Parliament on the basis of microsimulation models. In the United States, the TaxSim model is developed by the National Bureau of Economic Research (NBER) and is accessible to researchers.

However, ex post uses of microsimulation are also possible. The principle is identical to the ex ante methods but they apply to public policies that are actually implemented. The advantage of this use is that it no longer requires assumptions to be made about the state of the economy; the simulations are then applied to observed data over the study period and the counterfactual situation is compared with the actual situation. Ex post microsimulation is the subject of studies in the socio-fiscal field, which is described in the following section.

III. Two examples of the use of this method: socio-fiscal policy and pension policy

The Ines model, which was co-developed by INSEE and DREES at the time, played an active role in the creation of the activity allowance (prime d'activité, individual allowance given to low wage workers) in 2016, as well as the active solidarity income (Revenu de solidarité active or RSA, allowance given lowest income households) in 2009. The first step was to design this benefit on the basis of the legislator's objectives. Numerous scenarios were calculated. Once the principle of the benefit and the target budget had been set, microsimulation was used to determine the amount of the scale, in this context the individual bonus, corresponding to the criteria. It is with these means that the impact study of the law is then drafted. Microsimulation has thus made it possible to construct the scale of a social policy.

Two cases of widespread use of microsimulation for policy evaluation are the study of the pension system (see Cheloudko and Martin, 2020) and that of socio-fiscal reforms (Fredon and Sicsic, 2020). On this subject, the Ines model team publishes an annual assessment of the past year's socio-fiscal reforms and draws up a redistributive balance sheet based on a precisely defined methodology (André et al., 2015). The most recent study (Buresi et al., 2022) states that "the new social and fiscal measures introduced in 2020 and 2021, once fully implemented, increase the standard of living of people living in metropolitan France by 1.1% compared to a situation without their implementation. The average gain is 280 euros per year per person: 240 euros for the 2020 measures and 40 euros for those of 2021. This increase mainly benefits the wealthier half of the population, which is particularly affected by the main permanent reforms implemented."

In a similar exercise, the IPP and the OFCE published evaluations of reforms, ex ante in the context of the budget (Fabre et al., 2020) or sometimes ex post over a five-year period (Madec, Plane and Sampognaro, 2022). These analyses are often taken up in the public debate, whether in the media, with numerous press articles based on them, or in the context of parliamentary activity, with quotations in reports or in statements by political representatives. This is also the case for the reform of the taxation of wealth with the transformation of the solidarity tax on wealth (ISF) into a tax on real estate wealth (IFI), which has seen a debate on the population actually concerned by this reform and the amounts involved.

IV. What are the criteria for judging the quality of the mobilisation of this method?

The quality of a microsimulation method depends both on the quality of the underlying data and on the quality of the model used. The representativeness of the survey or the administrative databases ensures the external validity of the results, i.e. the capacity of the model to estimate the effects on the entire target population. The richness of the variables in the employment survey produced by INSEE, for example, allows representations to be made from different angles (activity status, diploma, etc.), whereas the fine granularity of the administrative data provides large samples in order to represent the results for specific populations (the wealthiest 1%, for example).

A systematic method of comparing model results with external sources guarantees the quality of the simulations. The Ines model is thus subject to an annual “validation” note. Each benefit and tax is compared to the real administrative aggregates. For income tax, for example, the number of taxable individuals, the total and the average amount are quality criteria for the simulations. Precise documentation and the availability of the source code in open format, i.e. accessible to all, are also a guarantee of transparency and therefore of the quality of a microsimulation model.

Finally, disparities may appear between the results of different models. The comparison of results, as well as the explanation of the differences, makes it possible to judge the advantages and disadvantages of the different models (André and Sicsic, 2020).

V. What are the strengths and limitations of this method compared to others?

The main strengths of microsimulation lie in the very reason for its creation: the models allow for the great diversity of individual situations. Writing legislation in an integrated way makes it possible to simulate ‘detailed effects of policies whose rules depend on a large number of individual characteristics, very often non-linear, for example because of threshold or ceiling effects’, such as housing benefits or income tax (Blanchet, 2020).

The main limitations are based on the exercise without equilibrium effects: the units in the models are assumed not to change their behaviour (especially in static models) or to interact other than through the limited assumptions of the model (demographic or pension choices in dynamic models). The assessments are thus described as “first round”, i.e. they do not take into account macroeconomic closure effects (such as labour market effects) or behavioural responses (such as savings or consumption adjustments). Nevertheless, the inclusion of non-use of certain social benefits is sometimes taken into account in evaluations and in static models and thus constitutes an integration of household behaviour in relation to social and fiscal policies.

Some studies aim to take these limitations into account and integrate behavioural effects following the estimations of microsimulation models (Paquier and Sicsic, 2021) or second-round effects (André and Biotteau, 2021).

Cited references

André, Mathias. and Biotteau, Anne-Lise. 2021. “Effets de moyen terme d’une hausse de TVA sur le niveau de vie et les inégalités: une approche par microsimulation”. *Économie et Statistique*, n°522-523.

André, Mathias. and Cazenave, Marie-Cécile. and Fontaine, Maëlle. and Fourcot, Juliette. and Sireyjol, Antoine. 2015. Effet des nouvelles mesures sociales et fiscales sur le niveau de vie des ménages : méthodologie de chiffrage avec le modèle de microsimulation Ines. Insee, Documents de travail, n°F1507.

André, Mathias. and Sicsic, Michaël. Évaluation des effets redistributifs des réformes socio-fiscales : comment s’y retrouver ?, blog de l’Insee, 2020. https://blog.insee.fr/evaluation-des-effets-redistributifs-des-reformes-socio-fiscales-comment-sy-retrouver/

Blanchet, Didier. 2020. Des modèles de microsimulation dans un institut de statistique : Pourquoi, comment, jusqu’où ?, C*ourrier des statistiques*, n°4.

Buresi, Gabriel. and Cornetet, Jules. and Cornuet, Flore. et Doan, Quynh-Chi. and Dufour, Camille. and Trémoulu, Raphaël. 2022. 'Les réformes sociofiscales de 2020 et 2021 augmentent le revenu disponible des ménages, en particulier pour la moitié la plus aisée'. France portrait social, Insee références.

Cheloudko, Pierre. and Martin, Henri. 2020. Une décennie de modélisation du système de retraite – La genèse du modèle de microsimulation TRAJECTOiRE, *Courrier des statistiques*, n°4.

Fabre, Brice. and Guillouzouic, Arthur. and Lallemand, Chloé. and Leroy, Claire. 2020. Budget 2020 : quels effets pour les ménages ?, Note IPP n°49.

Fredon, Simon. and Sicsic, Michaël. 2020. Ines, le modèle qui simule l'impact des politiques sociales et fiscales, *Courrier des statistiques*, n°4.

Landais, Camille. and Piketty, Thomas. and Saez, Emmanuel. 2011. *Pour une révolution fiscale*, La Découverte.

Madec, Pierre. and Plane, Mathieu. and Sampognaro, Raul. 2022. Une analyse macro et microéconomique du pouvoir d'achat des ménages en France : Bilan du quinquennat mis en perspective. OFCE Policy Brief, 104: 1-18.

Paquier, Félix. and Sicsic, Michaël. 2021. Effets des réformes 2018 de la fiscalité du capital des ménages sur les inégalités de niveau de vie en France : une évaluation par microsimulation, *Économie et Statistique*, n°530-531.

Some bibliographical references to go further

Journal issues dedicated to microsimulation:

Courrier des statistiques n°4, avril 2020;

Économie et statistiques (n°481-482, 2015) and Revue économique (vol. 67, 2016);

Économie et prévision (n°160-161, 2003).

General references providing an overview of the method:

Bessis, Franck. and Cotton, Paul. 2021. *La réforme, le chiffrage, son modèle et ses données*, Politix 2021/2 (n°134).

Legendre, François. 2019. L'émergence et la consolidation des méthodes de microsimulation en France. *Économie et Statistique*, n°510-511-512: 201-217.

Bourguignon, François. and Landais, Camille. 2022. Micro-simuler l'impact des politiques publiques sur les ménages : pourquoi, comment et lesquelles ?, *Les notes du conseil d'analyse économique*, n°74, septembre 2022.

O'Donoghue, Cathal. 2014. (ed), *Handbook of Microsimulation Modelling*, Emerald Publishing Ltd.

6. Experimentation in the Laboratory

LOU SAFRA

Abstract

Laboratory experimentation makes it possible to directly measure the attitudes and behaviour of individuals and to evaluate the causal effect of a variable on these attitudes and behaviour. To do this, individuals are put in a situation where they are asked to perform a certain number of tasks for which as many elements as possible are controlled (such as the duration of the task and the type of information given to participants). This approach can help to anticipate ex ante how individuals will respond to an intervention or can be used ex post to measure changes in behaviour following an intervention. It is particularly useful for uncovering non-conscious behavioural biases.

Keywords: Quantitative methods, within-/between-participant method, laboratory experimentation, causal effect, behaviours, attitudes, non-conscious behavioural bias, internal/external validity, automatic/non-automatic response

I. What does this method consist of?

In a simple way, in a laboratory experiment, participants perform a given task, designed to measure their behaviour. The first step in laboratory experimentation is therefore to establish an experimental protocol for measuring the individual's behaviour. Classically, these experiments rely on a computer task, which will make it possible to measure not only the participants' choices but also other data that may prove particularly informative, such as response time. These tasks can be aimed at measuring participants' preferences and perceptions as well as the way they learn or reason. Thus, laboratory experiments are particularly used by fields that are directly concerned with people's behaviours and perceptions, such as cognitive science and psychology, including social, developmental and political psychology, as well as economics and educational science. Most of these protocols are based on measuring participants' choices between different options or their evaluations of these options on a scale. For this purpose, different types of material (or stimuli) can be presented to the participants (images, texts, videos, sounds etc.). Thus, this method makes it possible to measure attitudes and behaviours directly, which can be particularly useful when it comes to behaviours or attitudes that participants tend not to report or of which they are not aware, even though these attitudes may have a significant influence on their behaviours, as is the case for implicit gender bias.

In addition to offering the possibility of directly measuring behaviour, laboratory experiments also make it possible to measure how behaviour can be influenced by a specific context. This is the core of the scientific experimental method: by comparing participants' behaviour in different conditions, one in which the factor of interest (the one whose influence is being studied) is present and one in which it is absent, it is possible to assess the causal link between this factor and the behaviour being studied. However, as these studies are conducted in laboratory settings, this factor of interest must be extracted from the real context to be studied experimentally. For example, when studying the acceptability of a new drug, its price, efficacy and side effects can be studied together or separately using fictitious choices in order to estimate their influence on the participants' perceptions. Thus, laboratory experiments require a thorough analysis of the factors that may affect the behaviour of interest. This notion of comparison extends beyond the choices themselves and can also be applied to different contexts or conditions. For example, comparing a condition in which participants have access to information on the percentage of female students in each secondary school stream with a condition in which this information is not given allows one to estimate the effect of this type of information on students' orientation choices.

These comparisons can be made by presenting all contexts or choices to each participant or by presenting only one type of context or choice to each participant. The first method, called within-participant, allows an accurate estimation of these effects by ruling out the possibility that the observed differences are due to factors other than those manipulated in the experiment (such as demographic factors). On the other hand, the second method, called between-participant, does not completely rule out the existence of non-measured explanatory variables, but is necessary when the two manipulated conditions are incompatible. For example, once participants have received information on the percentage of female students in each stream, their choices will most likely be influenced by this factor even if this information is no longer available.

The implementation and use of laboratory experiments therefore require several stages of theoretical reflection, requiring both an understanding of this method and a detailed analysis of public policies, in order to guarantee the quality of the data collected (the internal validity of the experiment) and their capacity to explain behaviours and situations relevant to public policies (the external validity of the experiment).

II. How is this method useful for policy evaluation?

The laboratory experimentation method has a dual purpose for policy evaluation. Firstly, it offers a new tool for measuring the target behaviours of policies (the behaviours that the policies seek to modify), offering complementary measures to existing tools such as questionnaires. It can therefore be integrated into the panel of tools that can be mobilised ex post to measure changes in behaviour following the implementation of an intervention or a public policy.

It also allows for a better understanding of the behaviour of interest, to evaluate its key components and to inform the development of public policies. It thus empirically enriches ex ante knowledge of target behaviours to enable the development of better adapted and thus potentially more effective public policies.

III. Examples of the use of this method for policy evaluation in the fields of education, anti-discrimination and urban cleanliness

Laboratory methods have been used in the field of education to evaluate the effectiveness of different interventions, such as sports, meditation or drama, on the executive functions of children and adolescents. Executive functions are a concept from cognitive science that encompasses the psychological processes involved in performing goal-directed actions, requiring, among other things, the use of action planning, inhibition of competing behaviours, and the smooth transition from one action to another. They have been shown to be associated with several measures of school, academic and occupational success, leading to the development of interventions specifically aimed at improving them in children and adolescents. As executive functions are robustly measured by laboratory experiments, such as tasks in which participants must inhibit an automatic response in order to provide a non-automatic response, laboratory experiments have been used to assess the effectiveness of these interventions. For example, to assess the effectiveness of a four-week meditation programme for 9-11 year old students, Parker and colleagues used a Flanker task, a well-known executive function task, to compare the correct response rates of participants in different conditions: when they were asked to indicate the orientation of a target image surrounded by other similar images and when they were asked to indicate only the orientation of these other images (Parker et al., 2014). While this example illustrates how cognitive science concepts and the associated methods can be mobilised for public policy evaluation, it is important to note that these methods can be combined with tools from other fields such as questionnaires. For example, several interventions aiming at reducing racist bias have combined measures of explicit racism, obtained through questionnaires, and of implicit racism, measured through laboratory experiments, in order to get as complete a picture as possible of the effects of these interventions. An example of this is the study published by Devine et al. in 2012, which these researchers found that an intervention combining an explanation of the existence of implicit racist biases and the presentation of strategies to reduce these biases that was conducted on American students did not have a significant effect on implicit biases but did lead to a reduction in racist biases over a two-month period (Devine et al., 2012).

Laboratory experimentation methods have also been applied to assess ex ante the possible effects of new policies. For example, drawing on the policy literature on the importance of bin visibility in reducing street litter, Abdel Sater and colleagues evaluated the potential effectiveness of an intervention to change the colour of street bin bags in a laboratory setting. To do this, they compared the ability of participants to detect bins in street photos based on the colour of the garbage bags. The colour of the bags was manipulated by computer from real photos, so that the experimental task was as close as possible to real conditions, but also as controlled as possible: only the colour of the bags differed between the photos with the grey bags and those with the red bags. This study demonstrated the potential effectiveness of this simple, low-cost intervention on bin visibility (Abdel Sater et al., 2020). Although this example has not yet been translated into the implementation of a real intervention, it illustrates how laboratory experiments can be integrated into the public policy cycle.

IV. What are the criteria for judging the quality of the mobilisation of this method?

Whether its use is *ex post* or *ex ante*, the first element to consider in assessing the relevance of using laboratory experimentation for policy evaluation is the alignment between the behaviour of interest, that which is directly related to the policy question, and the behaviour measured in the laboratory. This idea is fundamental for laboratory experiments to be truly useful for policy evaluation and not just a marketing tool. More precisely, laboratory experiments sometimes use abstract tasks, often initially designed to assess fundamental psychological mechanisms such as motivation. It is therefore necessary to ensure that the behaviour measured experimentally is robustly associated with the behaviour of interest as observed in real-life situations. This question is all the more important as laboratory experiments make it possible to measure not only explicit attitudes, those that participants are prepared to report in interviews or surveys, but also implicit attitudes, of which the participants themselves are not necessarily aware. While the latter type of attitude is of great theoretical interest, it is only weakly predictive of people's behaviour in everyday life and only predicts behaviour in specific situations, such as when people have to make a decision extremely quickly. Thus, an intervention may not have a significant effect on implicit attitudes but still change participants' behaviour. Both levels of measurement can be useful for in-depth policy evaluation and anticipating potential unpredicted effects but using only an implicit level of measurement for policy evaluation can lead to misinterpretations of policy effectiveness.

On the other hand, it is important, as with any tool used in the framework of public policy evaluation, to consider the size of the effects obtained. Indeed, the artificial context in which effects are observed in laboratory experiments calls for caution when mobilising these results for the evaluation of public policies. These often highly artificial conditions and tasks, although they make it possible to isolate the behaviour and factors of interest as much as possible, can also lead to biased interpretations when it comes to generalising these results to real situations. Indeed, an experiment in which only one type of information is given (for example, the name of the newspaper in which an article was published), can lead to an overestimation of the weight of this type of information in the decisions of individuals, because unlike the experimental context, in a real context individuals can base their choices on a multitude of information. The mobilisation of laboratory experimentation for the evaluation of public policies therefore requires taking into account the experimental protocol used as a whole, i.e., not only the type of choice that was measured, but also the type of information to which the participants had access.

Finally, in the case of ex ante use, it is also important to consider the population on which the results were obtained in order to assess whether these results can be used for the target population of the public policy. Indeed, behavioural results obtained only on a particular population may not be valid in another population. These differences between populations are notably important to take into account when the analysis specifically aims at comparing different populations and when the experimental protocol is used in a different population from the one on which it was initially tested. In both of these cases, it is necessary to consider that variations in behaviour observed experimentally may be due to the structure of the experimental design itself and not to differences in the behaviour of interest. For example, differences in the participants' level of concentration on the experimental task may generate differences in behaviour that do not reflect real differences in the target behaviour. It is therefore crucial that the type of experiment chosen be consistent with the target population(s) so as not to artificially create differences in behaviour between populations or to underestimate or overestimate the existence of certain behaviours in these populations.

Finally, in addition to these elements directly linked to the mobilisation of laboratory experiments for the evaluation of public policies, there are general criteria for evaluating the quality of laboratory experiments. These criteria are based in particular on the evaluation of the sensitivity of the experiment and its results to the influence of behavioural bias and randomness. To this end, the use of specific types of formulation, the repetition of each question, the use of a variety of experimental material controlled on key elements (such as the use of a series of different but similarly expressive women's and men's faces to assess gender bias) and the randomisation of the presentation of the different elements of the experiment (the order of presentation of questions and conditions for example) are classically implemented to ensure the reliability of the results of experiments conducted in a laboratory.

V. What are the strengths and limitations of this method compared to others?

The two main advantages of the experimental laboratory method are, on the one hand, that it makes it possible to test the existence of causal links between a factor or context and a behaviour and, on the other hand, that it offers a specific tool for measuring behaviour and attitudes. However, it is important to note that the criteria necessary to conduct a reliable laboratory experiment make this method sometimes more restrictive than other methods. For example, laboratory experiments are often longer than questionnaire surveys, making this method more expensive. At the same time, the need to control for a large number of factors limits the exploratory nature of this method and makes it more appropriate for measuring a specific behaviour or evaluating a given hypothesis.

Furthermore, the highly controlled context of laboratory experiments limits the possibility of directly interpreting the results of these experiments in terms of behaviour outside the laboratory. Indeed, the behaviours of interest are sometimes better predicted by explicit responses than by measurements made during laboratory experiments. However, laboratory experiments make it possible to measure behaviours that are difficult or impossible to identify in interviews or to measure in traditional surveys. Indeed, they offer the possibility of measuring implicit behaviours and are less sensitive to the recurrent biases observed with other methods, in particular the social desirability bias, i.e., the desire of participants to show themselves in the best light and to respond according to what they perceive to be a social norm, although this remains a risk in laboratory experiments. Thus, laboratory experiments hold particular promise for assessing the effectiveness of public policies in changing not only the behaviour of individuals but also implicit biases that can have important long-term effects.

Some bibliographical references to go further

Bordens, Kenneth. and Abbott, Bruce. 2014. *Research Design and Methods: A Process Approach*. McGraw Hill.

Reis, Harry. and Judd, Charles. 2000. *Handbook of research methods in social and personality psychology*. Cambridge University Press.

Gawronski, Bertram. 2009. Ten frequently asked questions about implicit measures and their frequently supposed, but not entirely correct answers. *Canadian Psychology/Psychologie canadienne*, 50(3): 141-150.

Cited references

Abdel Sater, Rita. and Mus, Mathilde. and Wyart, Valentin. and Chevallier, Coralie. 2020. *A zero-cost attention-based approach to promote cleaner streets*.

Devine, Patricia. and Forscher, Patrick. and Austin, Anthony. and Cox, William. 2012. Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of experimental social psychology*, 48(6): 1267-1278.

Parker, Alison. and Kupersmidt, Janis. and Mathis, Erin. and Scull, Tracy. and Sims, Calvin. 2014. The impact of mindfulness education on elementary school students: Evaluation of the Master Mind program. *Advances in School Mental Health Promotion*, 7(3): 184-204.

7. Testing

NICOLAS JACQUEMET

Abstract

Correspondence testing is a quantitative method aimed at measuring discrimination. It involves sending fictitious applications in response to real offers (for example, job offers). By providing and objective measure of discriminatory behaviour, this method is very useful, from a prospective point of view, for designing anti-discrimination policies.

Keywords: Quantitative methods, correspondence, discrimination, anti-discrimination policies, experimental applications

I. How is this method useful for policy evaluation?

Discrimination refers to an unequal treatment on the basis of individual characteristics that should not be relevant to the decision to be made: favouring a male candidate with superior professional skills is not discriminatory; but rejecting a female candidate on the basis of the suspicion that her availability will be less than that of a male candidate with an equivalent profile is indeed discriminatory, because there is no reason to believe that this female candidate matches this stereotype. As such, discrimination is a major source of inequality. Even more than other types of inequality, discrimination is both very costly in economic terms, by depriving the economy of some of its talent; and persistent, because the anticipation of such inequalities of treatment might discourage the discriminated persons and lead them to make choices (of level and branch of education, of career path) which only amplify these initial inequalities.

Despite the importance of the issues at stake, the development of public policies aimed at combating discrimination suffers from a lack of diagnostic elements due to the great difficulty of measuring it. This is the objective of the ” correspondence testing” method.

II. What does this method consist of?

Although this method can be applied to many different sectors of economic activity (housing search, seasonal rentals, master’s students’ applications) and to many different sources of discrimination (religion, sexual preference, socio-economic background, place of residence, disability), this presentation focuses for simplicity on its application to the measurement of hiring discrimination based on gender and/or origin.

This method is designed to provide a measure of the success of different types of applicants according to their socio-demographic characteristics, while at the same neutralising the effect of the intrinsic quality of the applications. Each of these two aims has its own methodological implications.

Constitution of fictitious applications

The success of different types of candidates is observed through the use of artificial applications, sent in response to real job offers circulating on the labour market. The method combines three ingredients: identities, applications and job offers.

The socio-demographic characteristics whose effect is to be measured are conveyed by the identity of the applicant. In order to test both gender discrimination and discrimination affecting applicants from, for example, North-African origin, a list of four fictitious identities (or four different categories of identities) will thus be created: two French-sounding surnames, one associated with a male first name and the other with a female first name, and two surnames that suggest that the person is from North-African immigration, associated with the same variations of the first name. Each of these identities is given a unique telephone number and an email address to contact the applicants. These first and last names and contact information correspond to the identity block of the applications.

In order to respond to job offers, these identities are put forward on applications that most often combine a CV and a cover letter. The aim is to construct applications that are as credible as possible and thus allow to distinguish the success of different identities. The process of constructing applications must therefore lead to a quality that is neither too high nor too low in comparison with the real applications that will be received, because any application that leads to an undifferentiated treatment of applicants, whether positive or negative, makes it impossible to identify the characteristics that favour the success of experimental applications.

The construction of the CV requires filling the training and experience sections with contents that are realistic and compatible with the job for which the application is sent, as well as a section dedicated to extra-professional activities. In order to ensure that these CV elements correspond to the intended occupations, most studies collect real CVs (available, for example, online), then mix the information from several CVs to construct a unique experimental CV and then modify the resulting 'experience', 'education' and 'extra-curricular activities' sections. The content of the CV is completed by a block containing personal information which includes at least the postal address, but may also mention the marital status, the presence of children, the age or the date of birth. The formatting of this information requires choosing as many predefined templates as there are different CVs, which will determine the order of the sections, the font used and the organisation of the different information (many templates in different file formats can be easily found online). Cover letters are constructed in the same way, combining the content of existing cover letters. The gender arrangements (if any, depending on the language) will be adapted according to the gender of the identity on the application (if gender is one of the characteristics tested, it is advisable to choose formulations that allow for as many occurrences of gender arrangements as possible). The combination of a CV and a cover letter is an application.

The job offers to which these applications will be sent are collected from public information sites (in France, many studies use the Pôle Emploi site, but depending on the profession targeted, it is sometimes necessary to use more specialised ones). These offers are filtered to check that they correspond to the predefined inclusion criteria, which primarily concern the occupation and the location of the job, but also, for example, the requirement of specific experience or skills. The remaining vacancies for which it is not possible to send an application according to the predefined modalities (often by e-mail, but also when, for example, the submission of an application requires the completion of an online questionnaire) are systematically excluded. For the remaining job offers, which will be included in the study, all characteristics of the offer (duration, type of contract, salary, etc.) are carefully recorded in order to build up a database to document the observed heterogeneity of job offers.

The number of experimental applications to be sent in response to a given job vacancy (which goes hand in hand with the number of different applications that need to be constructed) is a delicate choice. From a statistical point of view, it is very advantageous to be able to compare the success of different applications in response to a given job offer (i.e. "intra-offer" comparisons), as such comparisons will eliminate the effect of all unobserved elements that are specific to the job offer hence improving the statistical accuracy of the measures. Sending several applications associated to a given socio-demographic group also makes it possible to measure more finely the characteristics of the distribution of discrimination across job offers (see the results presented in Kline et al. 2020). While it is therefore desirable to send several applications in response to each vacancy, the maximum number is limited by two factors. On the one hand, the multiplication of applications increases the disruption caused by the survey on the functioning of the labour market and, above all, the risk of detection. This risk can be contained by taking care to allow sufficient time between the sending of two applications, but this delay goes hand in hand with a reduction in the probability of success for the latest applications, all the more so when the occupation attracts a large number of applications. On the other hand, some recent work (Philips, 2019) shows that the portfolio of applications sent in response to a given offer is likely to affect the relative success of experimental applications. Increasing the number of applications increases the risk of such a bias.

These two factors together imply to be more restrictive with regard to the number of applications sent, the tighter the occupation. The combination of these different factors leads most studies to limit themselves to sending a maximum of four applications in response to each job offer, sent no later than 24 hours after publication. For this purpose, each identity is associated with a single application (a CV and a cover letter), leading to as many unique and distinct experimental applications as the number of applications sent in response to each vacancy.

Measuring the success of the experimental applications requires keeping an accurate, time-limited record (ignoring, for example, responses received more than 3 months after they were sent) of employers' communication with applicants by archiving all written correspondence and transcribing the content of telephone messages received. These responses are then classified to distinguish between refusals, non-responses, requests for further information and invitations to an interview (sometimes referred to as 'expressions of interest'). For obvious ethical reasons, it is imperative to decline any expression of interest as quickly as possible, preferably through the same contact channels and following a predefined script (which most often refers to the previous acceptance of a job offer).

Assembly protocol

The combination of all these ingredients provides a measure of the success of fictitious applications that differ, among other things, in the socio-demographic group to which the application identity is associated. Of course, such differences in success can also be linked to the content of the application itself, which is all the more likely when the applications are clearly different from one another. One solution might therefore be to ensure that the applications are as close as possible to each other. But apart from the fact that any difference, however small, between applications would lead to the same conclusion, it is particularly difficult to distinguish between insignificant differences and more important differences, because the differences that are relevant are the subjective variations in the quality of applications that are perceived by employers.

The protocol that allows correspondence studies to neutralise the effect of any potentially confounding characteristic of the experimental applications (i.e. whose impact on the success rate would lead to erroneous conclusions about discrimination) is to systematically rotate the association between identities and applications. If, for example, identity a appears on application A and identity b on application B in the first mailing, these associations will be reversed in the next mailing (identity a now appearing on application B) before returning to the initial association in the third mailing, and so on. This rotation does not eliminate the effect of the perceived quality of the application: if application A is found to be of better quality, the success of the identity associated with it will be affected accordingly. But rotation ensures that any systematic difference in identity success across all mailings can no longer be attributed to the content of the application itself. From a statistical point of view, any characteristic for which a systematic rotation is organised becomes a source of noise in the measurement of discrimination related to the characteristics of interest, i.e. a source of variation in the success rate between applications belonging to different categories that is not attributable to discrimination. By construction, this noise is independent of the characteristics whose effect the correspondence study seeks to measure and therefore does not affect the ability of the method to measure discrimination. But it does, however, make it more difficult to detect. These consequences of noise in the measurements can be reduced by adapting the statistical analysis accordingly (in the form of offer fixed effects), but such modelling assumes a homogeneous effect of the quality of applications on all employers.

In sum, the correspondence study method is therefore based on three principles: multiplying the number of experimental applications in order to measure the effect of the socio-demographic characteristics by which they are distinguished, ensuring that these applications are as homogeneous as possible in order to reduce the noise that will affect the measurement of their effect, and organising a systematic rotation of the association between socio-demographic profiles and any other characteristic likely to affect their success. These three principles constitute a toolbox that can be applied to many aspects of the functioning of the labour market. For example, one can measure the effect of unemployment spells in the career path by experimentally modifying the "experience" section of applications, of the distance between the place of residence and work by manipulating the applicant's address, or of the family situation by varying the identity block according to, e.g., the presence of children or the marital status.

III. An example of the use of this method

A recent study carried out jointly by the Institute of Public Policies and ISM Corum under the aegis of DARES is one of the first large-scale studies to provide an overview of inequalities in access to employment according to gender and origin in the French labour market´(Dares IPP and ISM Corum, 2021a and 2021b)). These results confirm that ethnic discrimination is both strong and cross-cutting across all the occupations studied, leading to a penalty of around 30% in the chances of receiving a positive response. This study also highlights the lack of discrimination linked to the gender of the applicant, suggesting that, contrary to a persistent received idea, the strong career inequalities that exist on the labour market between men and women cannot be attributed to hiring decisions.

IV. What are the criteria for judging the quality of the mobilisation of this method?

The level of the callback rate for a given type of application provides little information about the functioning of the labour market. The results of a correspondence study are rather based on comparisons of callback rates between different types of applications. These comparisons will only manage to detect the difference in success of different types of applicants if the callback rate among reference applications is sufficiently high.

The variations in the socio-demographic characteristics of applicants are introduced through their identity, which is assumed to affect employers' perceptions. To ensure this, it is increasingly common in testing studies to first run a preliminary survey in which a sample of respondents is asked to associate a gender and/or origin with each of the identities presented to them. This survey provides an empirical measure of the quality of the perceptions induced by the identities, and can be used to select the identities included in the study by retaining those whose perceptions are most consistent with the desired group. Such a survey can also be an opportunity to collect additional information on the perceived profile of the identities presented: recent work shows that identities convey many stereotypes linked, for example, to social class or area of residence, which may contribute to the observed differences in success of applications from different categories (Gaddis, 2017).

Finally, observed differences in callback rates are subject to the famous criticism known as the 'Heckman critique', according to which differences in perceived skill variance within different population groups would be sufficient to produce systematic differences in average callback rates, and would be misinterpreted as a systematic bias against these population groups. This critique can be addressed if enough differences in quality are implemented across experimental applications: the statistical analysis can then allow for group-specific variances in unobserved heterogeneity (Neumark, 2012).

V. What are the strengths and limitations of this method compared to others?

Thanks to its design, the correspondence testing method provides a precise and convincing measure of the extent of discriminatory practices and the specific effect of applicants' socio-demographic characteristics on their successful integration into the labour market. As such, its objectifies a phenomenon that is more difficult to reveal though qualitative approaches: these practices are not easily verbalised in a semi-structured interview, for example, because they are illegitimate and sometimes unconscious. Its main advantage is that it guarantees by construction the independence of these characteristics from all the other elements embedded into the application. The main alternative is to use survey data to study differentials in career paths on the labour market between different categories of the population. But such studies require strong, and often not very credible, statistical assumptions that are needed to neutralise the effect of differences in education or career paths that distinguish these groups and contribute to the observed differences in labour market success.

The scope of the results produced by this method is nevertheless limited by two important factors.

The first is that the success of applications is measured in terms of whether or not they are invited to a job interview. This, however, is a rather imperfect reflection of the final outcome of the recruitment process: the existence of discrimination at this stage of the process only predicts discrimination at the actual hiring stage if all invited candidates are treated equally. If, on the contrary, additional discrimination occurs during the interview process, the measures of discrimination provided by this method underestimate the phenomenon. If, on the other hand, it turns out that the populations discriminated against in the selection of applications are favoured in exactly the opposite proportions when the final candidate is chosen, then these measures distort the reality of discrimination. Audit methods, which consist of using actors playing the role of experimental but real candidates, make it possible to overcome this limitation, but they have the disadvantage of involving a very broad set of factors (physical appearance, voice) that are likely to influence the recruitment process but cannot be distinguished from the socio-demographic characteristics that are apparent.

The second limitation is common to any empirical study but is particularly acute in the case of testings: as discussed above, the more homogeneous the applications, the more accurate the measurements. There are also practical reasons, linked to the fact that the number and specificity of fictitious applications increase with the diversity (geographical or in terms of occupations) of job offers. As a result, testing studies are often limited in scope, and their results can only be conditional on the scope of the study in terms of type of job, sector of activity, geographical area, age range of applicants, etc. The generalisation to the entire labour market of the results observed in this type of study is therefore based on the assumption that the scope chosen does not present any specificities in terms of propensity to discriminate (recruiters' preferences, degree of competition in recruitment, etc.) or, more convincingly, on the accumulation of concordant studies on different spheres of the labour market.

Some bibliographical references to go further

Adamovic, Mladen. 2020. "Analyzing Discrimination in Recruitment: A Guide and Best Practices for Resume Studies." *International Journal of Selection and Assessment* 28, no4 (2020): 445-64.

Adida, Claire. and Laitin, David. and Valfort, Marie-Anne. 2010. Identifying barriers to Muslim integration in France. *Proceedings of the National Academy of Sciences*, 107: 22384–22390.

Dares IPP and ISM Corum. 2021a. "Discrimination à l'embauche selon le sexe: les enseignements d'un testing de grande ampleur." *Dares Analyses* n°26/Note IPP n°67.

Dares IPP and ISM Corum. 2021b. "Discrimination à l'embauche des personnes d'origine supposée maghrébine: quels enseignements d'une grande étude par testing?" *Dares Analyses* n°67/Note IPP n°76.

Edo Anthony. and Jacquemet, Nicolas. 2013. La discrimination à l'embauche sur le marché du travail français. *Opuscule du CEPREMAP*, n°31, Editions rue d'Ulm.

Kline, Patrick. and Walters, Christopher. 2020. "Reasonable doubt: Experimental detection of job-level employment discrimination." *Econometrica* 89, n°2: 765-92.

du Parquet, Loïc. and Petit, Pascale. 2019. "Discrimination à l'embauche: retour sur deux décennies de testings en France." *Revue française d'économie*. Vol. XXXIV, n°1: 91-132.

Gaddis, Michael. 2017. "How Black are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies." *Sociological Science*, 4: 469-89. https://doi.org/10.15195/v4.a19.

Neumark, David. 2012. "Detecting Discrimination in Audit and Correspondence Studies." *Journal of Human Resources* 47, n°4: 1128-57.

Phillips, David. 2019. Do Comparisons of Fictional Applicants Measure Discrimination When Search Externalities Are Present? Evidence from Existing Experiments. *Economic Journal*, 129: 2240–2264.

Fougère, Denis. and Rathelot, Roland. and Aeberhardt, Romain. 2011. “Commentaire: Les méthodes de testing permettent-elles d’identifier et de mesurer l’ampleur des discriminations?” *Economie et Statistique*, 447, n°1: 97-101.

8. Cost-effectiveness analysis

THOMAS RAPP

Abstract

Cost-effectiveness analysis is a quantitative method of comparing the "return on investment" of a given policy (the desired results it produces, in relation to its cost), with other possible policies. This method makes it possible to estimate the effectiveness of a policy, i.e. its capacity to maximise a result criterion for each euro of public expenditure. It is useful for guiding public policy choices and the allocation of public expenditure within a given sector.

Keywords: Quantitative methods, cost-effectiveness, efficiency, effectiveness

I. What does this method consist of?

Cost-effectiveness analysis is a method of exploring the efficiency of a public policy, i.e. in colloquial terms determining its 'return on investment'. It is a comparative method in which the intervention being evaluated is compared to several other options: existing policies, alternatives, etc. This comparison makes it possible to prioritise the different options and to determine which one allows for the optimisation of public expenditure, i.e. to obtain the best possible result for each euro invested. The prioritisation of the different options is based on a simple economic calculation, that of the "incremental cost effectiveness ratio". This calculation relates the difference in costs to the difference in effectiveness between the intervention and its comparators.

Five steps are necessary to implement this evaluation method.

First, it is necessary to choose a perspective for the analysis and a target population. This consists of determining which perspective is adopted for the calculation: that of the public policy financer (payer)? of its beneficiaries? of society as a whole? It should be noted that the perspective often chosen in cost-effectiveness analyses is the societal perspective, because it takes into account the impact of the policy for all stakeholders (payers, beneficiaries, etc.). However, while this collective perspective is more interesting for the public decision-maker, it is also more difficult to implement because it implies a comprehensive measurement of costs and effectiveness criteria. The choice of the target population is often dictated by the objective of the public policy. For example, a public breast cancer prevention campaign targeting specific ages. It is often relevant to identify specific subgroups of people within this population based on, for example, their access to the intervention (e.g., access to health centres) or their exposure to other measures (e.g., access to privately funded preventive care).

Second, the scope of the costs associated with the public policy being evaluated must be determined. The costs considered in the evaluation are those linked to the deployment of the intervention, also known as "direct costs": investment in infrastructure, equipment, salaries of staff dedicated to the intervention, information campaigns, etc. In the context of an evaluation covering several years, these costs are discounted to take account of inflation. Costs indirectly associated with the public policy can also be included in the evaluation. Indeed, if the policy has a strong impact on the domestic sphere, it is often appropriate to include this impact in the calculation of the efficiency ratio. For example, it can be expected that an increase in the generosity of public support for the autonomy of frail seniors will reduce absenteeism from work by family carers, who can substitute professional services for their care time.

The third step is to choose a criterion for the effectiveness of public policy. The choice of this criterion is decisive, as it must be "sensitive" enough to capture the impact of the policy. Two main categories of criteria are generally used in evaluations: outcome criteria and utility criteria. On the one hand, outcome criteria make it possible to measure the effectiveness of the policy being evaluated with numerical indicators: for example, unemployment durations for the evaluation of employment policies, mortality rates for health policies, school success rates for education policies, etc. On the other hand, the utility criteria make it possible to measure the impact of public policy on the well-being of households, measured using questionnaires. This is known as a cost-utility analysis, the objective of which is to measure whether the policy being evaluated optimises the average level of well-being of its beneficiaries.

Fourthly, it is necessary to determine the data sources to be mobilised for the evaluation and to estimate the impact of the public policy on the basis of these data. Evaluations use two main categories of databases: those from "randomised experiments" and so-called "real-life" data. Randomised experiments (see separate brief) involve comparing two different populations, one receiving the intervention being evaluated and the other receiving an alternative. The success of these experiments depends on the absence of "contamination" between the two arms being compared. Indeed, any interaction between the members of the two groups calls into question the evaluation of the effectiveness of the treatment, and therefore the entire cost-effectiveness evaluation. Typically, these studies cover a two-year period and include a few thousand participants. Evaluations based on real-life data consist of identifying ex-post in administrative databases the two populations included in the evaluation (always according to their exposure), and following their evolution over a longer period. They have the advantage of being exhaustive: sample size, number of years of follow-up, etc.

Finally, the cost-effectiveness analysis itself can be carried out. This analysis involves two steps: firstly, the average effects observed on the cost and effectiveness parameters must be estimated from the data, and secondly, the cost-effectiveness impact of the public policy must be modelled on the basis of the estimated effects. The estimation of the effects is carried out using econometric regressions, which make it possible to identify the average impact of the public policy in the sample of data used, and confidence intervals for this effect (upper and lower bounds which frame its average value). This estimation step is essential because it allows us to define whether the impact of the policy is significant or not, in other words, whether the policy is producing the expected results on average. The choice of estimation method is an essential aspect of this stage. This is followed by the modelling phase, in which the effect estimates obtained in the first step are used as inputs to the cost-effectiveness model. For example, different probabilities of emergency hospitalisations will be estimated as part of the evaluation of a falls prevention policy for the elderly. All possible impacts of the policy will be tested to take account of the uncertainty associated with its effects, i.e. the different possible values for the probabilities of hospitalisation. This is called "micro-simulation".

II. How is this method useful for the evaluation of public policies?

Cost-effectiveness evaluation sheds light on the efficiency of a public policy, i.e. its capacity to maximise a result criterion for each euro of public expenditure. It is therefore a decision-making tool that makes it possible to answer many evaluative questions: Is it more cost-effective to implement policy A than policy B? What are the incremental costs and benefits of adopting a public policy? Which specific populations can benefit most from the deployment of this public policy? How can the allocation of public expenditure in a given sector (health, education, security etc.) be improved? Do the expected effects of a public policy exceed the costs of its implementation? Can the efficiency of a policy vary according to the profile of its beneficiaries, the population covered?

The answer to these questions can be given ex ante, in the context of an evaluation exploring whether it is wise to generalise the deployment of a programme implemented in a given geographical area, or ex post, in the context of an evaluation that determines whether a public policy has had the expected economic effects on a given population over a defined period of time.

The main advantage of this method is that it allows all possible economic effects of a public policy on a given population to be considered in a comprehensive manner. Its use therefore improves the transparency of public policy decision-making criteria. By using the results of a cost-effectiveness evaluation, the public decision-maker can arbitrate not only on the basis of economic criteria (costs), but also on the basis of the effectiveness of the results of public policy. In other words, the use of cost-effectiveness evaluation encourages decision-making that is not solely guided by considerations of public expenditure control, particularly when it uses efficiency criteria measured in terms of individual well-being. This method is also identified by France Stratégie as a central tool for comparing the efficiency of different public policies (Desplatz and Ferracci 2016).

III. Examples of the use of this method in the evaluation of health policies

Recently, numerous scientific studies have been carried out to evaluate the efficiency of policies to combat the COVID-19 pandemic. A systematic review of this literature identifies that the main control measures against COVID19 (testing, wearing of masks, social distancing, quarantines) were mostly efficient, i.e. their return on investment was high, and all the more so when the virus reproduction factor was high and they were introduced in combination (Vandepitte et al. 2021). This study nevertheless warns of the existence of country-specific factors (population density and structure, organisation of the health system, etc.) that explain the greater efficiency of these measures in different countries. It shows that the results of an efficiency analysis of a policy carried out in a particular country/context cannot be easily transposed to another country.

Cost-effectiveness assessment can also be used prospectively to guide public decision-making. For example, for more than ten years, the French National Authority for Health (HAS) has been using this evaluation method to determine the efficiency of health innovations and to inform negotiations on pricing and reimbursement of these innovations between the health industry and the Economic Committee for Health Products (CEPS). They are published in a transparent manner on the HAS website. These analyses are one of the main tools used by the CEPS to assess the expected impact of a decision to set the price of a medicine. Manufacturers are responsible for producing evaluation models for their innovations, following a methodological guide designed and updated by the HAS's Economic and Public Health Evaluation Commission (CEESP). This guide details the criteria used to assess the quality of the analysis (HAS 2020). Once the analysis has been carried out, manufacturers submit an "efficiency report" describing the content of the model and its results. Efficiency opinions" issued by the EPHSC conclude on the impact of the efficiency of the introduction of a new treatment on the French market, or evaluate ex post the efficiency of a treatment after several months of use.

IV. What are the criteria for judging the quality of the mobilisation of this method?

The quality of this method depends on the quality of the data used to build the model, and the quality of the method of identifying the causal relationship to measure the effects of the policy. These two points are essential. Indeed, it is essential that the step of estimating the effects of the policy mobilises advanced identification methods (randomised trial, propensity scores, instrumental variables) which are often complex to implement. When these data are not available, it is necessary to mobilise data from the scientific literature to find comparable evaluations in other countries, and to use the data from these evaluations to "feed" the model. If such data are not available in the literature, qualitative interviews must be used, which may reduce the accuracy of the assumptions and the overall quality of the model.

V. What are the strengths and limitations of this method compared to others?

The main advantage of cost-effectiveness analysis is that it is a transparent and easily accessible policy-making tool. Indeed, the comparison of the efficiency of different programmes is carried out with the help of a graphical representation which allows for the easy identification of the most efficient measures, i.e. those with the most favourable cost-effectiveness ratio, as they are less expensive than a comparator for a higher efficiency. Moreover, these analysis methods are robust: they have been used for several decades in all areas of the economy (health, development, education, labour, etc.).

Nevertheless, this method has two main limitations. The first limitation is that this method does not really guide public decision making when the result of the evaluation shows that the policy is more effective but also costlier than another measure. This is often the case when evaluating the impact of a policy aimed at encouraging the deployment of an innovation on a market, which is often more expensive but also more effective than a comparator. The second limitation of this method is the sometimes high cost of implementing it. The implementation of a randomised trial requires a significant financial investment (a “small” experiment can cost several hundred thousand euros). Moreover, estimating the effect of the policy implies a long time follow-up, which is often disconnected from the political time. For this reason, the feasibility of efficiency evaluations often depends on the capacity of the evaluation body to conduct an experiment. In the past, this may have blocked the implementation of public measures. For example, the deployment of telemedicine in France, although desired by the public authorities, has long been partly blocked by the inability of market players and/or the health insurance system to conduct experiments that would allow conclusions to be drawn about the efficiency of these devices.

Some bibliographical references to go further

Desplatz, Rozenn. and Ferracci, Marc. 2016. “How to assess the impact of public policies? A guide for policy makers and practitioners.” Paris, France: France Stratégies.

HAS. 2020. *Choix méthodologiques pour l’évaluation économique*. Saint-Denis.

Vandepitte, Sophie. and Alleman, Tijs. and Nopens, Ingmar. and Baetens, Jan. and Coenen, Samuel. and De Smedt, Delphine. 2021. “Cost-Effectiveness of COVID-19 Policy Measures: A Systematic Review.” *Value in Health* 24(11): 1551-69. https://doi.org/10.1016/j.jval.2021.05.013.

PARTIE II
QUALITATIVE METHODS

9. Direct Observation and Ethnography

NICOLAS FISCHER

Abstract

Direct observation or ethnography is a qualitative method that consists in directly observing the social situation under study – for example, the implementation of a public policy – implying a physical presence of the researcher in the situation at hand. It is a demanding method in terms of the commitment it requires (long-term physical presence in the field, systematic note-taking). It is particularly useful to account for the reality of practices and interactions, at a distance from official discourse.

Keywords: Qualitative methods, ethnography, direct observation, policy implementation, semi-structured interview, interactions, case study

I. What does this method consist of?

Direct observation derives from the practice of ethnographic observation, which is an old tradition in the social sciences, particularly in anthropology. It is part of qualitative evaluation methods. It thus aims to overcome the limitations of quantitative surveys, which are based solely on statistical analyses: the latter provide an overall numerical picture of the results of a policy, but they say nothing about how it is implemented and the concrete difficulties that are responsible for its failures or unexpected effects. Direct observation, on the other hand, allows us to grasp the practical situations that constitute policy implementation on the ground: we then have a first-hand description of the implementation of a given programme, but also of the material conditions of its success or failure.

The direct observation of social practices has a long history. First of all, it is inseparable from anthropology and ethnology: when these disciplines fully constituted themselves as sciences during the 19th century, they progressively theorised ethnography as their main method of data collection. At the time, the aim was to study populations that were geographically and culturally distant. Observation made it possible to reduce the social distance with the subjects of the investigation through immersive research, which involved prolonged stays in the field, learning local languages, and a series of methodological precautions designed to avoid any ethnocentric judgement on the part of the ethnographer. At the end of the 19th century, and in a perspective closer to evaluation, the social surveys conducted in Europe among working-class or marginalised populations also used observation, which again was intended to reduce the social distance separating the ethnographer from the environment he or she was observing. Finally, in the 20th century, observation was used in sociology, and later in political science, to study 'close' objects (public services, political parties, organisations). The challenge is to 'unfamiliarise' these known practices, as the observer's position invites us to decentralise our gaze and to question the causes and social mechanisms of activities that are taken for granted.

Within the qualitative family, observation is often combined with semi-structured interviews (see dedicated chapter on semi-structured interviews), both with administrative agents and with the publics they encounter. Here again, observation makes it possible to reconstruct what these interviews cannot say: first of all, it makes it possible to circumvent the self-censorship that informants often impose on themselves in interviews, particularly when it comes to talking about the quality of their work and the performance of their missions. It also makes it possible to describe precisely certain aspects of public policies that the evaluators and the evaluated would not think of mentioning in an interview. Local routines and habits, the practical organisation of work, postures and attitudes or non-verbal communication with users – and all that they reveal about the social relations and inequalities involved in the relationship between civil servants and their publics – are then made directly visible (Perret, 2008). This type of approach can be particularly useful when the policies evaluated target sensitive populations (precarious or socially marginalised people, people with disabilities, etc.), with whom interactions require specific skills on the part of civil servants: self-presentation, the ability to explain the administrative process or to manage the anxiety or anger of the public encountered.

Conducting ethnography requires special preparation (Becker, 2002). While it may seem easy to go to a place to observe it, it is necessary for the observer's outlook to be informed, and thus to constitute the space(s) studied as a scene of observation. A great deal of theoretical and documentary work is therefore essential to identify the relevant observation sites: which offices to observe, in which location (rural, urban, rich or disadvantaged municipality)? What activities and dimensions should be focused on? Should one try to compare the same moment of policy implementation in different places, or on the contrary analyse the different stages of a single administrative chain? After answering these questions, the ethnographer must go to the field and confront the inevitable tension between closeness and distance from the respondents. Observation implies sharing the daily life of the people being surveyed over a long period of time, while minimising the distance that potentially separates one from them. It is therefore necessary to align one's appearance, speech and body language as much as possible with that of the people being observed. Conversely, it is also advisable to regularly leave the field of observation in order to "retreat" into a space specific to the reflection on the activities observed: in this case, it is a matter of avoiding too strong an immersion in the practice, and thus of reinstating the external position of observation.

Throughout the observation, the observed activities are regularly recorded in a fieldwork diary, in written or recorded form. Although there is no standardised form or method for writing it, this diary must combine not only the description (of the places observed, with plans and sketches, and of the activities taking place there), but also the ethnographer's reactions: surprise, indignation or sympathy in the face of the phenomena observed provide information on the sensitivity of the observer, but also on the divergent sensitivity of the people being observed: it highlights the production of local representations of what is 'normal', 'acceptable' or 'problematic', representations that are not (yet) shared by an outsider who discovers the situation. From a methodological point of view, recording one's reactions during the observation also makes it possible to objectify them in order to analyse them, thus limiting the impact of the ethnographer's subjectivity on their observations.

II. How is this method useful for policy evaluation?

As Stéphane Beaud and Florence Weber (2012) note, the adoption of the ethnographic method results from dissatisfaction with the discourse that a group – in this case an administration – holds about itself: it is a question of going beyond the official presentation of an activity, what the legal rules, instructions or presentation brochures say about it, to analyse the reality of its practice. Such direct observation can therefore take place *ex post*, at the stage of implementation of public policies, which we know often corresponds to a real re-elaboration of the policy by administrative agents. It is particularly justified when it comes to evaluating a policy format that is difficult to quantify (reception at an administration counter, for example, see next section). Such an approach makes it possible to observe the diversity of local investments in the same policy, and its adaptation to the local conditions of its implementation (specificity of the public, of the socio-economic or political context) or of the actors who carry it out (legacy of local routines specific to a department, an office or a municipality). Such a perspective opens up two potential evaluative logics: highlighting the local innovations of which street-level bureaucrats are capable in order to deal with situations not provided for by the texts, and also considering the multiple logics that can possibly cause a public policy to deviate from its stated objective. Typically, this involves evaluating how a policy and the material resources allocated to it adjust with the realities encountered on the ground, identifying the issues neglected during its design, and isolating the practices that need to be modified to enable public action to produce its full effects.

III. An example of the use of this method: the evaluation of the reception policy in public services

Although it is already old, the report submitted to the Prime Minister in 1993 on *Les services publics et les populations défavorisées: évaluation de la politique d'accueil* (Paris: la Documentation française, 1993) [Public services and disadvantaged populations: an evaluation of reception policies] is a good example of the usefulness of the ethnographic method for evaluation. It illustrates first of all the interest of observation in order to carry out a detailed approach to the question initially posed in 1990 by the Interministerial Evaluation Committee: in a context marked by the development of the theme of the modernisation of public services, the challenge was to evaluate the capacity of local public service counters to effectively deal with the difficulties encountered on a daily basis by the most precarious populations. Such an analysis could not be carried out through a purely quantitative evaluation, nor by a simple interview survey: the objective was indeed to take an interest in interactions – that of state services located on the 'front line' with the publics who most depend on the benefit they allow – and to try to evaluate their quality – in particular to judge the capacity of users to effectively assert their rights. The aim was to examine the implementation of reception services, the quality of information provided to the public, the impact on the effectiveness of their rights, the possibility of implementing satisfaction indicators and, ultimately, the appropriateness of adopting selective reception policies, some of which would be adapted to disadvantaged groups.

This report also highlights the fact that observation is often combined with other methods to shed light on ethnographic findings and to connect them to more general statements on the observed administration: in this case, the qualitative survey is combined with a quantitative component (questionnaires sent to users to select them according to their socio-demographic characteristics). Within the qualitative component, the observations made at the counter were supplemented by qualitative interviews with users, reception staff and 'social intermediaries' (associations or civil servants from the social services who facilitate access to public services).

The research required the joint work of the administration's inspection services and consulting firms or academic research centres (3 private firms and a university centre), and a preliminary work of identifying the relevant observation scenes: each fieldwork was prepared by a mapping of all urban services, which made it possible to identify eight public services considered central to the problem of reception (police, hospital emergencies, town hall, etc.) The localities surveyed were selected because of their pre-existing classification as "disadvantaged areas".

These methodological choices are not without bias and illustrate in passing one of the difficulties of ethnographic research and the joint importance of the initial question, and of the observation protocol designed to answer it. In this case, the report concludes that it is necessary to adapt reception policies to disadvantaged populations, in particular by creating platforms or "public service centres" that bring together in the same place, within marginalised neighbourhoods, the offices of different public services (post office, town hall, etc.). These conclusions have been criticised by academics who have conducted their own ethnographic studies of precarious counter users (see Siblot, 2005; also Dubois, 2003): by focusing solely on the dependence of users on public services, the evaluation remains blind, in their view, to the multiple ‘coping’ strategies that precarious populations are able to develop in order to assert their rights, and which an in-depth ethnographic survey reveals. Similarly, the evaluation is accused of making an abusive generalisation by asserting the dominated nature of the users, whereas they are unequally endowed with different sorts of capital, particularly educational capital, and some of them may be in a position to interact on an equal footing with the reception staff.

IV. What are the criteria for judging the quality of the mobilisation of this method?

Ethnographic observation will be all the more useful if the observers have been able to carry out ‘casework’: in other words, to constitute the always singular situations observed in the field into ‘cases’ that can corroborate or invalidate a theory. The challenge is then to ’empirically delimit what is a problematic relationship between ideas and evidence, between theory and data’ (Hamidi, 2012). Maintaining this relationship requires ethnographers to pay constant attention to the practices observed in the field: they regularly bring up unexpected logics or themes, which must lead to enriching or modifying the initial theoretical question. This is an important issue in policy evaluation, where the initial design of the evaluation mission may be modified to avoid neglecting certain realities in the field (a problem raised in particular in the case of the reception of underprivileged populations in public services, cf. previous section).

The complexity of the ethnographic exercise then lies in the ability of observers to articulate, in the same research, cases of different status (Hamidi, 2012, referring to the extended case theory of the Manchester

School). We can thus associate ‘exemplary’ cases for which we can expect, given the context and the populations concerned, that the theoretical hypotheses will be fully validated (to keep the previous example: a post office counter in a working-class neighbourhood of a neglected urban area), and ‘borderline’ cases in which they will only be partially confirmed (another counter located in a less isolated neighbourhood, or located in an area with closer community solidarity or a narrower network of associations). The various factors that can influence policy implementation are unevenly present in these different cases: bringing them together therefore makes it possible to identify with precision those that have a full impact on public action and those that are more secondary.

V. What are the strengths and limitations of this method compared to others?

As we have seen, direct observation makes it possible to grasp ex post the material conditions of the implementation of a policy on the ground, away from official presentations. The identification of observation scenes that illustrate different configurations of implementation of the same policy can allow for a particularly detailed evaluation of the effects of a given policy.

As we have also seen, observation is most often intended to be combined with other methods and complementary approaches. A classic criticism of direct observation concerns the possibility of generalising its results (external validity): observations, carried out in a specific area and necessarily situated, would only concern the local context they describe and would not make it possible to move from the micro-sociological scale to the macro-scale, that of a more global evaluation of the public policy under study. This objection has been partly overcome in recent work, which has emphasised the need to supplement ethnography with other methods, in order to connect the practices observed locally with their institutional framework and its history. This link can be established differently depending on the approach: in Vincent Dubois (2003)'s research on family benefit offices (CAF), the interviews conducted with the staff make it possible to link the observation of interactions at the counter with the career paths of the civil servants, and beyond that with the institutional conditions of their recruitment (absence of a clear definition of the counter staff's mission and job description, etc.). On the same theme, Jean-Marc Weller's research (1999) focuses on the material organisation of reception in administrations and what it reveals (budget cuts, withdrawal of the welfare state and a new managerial conception that turns users into 'clients') in order to link the interactions observed in the field to the global reforms of public action, of which they are the reflection.

Another limitation of the ethnographic method is the investment in time and personnel that it requires. While observation is technically inexpensive – it requires neither recording equipment nor computer processing of the data collected – it does require the presence of an observer, or more often a group of observers working in a concerted manner on several scenes and for long observation sequences (several months), alternating periods of 'withdrawal' and then 'return' to the field. The aim is to capture changes in practices (particularly when evaluating the implementation of a recent reform, which field officials are discovering and then gradually appropriating), but also, as we have seen, to allow the evaluators to regularly withdraw from fieldwork in order to compare their conclusions during the course of the survey and to clarify or modify the general observations they intend to make about the policy being evaluated. Although this long investigation period may therefore seem time-consuming, it is clear that it does not only refer to "field" work and observation: it also corresponds to a period of (re)drafting the final evaluation report and the general conclusions it will propose.

Some bibliographical references to go further

Beaud, Stéphane. 2010. *Guide de l'enquête de terrain: produire et analyser des données ethnographiques*. Paris: La Découverte, Grands Repères Guides.

Dubois, Vincent. 2003. *La vie au guichet. Relation administrative et traitement de la misère*. Paris: Economica.

Hamidi, Camille. 2012. "De quoi un cas est-il le cas? Penser les cas limites." *Politix*, n°100, vol. 4: 85-98.

Jeannot, Gilles. 2008. "Les fonctionnaires travaillent-ils de plus en plus? Un double inventaire des recherches sur l'activité des agents publics." *Revue française de science politique* 58, n°1: 123-40.

Siblot, Yasmine. 2005. « "Adapter" les services publics aux habitants des "quartiers difficiles". Diagnostics misérabilistes et réformes libérales », *Actes de la recherche en sciences sociales*, 159, n°4: 70-87.

Weller, Jean-Marc. 1999. *L'Etat au guichet. Sociologie cognitive du travail et modernisation administrative des services publics*. Paris: Desclée de Brouwer.

10. Semi-structured Interview

CLÉMENT PIN

Abstract

A widely used qualitative research technique, the semi-structured interview consists of a verbal interaction solicited by the interviewer from a respondent, based on a grid of questions used in a very flexible manner. The interview aims both to collect information and to give an account of the person's experience and view of the world, from a comprehensive perspective. It is useful for various types of public policy evaluations, including clarifying the objectives of a policy, analysing its implementation or studying its reception.

Keywords: Qualitative methods, semi-structured interview, induction, empathy, case study, ideal-type, realist evaluation

I. What does this method consist of?

The semi-structured interview is a data collection technique widely used in qualitative research in the social sciences. In very general terms, it is radically different from a questionnaire survey, which aims to produce standardised data on a vast population in order to search for regularities in the variation of opinions or attitudes between groups of individuals by statistical processing. The practice of interviewing, whatever its specific form, is used to produce data that allow us to better understand the singularity of the experience that individuals or groups of individuals have of their relations with others, with institutions, or more broadly of social phenomena. While the qualitative and in-depth study of a singular case may in itself give rise to knowledge with a certain degree of generalisation, this knowledge is usually derived from data processing using several case studies and ideal-types, as well as from cross-checking with data collected by means of the other two classic qualitative techniques, namely observation and the analysis of written sources. Qualitative techniques can also be used in mixed method research.

The practice of interviewing emerged in the nineteenth century in the field of clinical psychology and social enquiry for medical and political purposes respectively. It developed as a research technique in its own right during the 20th century in the United States and then in Europe in a comprehensive sociological approach in the wake of the work of Max Weber. The function of the interview is to gather the words of individuals, the general theoretical postulate being that social phenomena cannot be understood and therefore explained independently of the meaning that individuals give to their actions. On this common basis, several scientific interview practices have been progressively formalised, the main ones being the ethnographic interview, the non-directive interview and the semi-structured interview. However, it is the latter that has become the most widely used technique in policy analysis in recent decades, particularly in France (Pinson, Sala Pala, 2017). In this context, it is often used, if not as an exclusive mode of data collection, at least as a privileged one, on the grounds that it allows for the production of data with intrinsic value (and not only by cross-checking with observations or documentation).

Like other forms of social science interviewing, the semi-structured interview is a verbal interaction solicited by the interviewer from a respondent. However, in the case of the semi-structured interview, the interaction situation is special in that the respondent is initially placed in the role of informant, the holder of valuable (common, non-scientific) knowledge on the topic of interest to the interviewer.

Epistemologically, the semi-structured interview is part of a scientific mode of reasoning in which the fieldwork is not simply meant to verify pre-existing theories developed in the abstract, but rather the basis for developing the research question and hypotheses: the theory is produced by induction from the field data, according to the principle of grounded theory popularised by Anselm Strauss.

What is meant by a 'semi-structured' interview? Although the interviewer should prepare an organised grid of questions to guide the interview, the use of this grid is not rigid. The challenge is for the respondent to provide as much information as possible, both objective (on the phenomena, institutions or processes studied) and subjective (on his/her representations, value system, beliefs). It is therefore necessary to interact with the interviewee in such a way that he or she actively assumes the role of informant, in a conversational manner rather than a questionnaire administered "from above". The quality of a semi-structured interview thus depends to a large extent on the interviewer's attitude of empathy and attentive listening, which will enable him/her to make the most appropriate use of his/her grid of questions in the situation (Kaufmann, 2016).

The application of these methodological principles will never have the effect of cutting short the debates specific to the field of qualitative research and the different forms of interviewing, whether these debates concern the validity of the data collected (their degree of objectivity/ subjectivity, their veracity/factuality, their partiality, etc.), or between scientific paradigms (constructivism/critical realism), so that there is no good use of semi-structured interviewing that is not reflected upon, methodically elaborated, and explained.

II. How is this method useful for policy evaluation?

Semi-structured interviews can be used to address three main types of evaluation questions. First, it can help to make the often complex set of initial objectives of a public policy understandable. Secondly, semi-structured interviews can be used in an evaluation process that aims to trace the processes of policy implementation, to understand how its objectives are concretely translated into the interventions and practices of administrative agents. Finally, although less recognised for this purpose in the French context, the semi-structured interview survey can contribute to producing evaluations by documenting the reception of a policy by its beneficiaries and, more broadly, by the individuals it targets. If these three uses can be combined in the same evaluative research, we will specify their respective contributions in turn.

From the perspective of clarifying the objectives of a policy, the semi-structured interview appears to be one of the rare means of empirically approaching the work of the government and, more precisely, the decision-making processes involved in putting public problems on the agenda and defining policies to deal with them. Because of their highly political nature, governmental spheres remain difficult to access for observation. Written sources, because of their official and consensual character, remain poor in terms of information for capturing debates and controversies between policymakers driven by ideologies, institutional logics and particular interests. The semi-structured interview is therefore used as a technique for retrospectively accessing first-hand information that is indispensable for deciphering the issues that presided over the formation of compromises and trade-offs that are only very implicitly expressed in the official formulation of policy objectives.

In an evaluation approach centred on the study of the means effectively deployed (outputs) in application of a policy, the use of the semi-structured interview appears at first sight to be less central. On the one hand, since the necessary data are by definition of a pronounced administrative and technical nature, they are often available in written form. Moreover, as the agents' practices are considered more ordinary, they lend themselves more to observation, which can be a useful technique at this stage, in order to grasp the practices of adapting the rule to the diversity of situations and publics concerned (see separate chapter on direct observation). However, the semi-structured interview can be used as a complement to cross-check the explanatory hypotheses concerning the agents' practices with the accounts they give of their work situations and the expert representations they develop about the publics they interact with.

The use of semi-structured interviews in the study of the effects (outcomes) of a policy is conceivable if we do not reduce this study to the only (quantitative) measurement of impacts but seek to understand (qualitatively) the process of producing these effects. This type of analysis, formalised in the 1990s by the pioneers of qualitative evaluation such as Michael Patton, emphasises that the same policy can have different meanings depending on the populations concerned, and that this diversity produces significant variation in its effects. The concept of reception (Revillard, 2019) helps to analyse the interactions between the logics of appropriation (cognitive and practical) and the effects (symbolic and material) of a policy. The empirical study of reception involves conducting semi-structured interviews, the particularity of which is to give primacy to the comprehension dimension rather than the information dimension, the examination focusing primarily on the subjectivity of the recipients. Another, less subjectivist, practice of semi-structured interviews is also developed in the realist evaluation. We present it in the next section.

III. An example of the use of this method in the evaluation of educational policies

Theorised by the sociologist Ray Pawson, realist evaluation is now well recognised in international scientific literature and is used by many governmental organisations (see separate chapter on realist evaluation). Its main characteristic is to replace the ordinary question “does this policy work? (in the sense of does it produce the intended effects?) with a more detailed questioning of “what effects does it produce? for whom? in what contexts? under what conditions? The (critical) realism of this approach lies in the postulate that measuring the impact of a policy is insufficient to grasp its effects, and that these are so different depending on the target audience and the context that it is essential, in order to evaluate it, to understand the variety of processes that it activates. Evaluating a policy is therefore a matter of formulating and empirically examining hypotheses about the way in which contexts, mechanisms and outcomes interact (the ”contexts-mechanisms-outcomes“ analysis scheme – CMO).

The work of formulating and examining hypotheses is based centrally on the conduct of semi-structured interviews designed according to a logic described as a teacher-learner function (Pawson, 1996), halfway between the structured and unstructured interview. The informational dimension of the interview is dominant, with the exchange with the interviewee focusing less on his or her experiences and representations than on a reflection on research hypotheses (theory-driven). This interview practice cannot, however, be described as directive insofar as, depending on the phase of the survey, the interviewer and the respondent will alternately play the roles of teacher and learner. In order to help anticipate and control this role switching, Ana Manzano (2016) distinguishes three phases in interview uses. The first set of interviews performs a theory gleaning function, i.e. it identifies provisional hypotheses from the actors about the effects of contextual circumstances on the functioning of the programme studied. In a second phase, certain theories are discarded and the selected theories are examined in greater detail by means of less standardised interviews in order to question the interviewees in a variety of ways with a view to refining the theory (theory refining). In the third phase of theory consolidation, the evaluator acts as a teacher by presenting his or her contextualised understanding of the programme to the respondent, to which the respondent can react by using examples in a logic of verification or falsification.

A recent example of an evaluation conducted in the field of educational policy illustrates this practice of semi-structured interviews particularly well. In order to evaluate the Colombian policy aimed at reducing regional inequalities in educational success by extending the length of the school day universally (Jordana Unica programme), Juan David Parra (2022) carried out a qualitative study consisting of 31 interviews (11 with officials from central and deconcentrated state services, 20 with school headmasters and educators), 20 focus groups (10 with parents, 10 with students) and 40 hours of non-participatory observations in schools. He also administered a questionnaire to a representative sample of school headmasters (N = 681). This survey enabled him to formulate, refine and then consolidate hypotheses on the implementation, reception and effects of this policy by emphasising the importance of reasoning at three levels: the decentralisation of educational policies, the well-being of children and adolescents, and the motivation of pupils.

IV. What are the criteria for judging the quality of the mobilisation of this method?

A first element conditioning the quality of a study based on semi-structured interviews concerns the number and choice of interviewees. As the representativeness of the sample is not a criterion of validity, the principle is rather to carry out a sufficient number of interviews (generally estimated at between 20 and 30) to gather the testimony of people who, from a formal or informal point of view, occupy different positions and are in different situations with regard to the object studied, so that they may have different points of view, in other words, varied experiences, practices and representations about it.

A second quality criterion is the way the interviews are conducted. The semi-structured interview must alternate between moments intended to collect narratives or stories freely produced by the respondent (generally at least at the beginning of the interview) and moments of greater directivity aimed at collecting information previously targeted by the interviewer. This art of interviewing is prepared beforehand by drawing up an interview guide, which evolves over the course of the research and can be adjusted according to the interviewees. This guide not only includes the formulation of initial instructions and general themes for discussion, but also establishes a series of follow-up questions that make it possible to obtain the information sought. Conducting interviews also depends on the posture the interviewer and respondent adopt in the situation and the follow-up techniques used by the interviewer.

A third set of issues lies in the processing of the data collected by the interviews. This decisive stage aims to analyse the content of the interviews in a cross-referenced and comparative manner so as not only to synthesise and cross-check the information collected, but also to produce an interpretation that is both global and detailed of the object studied, with reference to the theoretical framework and the research hypotheses initially formulated. This phase of the work requires the data collected in each interview to be relatively decontextualised by analysing their content in terms of the categories of analysis relating to the functioning of the action system and/or the processes studied and the experience of the various actors concerned.

V. What are the strengths and limitations of this method compared to others?

The main advantage of semi-structured interviews is that they provide essential data for understanding the processes by which a public policy produces its effects, from the genesis of the multiplicity of its objectives and its content (means devoted, instruments developed), to the actual methods of implementation and the various ways in which it is received. These data relate to the practices and representations of all the actors involved or more widely concerned (a priori) by the same policy. Depending on the research stages and the types of respondents solicited (decision-makers, implementers, beneficiaries, recipients), the use of the semi-structured interview can be modulated to activate its informative or comprehensive dimension first.

Its main limitations are twofold. Firstly, in the context of a strictly qualitative evaluation, it is required that the administration of proof operates by crossing the use of interviews with other data collection techniques, namely observation and the analysis of written sources. Secondly, as a qualitative method, it is clear that the use of the semi-structured interview does not in itself allow for the production of quantitative evaluations, evaluations which are otherwise very useful in providing contextual data for the design of the questioning of a qualitative evaluation.

Finally, it should be noted that in the current context of quantitative impact evaluation development, semi-structured interviews can find their place in the framework of research adopting a mixed methodology (Pin, Barone, 2021). Semi-structured interviews can thus contribute to the design (upstream) and interpretation (downstream) of a randomised experiment. In this case, as in others, the use of the interview will be modulated according to the research stages. The semi-structured interview technique will initially be used in a "qualitative instrumentalized" way to help identify the various contextual conditions of implementation of a programme whose impact we are trying to measure and thus refine its implementation methods. The semi-structured interview can then be used in an ’empowered qualitative’ logic to construct ideal-types that provide a posteriori explanatory elements of a qualitative nature to understand the causal processes that led to the measured impacts.

Some bibliographical references to go further

Kaufmann, Jean-Claude. 2016. *L’entretien compréhensif*. Armand Colin.

Manzano, Ana. 2016. “The craft of interviewing in realist evaluation”. *Evaluation*, n°22: 342-360.

Parra, Juan David. 2022. “Decentralisation and school-based management in Colombia: An exploration (using systems thinking) of the Full-Day Schooling programme”. *International Journal of Educational Development*, n°91: 102579.

Pawson, Ray. 1996. “Theorizing the interview”. *British Journal of Sociology*, n°47: 295-314.

Pin, Clément. and Barone, Carlo. 2021. “L’apport des méthodes mixtes à l’évaluation”. *Revue française de science politique*, n°71: 391-412.

Pinson, Gilles. and Sala Pala, Valérie. 2007. “Peut-on vraiment se passer de l’entretien en sociologie de l’action publique?”. *Revue française de science politique*, n°57: 555-597.

Revillard, Anne. 2018. “Saisir les conséquences d’une politique à partir de ses ressortissants: la réception de l’action publique”. *Revue française de science politique*, n°68: 469-492.

11. Focus Groups

ANA MANZANO

Abstract

Focus groups are a qualitative method which consists of a researcher leading a collective conversation with a group of (commonly 4 to 8) people, guided by questions for them to comment on. Focus groups can be a way of involving users and diverse stakeholders' viewpoints and experiences in the evaluation of a given intervention. They are suitable at different stages of the policy process and for different evaluation approaches, often in combination with other qualitative and/ or quantitative methods.

Keywords: Qualitative methods, focus groups, group interviews, participatory evaluation

I. What does this method consist of?

Focus groups consist of one or more conversations with a group of people assembled for discussion. Focus groups are led by a researcher (and often include an observer), aiming to gain knowledge of different possible outcomes in areas such as selling (marketing), influencing decisions (politics, health behaviour), or assessing the worth of public interventions and policies (monitoring and evaluation). Trained researchers facilitate discussions with a set of unstructured and/or structured questions for the group to comment on. These conversations can be stimulated by prompts such as photos, videos, vignettes, games, etc and by decision-making techniques such as informal voting methods. Focus groups can be conducted in a variety of formats (digital, analogue, virtual) and with participants of similar characteristics of interest (referred to as “homogeneous focus groups”, such as a focus group comprising of teachers) or diverse characteristics (referred to as “heterogeneous focus groups”, such as a focus group comprising of teachers, students and parents).

Focus groups are a primary qualitative research method, belonging to the family of group-based discussion methods. Similar to other qualitative data, there is no agreement in the research methods literature on the optimum number of participants for a focus group or the suitable number of groups. At design stages, it is useful to present focus group sizes in ranges because there are many contingencies that can impact the number of participants attending groups. Some authors favour smaller groups (n=3-5) because they have greater potential to explore complex topics in depth. For example, richer information can be obtained by conducting two groups of four participants than one group of eight participants. Some recommend medium size groups (n=6-8), while others suggest bigger groups (n=6-12) to capture a greater variety of views. The duration of the discussion depends on group size and topic but, as a rule of thumb, 90 minutes are necessary for all discussants to have the chance to express their views. Durations longer than two hours can increase participant burden and will also increase the risk of deterring people from attending in the first place.

Many expressions are used in evaluation to describe group data collection methods and there are geographic preferences between the use of “focus group” vs “group interview.” A key distinction is that focus groups highlight the significant role of group dialectical processes (e.g. norms, dynamics, non-verbal communication) that can assist evaluators in gaining knowledge about group views and subgroup agreements and disagreements. Conceptual differences between many of the group data collection terms are often unclear and there is not always consensus on how they are different from each other. Often group/stakeholder sessions (community group meetings, advisory groups, public engagement workshop consultation events, knowledge cafes, expert meetings, collective facilitated conversation with groups, etc) are not designed or conducted as per standard focus groups discussion format. Although sometimes these sessions can be directed at particular policy beneficiaries' groups, all those require less preparatory work, no structured facilitation from researchers, and lack post-event content/ transcript formal analysis.

II. How is this method useful for policy evaluation?

Although policy evaluation has been dominated by the search for hard facts through experimental and quantitative approaches, policy makers also have a preference for user/customer involvement, and focus groups have the potential to support this participatory aim. Alongside in-depth interviews, focus groups are one of the most used qualitative social research methods in policy evaluation. The distinctive features of focus groups are attractive to policy makers, such as exploring contrasting meanings, values, experiences, viewpoints and behaviours from different subgroups of stakeholders, and capturing the complexity of policy implementation contexts and processes. The political value of focus groups is often as important as the specific information about values and multiple viewpoints that the groups can provide.

On their own or combined with other research methods, focus groups are used in many evaluation approaches (e.g. theory-driven and theory-based, process and outcome/impact evaluations, developmental, participatory and empowerment evaluation), for a range of purposes, and at different stages of the policy process (planning, implementation, monitoring, assessment, successive programming cycles). They are suitable for ex ante and ex post evaluation approaches, and are often used in evaluability assessments, needs assessments, programme theory development, instrument and survey development, implementation, utilisation-focused and formative evaluations.

Focus groups are mostly useful to answer exploratory evaluative questions (why and how) because they provide a dynamic means to portray policies in action. They have the potential to increase understanding of:

- a problem, or a policy to approach a problem, and how it is perceived and experienced by different stakeholders (users, front-line staff, management), their expectations and solutions proposed by them.
- to get feedback on quality, use and satisfaction related to the activities and resources delivered by the policy. What worked well, for whom, and what did not work as intended, why and in what circumstances this happened.
- the policy implementation process (e.g. the management, the partnerships with other institutions/departments, the delivery of policy activities and resources).
- to discern the types of changes assumed/expected (theories of change) and produced (if any) from different user perspectives and in different policy contexts across time and space.
- to explore evaluation indicators/criteria when they are not clear or alternative criteria are sought.
- to understand people's experiences of outputs and short-medium-long term outcome patterns (intended and unintended) observed in different macro-meso-micro policy contexts and/or as a consequence of the changes (activities and resources) brought about by the policy.
- to develop and pre-test other qualitative and quantitative data collection instruments such as interviews, experiments and surveys.

III. Two examples of the use of this method: developing indicators and assessing the implementation of a childhood development programme

Focus groups should be used in policy evaluation according to the type of evidence to be generated. For example, focus groups – combined with other primary and secondary methods – are often used to develop evaluation indicators (e.g. participation rates, incidence) that can help answer evaluation questions by marking accomplishments (outputs/ outcomes) in a specific and measurable way. Involving beneficiaries and other stakeholders to develop indicators could make the policy relevant to them and enhance buy-in of evaluation findings. EVALSED (European Commission, 2008), a resource providing guidance for the evaluation of socio-economic development policies in the European Union, gives an example of using focus groups with policy beneficiaries (e.g. representatives of regional enterprises) to develop evaluation indicators in an economic development policy in Benton Harbour (Michigan, USA).

Formative evaluations, which aim to develop policies by examining their implementation, often use focus groups. The formative evaluation of the UNICEF Early Childhood Development (ECD) Project of the Integrated Maternal and Child Health and Development Programme (2017-2020) in China (Zhou Hong et al. 2022), employed a theory-based, utilisation-focused, mixed-methods design that included focus groups. The evaluation results provided evidence to advocate for the national scale-up of the ECD model and informed the design of the National Health Commission-UNICEF Scaling up of Early Childhood Development Program 2021-2025. Focus group discussions with younger parents/ caregivers identified needs for nurturing care skills and this evidence was a driving force for recommending the scale-up of the ECD. Stigma in home visits was also raised in some group discussions and additional attention to privacy protection was recommended for scaling-up. Focus groups with administrators reinforced recommendations to increase funding for the implementation of three types of services, guarantee service frequency and increase service coverage.

IV. What are the criteria for judging the quality of the mobilisation of this method?

Since there are multiple evaluation approaches that differ greatly in their philosophical and methodological premises, a single set of quality indicators for conducting focus groups in evaluation does not exist. This is because each of those approaches has diverse and contradictory assumptions and what matters in terms of 'quality' varies according to these assumptions.

Similarly, qualitative research is not a uniform approach, comprising many different qualitative traditions based on different paradigms, with diverse philosophical assumptions, that a single quality framework could not address. The area of "qualitative data quality criteria" is controversial, with various positions and many classificatory suggestions available, which range from a total rejection of the notion of criteria, to those who propose similar criterion for quantitative and qualitative research.

Consequently, although there are abundant quality criteria on when to use, how to design, recruit, conduct and analyse focus groups, there are no agreed standards for judging quality in qualitative research evaluations. Focus groups come in many formats, and this is why practical, design and quality issues can take on rather contrasting characters. For example, choice of venue is important for "real life" focus groups (as opposed to virtual reality focus groups) since successful recruitment may depend on venue accessibility and practicalities (travel costs, refreshments, audio recording); issues of duration and facilitation are always important, but they will be taken to another level in computer-mediated discussion forms (synchronous or asynchronous).

Spencer et al. (2003, 16) proposed a general framework to support quality indicators on four qualitative methods, including focus groups. This framework is based on four essential guiding principles: 1) To be contributory in advancing wider knowledge or understanding; 2) To have a defensible design and strategy that aims to answer the given evaluation questions; 3) To be rigorous through systematic, transparent data collection, and analysis and interpretation of data; 4) to be credible by offering justifiable, defensible and plausible arguments about the significance of the data generated.

Ryan et al. (2014) proposed that evaluators consider core questions to maximise their learning when conducting focus groups and to improve the credibility of focus group evidence, such as: "Did the focus group participants establish common ground in conversation or primarily act as individuals?"; "What were the power dynamics between the moderator and participants, both as a group and as individuals?"; "What were the relations among the participants – collective or dominant?"

V. What are the strengths and limitations of this method compared to others?

The strengths of using focus groups include:

- In a group, people can build upon/challenge one another's responses and think of ideas that they may not have thought of on their own. This rich blend of perspectives and disagreements can enlighten researchers on policy complexities, often not attainable from less dynamic methods.

- The flexible format is conducive to exploration of unanticipated outcomes and contextual differences.
- Focus groups are often recommended as a time saving and cost-efficient method but the evidence for such assertions is unclear.

Focus groups present the following limitations:

- They are better used in a mix of method designs instead of as a stand-alone method.
- They are not suitable for the discussion of overly sensitive and/or controversial topics because people are less likely to open up about those in a group and the promise of confidentiality and anonymity is compromised.
- They often do not provide a high level of nuance or detail.
- Focus groups can be challenging for people with physical and communication access needs. Inclusion strategies for all abilities are needed, such as choosing accessible locations and rooms, conducting discussions online, smaller sample size focus groups, etc.
- Culturally responsive focus groups must be mindful not only of language and cultural identities but in some cultures, they may be better replaced with other group conversational decolonised methods, such as sharing circles based on open-structured storytelling.
- Focus groups have unpredictable composition and dynamics. Some groups of stakeholders are notably hard to recruit for group discussions.
- Those who are comfortable speaking in front of a group are more likely to be recruited.

- Discussions can be side-tracked and/or dominated by vocal individuals/group opinion leaders. Status differences between researchers and participants, or among participants, will influence discussions.

Some bibliographical references to go further

Cohen Miller, Anna. and Durrani, Naureen. and Kataeva, Zumrad. and Makhmetova, Zhadyra. 2022. “Conducting Focus Groups in Multicultural Educational Contexts: Lessons Learned and Methodological Insights.” *International Journal of Qualitative Methods*, 21: 1–10. https://op.europa.eu/en/publication-detail/-/publication/752dd092-3ea5-429c-96dc-4507d0cda886.

European Commission. 2008. “EVALSED: The Resource for the Evaluation of Socio-Economic Development.” https://op.europa.eu/en/publication-detail/-/publication/752dd092-3ea5-429c-96dc-4507d0cda886

Krueger, Richard. and Casey, Mary Anne. 2000. *Focus Group: A Practical Guide for Applied Research*. Thousand Oaks, California: Sage.

Manzano, Ana. 2022. Conducting focus groups in realist evaluation. *Evaluation*, 28(4): 406-425.

Ryan, Katherine. and Gandha, Tyszha. and Culbertson, Michael. and Carlson, Crystal. 2014. “Focus Group Evidence: Implications for Design and Analysis.” *American Journal of Evaluation*, 35(3): 328–45.

Spencer, Liz. and Ritchie, Jane. and Lewis, Jane. and Dillon, Lucy. 2003. “Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence.” London. https://www.cebma.org/wp-content/uploads/Spencer-Quality-in-qualitative-evaluation.pdf.

Zhou, Hong. and Yang, Li. and Wang, Yan. and Liu, Shuang. and He, Hong. 2022. “Formative Evaluation of the National Health Commission-UNICEF Early Childhood Development Project of the Integrated Maternal and Child Health and Development Programme (2017-2020).” https://www.unicef.org/evaluation/reports#/detail/13964/formative-evaluation-of-the-national-health-commission-unicef-early-childhood-development-project-of-the-integrated-maternal-and-child-health-and-development-programme-2017-2020.

12. Group Interviews

CHARLOTTE HALPERN

Abstract

The group interview[1] is a qualitative method through which semi-structured interviews are conducted with several people at the same time. This method intends to artificially recreating a set of social interactions between a selected number of participants, for example different policy stakeholders. It is useful to various forms of policy evaluation, such as ex ante, ex post and process evaluation.

Keywords: Qualitative methods, interview, group interview, elites, plurality, constructivism, interpretivism

[1] My interest in group interviews as a method for public policy research derives from the experience accumulated thanks to the financial support provided by three EU funded research projects on sustainable mobility transition in European cities. Between 2015-2022, I organised some 20 group interviews across 14 cities and in Brussels with a variety of stakeholders. I am particularly grateful to the H2020 CREATE project (Grant n° 636573) partners, in particular Pr. Peter Jones, Charles Buckingham and Lucia Cristea, for having supported the idea of using this method to examine the role of policy developments in achieving a peak car, to the H2020 MORE project (Grant n° 769276) which provided an opportunity to strengthen the methodology, thanks to Dr. Jenny McArthur's suggestion to link group interviews with a stakeholders' mapping exercise and including the data thus collected in a larger dataset, and lastly, to the H2020 CIVITAS SUMP PLUS project (Grant n° 814881) during which I experimented with hybrid and remote group interviews in the context of the COVID-19 pandemic. All views are those of the author.

I. What does this method consist of?

Interviews conducted with several actors (or stakeholders) at the same time refer to a diverse set of well-known qualitative methods in social science research. Their specific use depends on the role and function they hold in a research strategy (Knott et al., 2022), as well as their properties (Duchesne, Haegel 2008). Among them, group interviews are of particular relevance for policy evaluation research. They are not to be confused with other techniques such as group discussions, focus groups and pre-tests, mainly because they do not require being tethered unto a common experience, nor for participants to share homogeneous professional and social statuses (Marier et al., 2020). In artificially creating a set of social interactions between a selected number of participants, they differ from ethnographic methods, including observations.

Group interviews are understood as a technique used in public policy research, to launch an informal group discussion with a small group of knowledgeable stakeholders and experts – also referred to as "elites" (Glas 2021) – whose contribution is thought relevant for the understanding of the issue under study, including the evaluation of a public policy programme. The added value of group interviews does not lie in the time saved by interviewing several people at the same time – this view is erroneous as group interviews require considerable preparatory work and data processing than a series of one-to-one interviews (see separate chapter on semi-structured interviews) – but in providing the opportunity to artificially generate social interactions among a diversity of stakeholders. It helps identify and make sense of a plurality of perspectives, interests, and values, as well as shedding light on contradictions and ambiguities. Following Frey and Fontana (1991, 183), group interviews « take advantage of group dynamics to produce new and additional data. In addition to the respondent-interviewer relationship, the evolving relations among group members can be a stimulus to elaboration and expression. »

Group interviews may play a decisive role in qualitative research designs in different ways. First, when introduced in an exploratory perspective in the earliest stage of the research, group interviews are particularly useful in the case of a little-studied subject, for which the sources are scarce and insufficiently diversified. Second, by drawing on a "group effect", group interactions may foster insightful perspectives on a given topic that would have remained hidden in observations or one-to-one interviews. As such, group interviews provide an opportunity to artificially generate a set of social interactions to express shared views or disagreements on a given topic (Morgan 1997), while leaving the possibility for additional one-to-one interviews with a selected number of participants. For those practitioners at the very top of their organisational structure, joining a group discussion constitutes a decisive factor for making time for the interview (Glas 2021). Third, during the earliest stage of research, they can be used to examine the robustness of the set of hypotheses stemming from the literature review and to refine them accordingly. Group interviews are of relevance in the context of a comparative research framework, with the same interview guide being applied across the cases under study to provide a first general comparative overview and generate some hypotheses on a case-by-case basis.

Furthermore, in selecting 8-12 participants, the organiser aims at bringing together a set of knowledgeable stakeholders and experts, representing a diversity of views on the object under study due to their respective background, roles and functions in their own organisations. Diversity may vary according to the policy context and the research question. It may refer to different training and professional backgrounds to ensure some cross-disciplinary discussions, to different roles and functions[2] to allow for a variety of concerns, contexts, and priorities to be addressed, or to reflect the large range of organisations and institutions that characterises

[2] Elected representative, technician, civil servant, NGO activist, business owner, etc.

this policy context. In case studies that cover a longer period of 40 to 50 years, diversity may refer to different generations of stakeholders and experts.

Depending on the evaluative research question, data availability and whether the data is collected in the same language, this qualitative dataset can be coded for analysis using a qualitative analysis software such as InVivo (see also Knott et al., 2022). It can thus be used for text or discourse analysis, but also to produce a stakeholders' mapping or a policy timeline, providing a strong basis for further developing the dataset and deepening the analysis through more targeted evaluative questioning.

II. How is this method useful for policy evaluation?

The extent to which group interviews may play a decisive role in a qualitative research design has already been addressed. In the context of public policy evaluation, it offers an opportunity to re-examine the boundaries of well-known policy problems as well as causal relations (Zittoun et al., 2021). Drawing on the constructivist-interpretative school of thought, this method takes a critical view on rationalist premises and highlights the constraints resulting from the various factors that may complicate evaluation activities (Wollman, 2006). It acknowledges that policy goals (as intended consequences) are often vague, ambiguous, potentially contradictory, or mutually exclusive. Public policy goals are understood in different ways by key policy actors, let alone by stakeholders, and while not necessarily accurate, these various understandings of policy problems and solutions are nonetheless fed back into the policy process, influencing its direction and future developments. This raises significant causality problems, more so for policy issues that are characterised by complexity and uncertainty, and at a time of crisis (Voss and Kemp, 2006). Based on these observations, group interviews seek to artificially generate a set of social interactions to critically examine causal relations between expected or observed changes and a given policy programme or measure.

Having this in mind, group interviews are useful to a variety of evaluative questions, such as ex ante, ex post and process evaluation, whether in combination with other evaluation methods or as a stand-alone. When it comes to ex ante evaluations, it can be drawn upon as an opportunity to examine (more or less) explicit causal relations between stated goals, the proposed selection of means, as well as expected results (see separate chapter on theory-based evaluation). In addition, it helps make sense of how alternative policy options are debated, what worldviews and arguments are being used, and what risk mitigation strategies are being developed to overcome expected resistances. This may, in turn, feed into decision-making and shed light on existing contradictions and ambiguities. Group interviews have also been particularly useful to feed into process evaluation, also in an accompanying (running in parallel) or an intervening mode. In this case, its function is to identify and make sense of interim effects while implementation is underway. Lastly, in the case of ex post evaluations, whether focusing on methods or findings, group interviews shed a complementary light to targeted evaluative questioning, often helping to make sense of potential disconnects between stated policy goals and their un(intended) effects, to discuss the use of a given set of indicators, and to spark a debate about future policy programmes. This, in turn, may contribute to examining learning processes, either as an object of research or of intervention.

III. An example of the use of this method

Group interviews have been used in a diversity of public policy research contexts, including evaluative questions. In the background of the climate crisis, it opens new avenues for the evaluation of transition and adaptation policies. As policies aim at achieving long-term goals, transition and adaptation policies refer to the change from the possible to the desirable, and progress is assessed in relation to policy futures that are not unequivocal. By contrast to technically clear problems, transition policy problems do not draw on a clear definition or solution, they are characterised by uncertain causal-effects relationships, and they bring together a wide range of stakeholders with conflicting values or interests, thus accounting for constant disagreements over the means to address the problems (Van der Steen et al. 2016). This fosters the need to draw on evaluative research designs in which degrees of divergence in values are purposefully examined and debated (Delahais et al., 2020).

Focusing on sustainable mobility transitions, Hickman and Banister (2014) examined the extent to which the future constitutes a challenge for policymakers, as well as the shortcomings of dominant methods as identified in the literature, such as forecasting and modeling in particular, or classic approaches used in scenario analysis. Reflecting on the work achieved together under the Urban Buzz Project[3], they account for how a backcasting approach to transport planning in London was set up with the explicit goal to assess the existing strategy's carbon efficiency and contribute to the development of a new strategy aimed at a 60% reduction in transport emissions by 2025 and 2050. The research design drew on a combination of methods, including group interviews, which took the format of workshops with policy-makers and, alternatively, with policy makers and stakeholders, at each stage of the process. The research design explicitly sought to bring the role of values back in the analytical framework, to assess the diversity of representations about transition futures, the hierarchy of values associated with transition processes, the range of implementation strategies at hand and the extent to which such choices were debatable. By contributing to the

[3] See the VIBAT London (Looking Over the Horizon: Transport and Global Warming - Visioning and Backcasting for Transport in London) project's website: https://www.ucl.ac.uk/urbanbuzz/projects_28.php (last consulted November 8, 2022).

development of a backcast scenario closely articulated with an implementation pathway, the project confirmed the relevance of examining stakeholders' values to address transport futures and fed into a changed approach to mobility in London.

IV. What are the criteria for judging the quality of the mobilisation of this method?

The simultaneous interview of stakeholders is not necessarily a timesaving research strategy. The logistics require a considerable amount of preparatory work and data analysis (Duchesne, Hagel, 2008). Being exploratory in nature, group interviews are, indeed, grounded in extensive preliminary research, such as a literature review, an assessment of data availability – grey literature, public reports, press clippings, political party manifestos, etc. – and a mapping of main stakeholders. This feeds into the production of an interview guide, which contributes to structuring the discussion while at the same time serving an exploratory purpose. It may include a small number of purposive questions to guide the discussion. In addition, small-group discussions may be encouraged through dedicated sequences, to produce a detailed and/or context specific understanding of working relationships across different organisations or to generate a precise understanding of a policy timeline, to be reflected on a paperboard.

The group interview organiser should also be aware that bringing together such a diverse group of stakeholders can be a perilous exercise, especially if the topic is contentious. While seeking to foster an informal and lively discussion, group interviews should take place in a formal framework. Also, participants may be reluctant to attend a group interview, fearing that it may only lead to a general and informal discussion. It is thus critical to clearly introduce it as a research method and to provide a (light) structure to avoid overly general and trivial discussions. While the discussion should not last more than 3-4 hours, accommodating time for a break will offer some opportunities for small talk. To avoid putting participants in a difficult position, participants must be informed in advance of the interview's main features and the list of participants, and must provide their informed consent. Decisions about anonymity or confidentiality, data storage and dissemination, are to be addressed when asking the participants' informed consent, whether in written or oral. Depending on the chosen approach for analysing the data, group interviews can be audio recorded and detailed notes can be taken during the discussion for the purpose of the research team. No public external to the interviews' organisers and participants should be admitted.

Group interviews thus require important preparatory work to decide on the selection of participants, the interview guide and whether accommodating small group discussions might be useful to explore a specific issue into more depth.

V. What are the strengths and limitations of this method compared to others?

To conclude, group interviews present several advantages to policy evaluation research and practice. When used in an exploratory perspective, at the earliest stage of research, they help examine the robustness of the set of hypotheses resulting from the literature review, to provide a first general comparative overview and to generate context specific hypotheses. By artificially generating a set of social interactions or "group effect", they provide an opportunity for participants to express shared views or disagreements on a given topic. As such, they are a powerful data collection technique, which provides a fresh look on a given topic that would have remained hidden in observations or one-to-one interviews.

By artificially generating a set of interactions, the “group effect” produces a highly original dataset, consisting of new information and evidence. By sharing their views and potential disagreements about a specific policy issue, its narrative, causal relations, and effects become debatable again, thus contributing to open new avenues for evaluative research or to inform policy making. Moreover, group interviews help generate a robust set of general and case-by-case assumptions, to question the relevance of external and internal drivers of change, to identify the effects of a given policy measure while at the same time taking into consideration wider policy considerations (and questioning its (unintended) effects).

Yet, they are ill adjusted for a targeted evaluative questioning. Other qualitative methods, such as focus groups would be better suited, mainly because group interviews do not require participants to share a common experience, homogeneous professional and social statuses. Also, group interviews seek to artificially create a set of social interactions between a selected number of participants in which they are encouraged to express their disagreements on a given topic, whether the diagnosis of the problem, the hierarchy of values to select a course for action, or its effects. As such, they also differ from ethnographic methods, including observations, and from one-to-one interviews.

Some bibliographical references to go further

Delahais, Thomas. and Sage, Kate. and Honoré, Vincent. 2020. Evaluators in Transition. *Zeitschrift für Evaluation* (ZfEv), 2: 239-260.

Duchesne, Sophie. and Haegel, Florence. 2008. L’enquête et ses méthodes: l’entretien collectif. Paris: Armand Colin.

Frey, James H.. and Fontana, Andrea. 1991. The group interview in social science research. *Social Science Journal*, 28(2): 175-187.

Glas, Aarie. 2021. Positionality, power and positions of power: reflexivity in elite interviewing. PS, *Political science & politics*, 54(3): 438-442.

Hickman, Robin. and Banister, David. 2014. Transport, climate change and the city. London: Routledge.

Knoll, Eleanor. and Hamid Rao, Aliya. and Summers, Kate. and Teeger, Chana. 2022. Interviews in the social sciences. *Nat Rev Methods Primers*, 2(73). https://doi.org/10.1038/s43586-022-00150-6.

Marier, Patrick. and Dickson, Daniel. and Dubé, Anne-Sophie. 2020. Using focus groups in comparative policy analysis. In: Peters, B. Guy. and Fontaine, Guillaume. Eds. *Handbook of Research Methods and Applications in Comparative Policy Analysis*. Cheltenham: Edward Elgar Publishing, 297-310.

Morgan, David L. 1997. *Focus Groups as Qualitative Research*, London: Sage.

Steen, Martijn van der. and Chin-A-Fat, Nancy. and Vink, Martinus. and Twist, Mark van. 2016. Puzzling, powering and perpetuating: Long-term decision-making by the Dutch Delta Committee. *Futures*, 76: 7-17.

Voß, Jan-Peter. and Kemp, René. 2006. Sustainability and reflexive governance: introduction. In: Voß, Jan-Peter. and Bauknecht, Dieter. and Kemp, René. Eds. *Reflexive Governance for Sustainable Development*. Cheltenham: Edward Elgar.

Wollmann, Hellmut. 2006. Evaluation and evaluation research. In Fischer, Frank, Miller, Gerald J., Sidney Mara S. *Handbook of Public policy analysis*. London: Routledge.

13. Case Studies

VALÉRY RIDDE, ABDOURAHMANE COULIBALY, AND LARA GAUTIER

Abstract

Case studies consist of an in-depth analysis of one or more cases, using a variety of methods and theoretical approaches. The choice of cases (single or multiple) studied is crucial. Case studies are particularly suitable for studying the emergence and processes involved in policy implementation and for contributing to theory-based evaluations.

Keywords: Qualitative methods, quantitative methods, mixed methods, case study, theoretical approaches, single/multiple cases, empirical triangulation, analytical generalisation

I. What does this method consist of?

Also used in anthropology, the case study approach has long been used in evaluation, where it is considered not as a method but as a research strategy (Yin 2018). By studying a policy in context and using multiple lines of evidence, the case study (single or multiple) seeks to answer 'how' and 'why' questions from a systems approach and with the support of theoretical approaches. Conducting a case study for a public policy evaluation follows a standard evaluation process: planning, drafting the protocol, preparing the field, collecting and analysing data, sharing results and making recommendations for policy improvement (Gagnon 2012). As with all evaluations, the choice of methods should follow the objectives and the evaluation question, not the other way around. A case study may thus mobilise qualitative, quantitative and different mixed methods designs.

The case study strategy is therefore appropriate when organising an evaluation of policy emergence, process, relevance or adaptation. It is often mobilised when evaluation teams have little or no control over the events and context that influence policy actions. This is often the case outside of experimental situations, which are rare in the field of public policy. It is therefore mostly recommended for understanding a contemporary, often complex, phenomenon organised in a real context.

The case study approach can be used to explain a public policy, describe it in depth or illustrate a specific situation, which can sometimes be original and enlightening for decision-making. The advantage of case studies is that they can be adapted to different situations where there are multiple variables of interest around a policy. It is also about being able to use multiple sources of data, both quantitative and qualitative, which allow for empirical triangulation. The case study strategy allows theoretical propositions and the state of scientific knowledge to guide data collection and analysis. It fits perfectly with, but is not limited to, theory-based evaluation approaches (see separate chapter on theory-based evaluation).

There are a myriad of proposals for the types of case studies that are possible. Firstly, it is possible to use single/single case studies (involving one policy) or multiple case studies (several policies in the same organisational context or one policy in different contexts). Secondly, these cases can be studied holistically (the policy as a whole) or at different levels of analysis (the dimensions of the policy that the intervention theory will have specified or the particular regional contexts). The choice of case studies should be heuristic (to learn from the study) and strategic (to have data available within the available budget, to answer useful questions). A key criterion for case selection is to have sufficiently relevant information to understand the policy in depth and complexity. Case sampling should therefore be explicit, rigorous and transparent. The selection of case studies can thus be critical, unique, typical, revealing, instrumental, etc. This selection can also be carried out in collaboration between the research and policy teams to ensure that the choices are relevant and feasible. The selection can also be based on prior quantitative analyses to obtain the starting situation of the cases and, for example, choose cases that are very contrasting or very similar in their performance with regard to the policy being analysed.

Sometimes it can also be useful to have a diachronic approach in order to produce longitudinal case studies. For example, analysing a policy over time can reveal the influences of changes in the context or in the strategies of those implementing it, or of those benefiting from it. Starting with cases with similar initial conditions and then studying their evolution is referred to as 'racing cases' by Eisenhardt (Gehman et al. 2018).

When analysing the data, the case study approach requires, in addition to the usual analyses specific to the methods (content analysis, thematic analysis, descriptive or inferential statistics, etc.), to mobilise a replication logic. The idea is to compare, in a systematic and rigorous way, the empirical data and the theory, be it the theory of the policy intervention or a theoretical or conceptual framework used to understand the policy. This process is referred to by Yin as analytical generalisation. When several cases support the same theory, it is possible to suggest the presence of a replication logic (Yin 2010).

Configurations can be heuristic tools for this analysis, whether they are organisational or rooted in critical realism (see separate chapter on realistic evaluation). Furthermore, finding similar patterns, or situations, in different contexts strengthens the ability to generalise the results of case studies. Yin believes that analytical generalisation requires the construction of a very strong case that will be able to withstand the challenges of logical analysis. Thus, it is essential to specify this theoretical rationale at the outset of the case study, either by mobilising a theory or from the state of the art without it being entirely specific to the public policy being analysed. At the beginning of a case study, it is therefore necessary to remain at a relatively high conceptual level, at least higher than the policy under study. Secondly, the empirical results of the case study must show how they align (or not) with the theoretical argument at the outset. Finally, it will be necessary to discuss how this theoretical thinking, based on this particular policy, can also be applied to other situations and policies in the particular case study. The fact that, even at the beginning of the case study, a counter-argument (rival hypotheses) was also formulated, and that empirical evidence was sought during the data collection process (which refutes them), reinforces the validity of this process of analytical generalisation. Finally, the power of multiple case studies is that this analytical generalisation is strengthened when the results of one case are similar to those of other cases.

Some research teams even propose that case studies can lead to theory-building, especially when analysing complex objects such as public policies.

II. How is this method useful for policy evaluation?

Before deciding to embark on a case study approach, two preliminary questions should be asked which will determine the appropriateness of the approach:

1. Does the phenomenon I am interested in need the case(s) to be understandable? (e.g., Theory-building case studies)
2. Does the case(s) represent an empirical window that informs the analysis of the wider phenomenon?

Once one or the other has been answered positively, the evaluative questions can be defined:

- Under what real-life conditions can public policy X, piloted in context A, be scaled up in contexts B, C, and D?
- How did the controversy about public policy Y in context B emerge?
- What are the success factors for the implementation of public policy X in context A?
- How were public policies Y and Z implemented in context B?
- Why did public policy X in context A and B fail, while it had positive effects in context C?
- Why did public policy X implemented in context A fail, while public policy Y implemented in the same context A succeeded?
- What is it about the characteristics of public policy Z implemented in contexts A, B, and C that informs μ theory-building case studies?

The case study can be used at any point in the evaluation process, ex ante (at the time of policy design), in itinere (during implementation), or ex post (e.g. to better understand the results produced).

III. An example of the use of this method in Burkina Faso

Simple and multiple longitudinal case studies were mobilised to study a public health financing policy in Burkina Faso (Ridde 2021).

The World Bank encouraged the government to test in a dozen districts a modality for financing health centres in addition to the state budget. The idea was to organise a performance-based payment system in which health centres and health professionals received additional funds based on the achievement of activity results. For example, for each delivery performed in the centre with a partographer, they received 3.2 euros to be shared between the structure and the staff, according to complex procedures and indicators. Verification and control processes were organised to ensure the reliability of payment claims.

To study the emergence of this new policy, we conducted a single case study (focusing on the policy) to better understand its origin, ideas, proposed solutions, people who proposed it, power issues, etc. We employed a literature review and 14 qualitative in-depth interviews with policy makers, funding agencies and experts on the subject. Using an analytical generalisation approach, we compared this emergence to understand whether what happened in Burkina Faso was also happening in Benin.

To study the implementation of the policy in Burkina Faso, we then used multiple longitudinal case studies. For reasons of time and budget, we selected three districts representing the diversity of situations in which the policy was implemented. Then, within each of these districts, we selected six cases from the primary health centres (about 30 per district) and one case that was the referral hospital (only one per district). The six cases were selected according to the three types of financing strategies that the policy wished to test, so two cases per type. We decided to select two cases with the greatest possible contrast within each of the three types: one very performant health centre and one not at all. Performance was calculated using a quantitative method (time series) on the basis of indicators of health centre attendance in the years preceding the policy. This etic analysis (from the external perspective) ranked all the health centres according to their order of performance to support case selection. The latter also benefited from the emic opinion (from the internal point of view) of local health system managers in order to take into account their own perception of the performance of the centres, beyond the quantitative approach which only gives a partial view of performance. Thus, for each of the seven cases selected per district (7×3 = 21), we used multiple sources of data to understand the challenges of policy implementation: analysis of documentation, formal qualitative interviews (between 114 and 215 per district) and informal interviews (between 26 and 168 per district), and observations of situations. A data collection grid was also used to measure the fidelity of policy implementation. In order to better understand the evolution of policy implementation, and in particular adaptations over time, three data collection moments were carried out over a 24-month period, thus following the longitudinal multiple case study approach.

Finally, these case studies have also been fruitful in studying, with a qualitative approach and a long immersion in the field, the unexpected consequences (positive or negative) of this policy. Although this dimension of the evaluation is still too little understood, its implementation in Burkina Faso has shown the relevance of this approach (Turcotte-Tremblay et al. 2017). Limiting oneself to the expected effects, which is often implied by an extreme focus on the sole theory of intervention developed by the teams that define the policy, reduces the heuristic scope of the evaluation. While successes are essential, challenges may also be necessary to improve public policies with the help of case studies.

For all these approaches, the analysis was carried out in a hybrid manner, both deductive (with respect to the intervention theory or a conceptual framework) and inductive (original empirical data). The comparison between cases, between districts and between countries allowed for an increase in abstraction in an analytical generalisation process.

IV. What are the criteria for judging the quality of the mobilisation of this method?

Judging the quality of a complex approach such as case studies requires a global vision, going beyond the specific but essential reflections of the usual methods (quantitative and qualitative). To this end, Yin (2018) proposes to study the quality of case studies in terms of four dimensions:

- Construct validity (studying the expected policy and not something else): using multiple sources of evidence, describing and establishing a causal chain, involving stakeholders in the validation of the protocol and reports;

- Internal validity (confidence in results): compare empirical data with each other and with theory, construct explanatory logics, account for competing and alternative hypotheses, use logical frameworks/ theories of intervention;

- External validity (ability to generalise results): use theories, use the logic of analytical replication;

- Reliability (for the same case study, the same findings): use a policy study protocol, develop a case database.

V. What are the strengths and limitations of this method compared to others?

The main strength of the case study is its ability to 'incorporate the unique characteristics of each case and to examine complex phenomena in their context', i.e. in real-life conditions (Stiles 2013, 30).

The case study strategy, due to the abundance and variety of the corpus of data mobilised, and the research methods employed (qualitative, quantitative or mixed), most often allows for a rich description of the public policy(ies) being evaluated and the contexts of implementation. This is particularly true of single case studies, which allow for in-depth analysis. With regard to multiple case studies, the main advantage is that it allows for more potential variation, which increases the robustness of the explanation. The downside is that these strategies require a significant time commitment. Thus, the sheer volume of work can be problematic, especially if the deadlines set by the sponsors are short. In addition, if there are several evaluative questions, or a question that invites the linking of implementation issues to outcomes, then it may be necessary to consider combining the case study (which may focus on process analysis, for example) with another complementary research strategy, such as quasi-experimental approaches (Yin and Ridde, 2012). Finally, several biases may arise – the biased choice of case(s), low statistical power when conducting quantitative analyses. These biases may erode comparability across cases or contexts. The rich justification of the choice of cases (public policies) (Stake 1995) and the description of the context(s), as well as the process of analytical generalisation, described above, help to reduce the impact of these biases.

With regard to theory-building case studies, both advantages and disadvantages of the case study are identified (Stiles 2013). The case study strategy here consists of comparing different statements from theory with one or more observations. This can be done by describing the few cases in theoretical terms. Thus, although each detail can only be observed once, they can be very numerous and therefore useful for theory building. However, the same biases mentioned above are likely to occur (biased case selection, low statistical power). Confidence in individual statements may be eroded by these biases. On the other hand, as many statements are examined – reflecting a variety of contexts and therefore possible variations – the overall strengthening of confidence in the theory may be just as important as in a hypothesis testing study.

Some bibliographical references to go further

Gagnon, Yves-Chantal. 2012. *L'étude de cas comme méthode de recherche*. 2nd ed. Québec: Presses de l'Université du Québec.

Gehman, Joel. and Glaser, Vern L.. and Eisenhardt, Kathleen M.. and Gioia, Denny. and Langley, Ann. and Corley, Kevin G.. 2018. "Finding Theory–Method Fit: A Comparison of Three Qualitative Approaches to Theory Building." *Journal of Management Inquiry*, 27(3): 284-300. https://doi.org/10.1177/1056492617706029.

Ridde, Valéry, éd. 2021. *Vers une couverture sanitaire universelle en 2030?* Éditions science et bien commun. Québec: Canada: Zenodo. https://doi.org/10.5281/ZENODO.5166925.

Stake, Robert E. 1995. *The Art of Case Study Research*. Thousand Oaks, CA: SAGE Publications.

Stiles, William B. 2013. "Using Case Studies to Build Psychotherapeutic Theories." *Psychothérapies*, 33(1): 29-35. https://doi.org/10.3917/psys.131.0029.

Turcotte-Tremblay, Anne-Marie. and Ali Gali-Gali, Idriss. and De Allegri, Manuela. and Ridde, Valéry. 2017. "The Unintended Consequences of Community Verifications for Performance-Based Financing in Burkina Faso." *Social Science & Medicine*, 191: 226-36. https://doi.org/10.1016/j.socscimed.2017.09.007.

Yin, Robert K. 2010. "Analytic Generalization." In *Encyclopedia of Case Study Research*, by Albert Mills, Gabrielle Durepos, and Elden Wiebe, 6. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc. https://doi.org/10.4135/9781412957397.n8.

Yin, Robert K. 2018. *Case study research and applications: design and methods*. Sixth edition. Los Angeles: SAGE.

14. Process Tracing

ESTELLE RAIMONDO

Abstract

Process tracing is a theory-based evaluation approach. Based on the formulation of a process theory-of-change, it collects evidence to ascertain how the intervention unfolded in a single case and whether it plausibly contributed to change in outcomes. Often described as a qualitative method, process tracing can in fact rely on a diversity of qualitative and quantitative methods. Particularly useful to evaluate complex interventions, it addresses the questions of "under what conditions, how, and why" an intervention worked, rather than how much impact it produced.

Keywords: Qualitative methods, processes, theory-based evaluation, process theory-of-change, causal principles, contribution pathways, evidence, fingerprints, Bayesian reasoning

I. What does this approach consist of?

When evaluators conduct Process Tracing (PT), they behave a bit like "detectives". When applying process tracing, evaluators are interested in explaining, rather than simply describing, change processes. To put it simply, evaluators seek to trace how the activities of actors/entities and their motivation are interlinked to trigger change in the behavior and action of others. Empirically, process tracing is also akin to "detective work" as it consists of assembling a body of evidence (what D. Beach calls 'fingerprints') to ascertain how the intervention unfolded in a single case and whether it plausibly contributed to change in outcomes. In slightly more technical terms, process tracing is a theory-based evaluation approach for studying how interventions worked in actual cases (see separate chapter on theory-based evaluation). As such, process tracing belongs to the family of methods that seek to answer the questions of "how, why and under what circumstances" programs and policies work by studying how they play out in the real-world. Visually, process tracing seeks to understand what is going on 'in-between' the arrow linking interventions and results, in a typical theory of change. Its comparative advantage over other methods is in fully opening the black box of change processes.

Process tracing is often considered a "qualitative" approach because it tends to rely on qualitative evidence (from interviews, observations, documents, etc.) but, like many other theory-based evaluation approaches, it resists simple classification and is better described as 'methods agnostic'. It can accommodate and use a range of methods of data collection and analysis, quantitative or qualitative, in seeking to assemble a body of evidence that is robust enough to adjudicate between the process theory of change under scrutiny and alternative explanations. In addition, more recently, some evaluators have mathematically formalized the use of process tracing through the application of Bayes' theorem (Befani 2021)

At its core, there are two main phases and a few unique features to process tracing that distinguish it from other theory-based evaluations which we will highlight briefly.

I.I. The first phase of process tracing consists of formulating a process theory-of-change (pTOC).

A pTOC is a detailed theory of how an intervention produced a contribution to an outcome of interest. It means unpacking the activities of actors/entities which together constitute the inner working of programs (the arrow). Actors are the people or organizations doing things, whereas actions are what they are doing. Understanding why the actions of one actor led other actors to do things requires trying to make as explicit as possible what Cartwright and Hardie (2012) term the causal principles.

To do that, initially, the evaluators brainstorm about what 'contribution' might have realistically been produced by an intervention and start laying out plausible contribution pathways between them. This might mean drawing from existing theoretical literatures in the social sciences on the topic or from repositories of evaluative evidence in the grey literature. It also means exploiting program and policy documents. In this sense, process tracing does not limit its investigation to the stated policy goals but to plausible intended or unintended pathways to outcomes of interest.

When determining what contribution might have been produced by an intervention, it is also important to explore competing explanations outside of the scope of program activities that could also account for the outcomes.

The number of details provided in a pToC varies. A more detailed pToC is required when the evaluation seeks to produce actionable knowledge that can help with project implementation. In contrast, if the goal is to understand how a type of intervention works across several cases, a simplified, mid-range pToC can be sufficient.

I.II. The second phase consists in testing the pTOC empirically and figuring out how it actually worked in a case.

Process tracing seeks to test and refine its theory by observing how the intervention worked in a single case. In process tracing a granular pToC is used as a scaffolding for the empirical assessment of how a contribution was actually produced. This means that before engaging in actual data collection, evaluators must anticipate the type of plausible “fingerprints” left by the change mechanism and figure out the type of evidence they need or want to see to boost their confidence in their theory. There are two types of useful evidence that the evaluators are looking for. Some evidence “need to be found” to avoid decreasing evaluators’ confidence/ disconfirming the pToC (sometimes called “hoop test”). Evaluators are also seeking evidence that they would “love to find” to significantly boost their confidence in the pToC (sometimes called “smoking gun test”).

When thinking about the evidence evaluators would need/love to see, they should cast the net widely in a search for a variety of different potential “fingerprints”. In process tracing, each individual piece of evidence typically tells us little, but combined, they might act as a unique, confirmatory signature that a given action and linkage took place in the case. Working with evidence therefore often involves a form of bricolage (for more on this, see Beach and Pedersen, 2019: 232-233).

Once the data collection has started, a critical assessment of the observations and evidence must take place. Bayesian reasoning is often used as the logical framework to assess the strength (probative value) of the evidence, either in an informal way, similar to how Bayesian reasoning informs evidence evaluation in criminal investigations (e.g. Beach and Pedersen, 2019), or more formally through the application of Bayes’ theorem and estimation of probabilities of finding/not finding evidence (e.g. Befani and Stedman-Pryce, 2017). Essentially, evaluators conducting process tracing must ask the following questions:

- If expected “fingerprints” are not found, did we have full access to the empirical record, and can we trust that our sources were not hiding something from us?
- If expected “fingerprints” are found, have we interpreted what our sources have told us correctly in this context, and can we trust them?

II. How is this approach useful for policy evaluation?

When process tracing made its way into evaluation practice, the field of impact evaluation had been dominated by (quasi-)experimental approaches with strong comparative advantages in establishing the average treatment effect of relatively straightforward interventions whose effect could be measured quantitatively. However, the need to expand the evaluators’ toolbox to other approaches that could answer different types of impact evaluation questions and investigate interventions that were more complex and less amenable to quantification and controlled comparisons became increasingly pressing. Process tracing emerged as a useful approach for evaluations that seek to explain change processes and are less concerned with the question of “how much” an intervention impacted a desired outcome, and more with understanding “under what conditions, how, and why” an intervention worked in the real-world.

Process tracing has been used to assess the impact of a range of interventions, but has a comparative advantage over other methods in studying ‘intangible’ or ‘soft’ interventions, such as the influence of knowledge and data work, advocacy and communication campaign, policy dialogue on decision-making, etc. It also works well to assess the impact of interventions that target behavioral changes among participants through sensitization and incentives mechanisms.

Process tracing can be used to serve various decision needs, but it fits particularly well for the adaptive management of interventions, when seeking to test and refine implementation modalities in various contexts. It can also be useful to use process tracing during a piloting or scale-up phase, to gauge whether the change mechanisms are triggered when interventions are replicated or scaled up. It tends to work well as an embedded or retrospective approach.

III. Examples of the use of this approach in the field of development

A few examples of real-world applications of process tracing in evaluation primarily drawn from the development evaluation include: the use of process tracing to assess the sustainability of budget support interventions (Orth et al. 2017), to study the impact of advocacy campaigns on the preservation of biodiversity (D'Errico, et al. 2017), and to understand the contribution of citizen engagement mechanisms in the improvement of public service delivery in the Dominican Republic (Raimondo, 2020).

In this latter example, the evaluation sought to respond to the intensification of aid agencies' efforts to put citizens front and center in defining their development agenda. The World Bank decided in 2014 to mainstream citizen engagement activities in all of its projects where direct beneficiaries could be identified. In making this policy commitment, the World Bank claimed that engaging citizens was not only the "right" thing to do, but it was also going to improve the effectiveness of its projects. The evaluation selected a typical case of using citizen engagement mechanisms to improve the delivery of health and education services for poor households in the Dominican Republic to test that claim. Unpacking and testing the causal mechanism underlying citizen engagement activities certainly enhanced the evaluation team's understanding of the behavioral, operational, and institutional inner workings of the intervention and the conditions under which citizen engagement could transmit causal power to change the quality of services. Based on this granular understanding, the evaluation made practical recommendations to the program in terms of how meetings with citizens should be facilitated and by whom to ensure an effective feedback loop and service improvement. However, process tracing needed to be complemented with cross-case comparisons to enhance the generalizability of the findings and their policy relevance for the entire program, which was implemented across regions.

IV. What are the criteria for judging the quality of the mobilisation of this approach?

The quality of process tracing's implementation hinges on how well theory and empirics are brought together. To arrive at a process tracing with high internal validity, the three following criteria should be kept in mind: (1) a more disaggregated and fine-tuned pToC that captures key episodes and mechanisms; (2) evidence that is highly unique found for each part of the pToC; (3) trustworthy sources and full access to the empirical record. On the other hand, if the pToC is too simple or abstract, if the evidence found is not unique or could be found for other explanations, or if the sources are too weak or not trustworthy, the internal validity will be low.

For some evaluations, it is also important for the lessons drawn from process tracing to travel to other contexts. Process tracing on its own does not have high external validity, but by combining it with cross-case comparisons it is possible to explore whether similar processes also work in other cases across contexts.

V. What are the strengths and limitations of this approach compared to others?

Key strengths of the approach when well implemented:

- If the three quality criteria laid out above are met, then applying process tracing significantly bolsters our capacity to establish a strong causal link between interventions and outcomes and at the same time have strong explanatory power behind the 'how' and 'why' of processes of change.
- Process tracing provides a clear scaffold for making transparent the process of evidence gathering and assessment as well as triangulating sources of evidence. This process goes far beyond typical case study approaches, and other theory-based approaches. Process tracing makes the theory of change vividly unfold in front of the eyes of the evaluator and allows them to reach strong confidence in their impact/contribution claims.
- It is also much easier to derive 'practical lessons' from a process tracing study than from many other types of evaluation approaches. Because it focuses the evaluator's mind on causal explanations and the linkages between actions and behavior change, it helps elaborate ideas about how such activities should be tweaked or changed to improve outcomes.

- Process tracing has a comparative advantage over other (impact) evaluation methods in assessing interventions that are not amenable to quantification or experimentation, such as policy dialogue, the contribution of research, knowledge and data work, advocacy and communication campaigns, etc.

Some (de)limitations of the approach:

- Process tracing is not adequate to answer 'how much of an impact an intervention had on average on an outcome of interest' and should not be used to fulfill this objective.
- While it needs not be overly technical, there is a steep learning curve to mastering the ropes of process tracing. Notably, evaluators need to become familiar with setting up 'empirical tests' to gauge the probative value (uniqueness and trustworthiness) of their evidence; they need to become more rigorous in how they reconstruct process Theory-of-change and leverage the existing literature to theorize about behavioral change linked to specific actions, etc.
- On its own process tracing has weak external validity and needs to be paired with a cross-case design, which can become onerous and time-consuming.

Some bibliographical references to go further

Beach, Derek. and Brun Pedersen, Rasmus. 2019. *Process Tracing Methods*. Ann Arbor: University of Michigan Press.

Befani, Barbara. and Stedman-Bryce, Gavin. 2017. "Process Tracing and Bayesian updating for impact evaluation". *Evaluation*, 23(1): 42-60.

Befani, Barbara. 2021. *Credible Explanations of Development Outcomes: Improving Quality and Rigour with Bayesian Theory-Based Evaluation*. Report 2021:03, Expert Group for Aid Studies (EBA), Sweden.

Cartwright, Nancy. and Hardie, Jeremy. 2012. *Evidence-based policy*: *A practical guide to doing it better*. Oxford: Oxford University Press.

D'Errico Stefano. and Befani, Barbara. and Booker, Francesca. and Guiliani, Alessandra. 2017. *Influencing policy change in Uganda: an impact evaluation of the Uganda Poverty and Conservation Learning Group's work*. PCLG Research Report. https://www.iied.org/sites/default/files/pdfs/migrate/G04157.pdf

Orth, Magdalena. and Schmitt, Johannes. and Krisch, Franziska. and Oltsch, Stefan. 2017. *What we know about the effectiveness of budget support*. Evaluation Synthesis, German Institute for Development Evaluation (DEval), Bonn.

Raimondo, Estelle. 2020. "Getting Practical with Causal Mechanisms: The Application of Process-Tracing under Real-World Evaluation Constraints." *New Directions for Evaluation*, Fall 2020: 45-58.

15. Comparative Historical Analysis

EMANUELE FERRAGINA

Abstract

Comparative historical analysis combines two major methodological tools of social science, comparison (the study of similarities and differences across cases) and history (the analysis of processes of change in their temporal dimension), to help explain large scale outcomes on a variety of topics. It is particularly useful to account for the definition of public policies (policy framing and policy change).

Keywords: Mixed methods, qualitative methods, historical analysis, similarities, differences, history, macro, comparison, critical junctures, path dependency

I. What does this approach consist of?

Comparative historical analysis (CHA) is more an approach than a method, and it is rooted in a long history from old seminal works, e.g. De la Démocratie en Amérique (Tocqueville 1960) and The Protestant Ethic and the Spirit of Capitalism (Weber 2001) to modern classics, e.g. The Social Origins of Dictatorship and Democracy (Moore 1966) and States and Social Revolutions (Skocpol 1979). The historical approach in social sciences offers explanations of large-scale outcomes on a wide range of topics, such as revolutions, the advent of democratic or authoritarian rule, path dependent institutional processes, policy continuity and change in various domains. This approach has several distinctive characteristics that have fostered its extensive use in social science research and public policy.

CHA explores similarities and differences across different cases – recalling John Stuart Mill's method of agreement and difference – with the aim to unveil causal mechanisms that determine specific outcomes (see separate chapter on case studies). Processes of change and their temporal dimension are at the core of sociology and political science, and for this reason CHA helped the identification of the origin of specific reforms, or the point of departure for significant institutional change. The cases analysed are often nation-states, but other entities (such as regions, social movements and organisations) have also been scrutinised (for an example of regional analysis, see Ferragina 2012; 2013). This approach attributes a big role to theory, and a very interesting debate has taken place on the American Journal of Sociology, with a symposium comparing the place assigned to theory in historical sociology and rational choice theory: "we're no angels: realism, rational choice, and relationality in social science" (see the contributions to this debate of Somers 1998; Kiser and Hechter 1998; Goldstone 1998; Calhoun 1998). The debate contrasted the use of these different perspectives, highlighting that CHA helps to test and generate theory through a macro-configurational, case-based and temporally-oriented approach.

The macro component concerns large-scale outcomes, i.e. state building, democratic transitions, societal patterns of inequality, war and peace. Researchers focus on large-scale causal factors, including both political-economic structures (e.g. colonialism) and complex organisational institutional arrangements (e.g. social policy regimes). This macro approach can also explain micro-level events and processes that should (or should not) be present within particular cases if the macro theory is correct. The configurational component refers to the way in which researchers consider how multiple factors combine to form coherent causal packages. One for example cannot study revolutions without analysing how various events and underlying processes constitute these social phenomena. Even when CHA scholars are interested in studying the effects of a specific variable they care a lot about the context and other potential causes.

Differently from other techniques commonly used in social science, CHA does not shy away from complex questions for which data are not readily available. One of the most regrettable trends in social sciences is the selection of questions on the basis of available data. As in the Nietzschean metaphor, it is as if researchers are like drunk people who search their lost keys only under the lamppost. For this reason, CHA focuses on real world puzzles and uses mechanisms-based explanations, following questions of this kind: why do cases that are similar on many key dimensions exhibit different outcomes on a dependent variable of interest? Or alternatively, why do seemingly disparate cases all have the same outcome? Moreover, real world puzzles may also be formulated when particular cases do not conform to expectations from existing theory or large-N research. CHA places emphasis on developing a deep understanding of the cases to adjudicate competing hypotheses.

Without the pretension of being exhaustive, it is important to mention here the most used conceptual tools in CHA, that is critical junctures, path dependency and other devices to capture gradual change. Collier and Collier (1991: 29) have defined critical junctures as periods of significant change that occur producing durable effects. Critical junctures unsettle previous institutional patterns and open to a new period of path dependency. Path dependency indicates that when a nation or another macro-unit of analysis has started to move in one direction, the costs to revert the trajectory are very high and this contributes to a sort of inertia that can be broken again only with a new critical juncture (Pierson 2004). In simple terms: history matters.

While critical junctures and path dependency are used to describe the succession of radical change and stability, other conceptual tools indicate the presence of a gradual change that can progressively produce conspicuous change. Streeck and Thelen (2005) classified this form of change into five categories: Displacement, that is when a traditional institutional structure is progressively discredited and put at the margins in favour of those that are more apt to satisfy present needs. Layering, that is when new elements are progressively added to the old structure. This form of institutional change is often observed in social policy, for example in the field of labour market and family policy (Daly and Ferragina 2018). Institutional change can also happen just because an institution becomes obsolete to respond to its original aims as it has not been adequately updated over time (this form of institutional change is called drift (Hacker 2004). Another form of institutional change is that of conversion, that is when an existing institution is redirected towards new objectives. A last form is that of exhaustion, that brings the institution to a progressive disappearance.

II. How is this approach useful for policy evaluation?

CHA can be employed to understand how to set up a policy evaluation study, recognize the origins of specific policies, better understand the context within which policies and outcomes change, and observe an institutional trajectory in the long run. In a nutshell, a CHA can help to situate specific policy evaluations within a context, illustrating for example the concatenation of policy changes that bring to a fundamental institutional change in the long run (in this respect see the example below about 'selective neoliberalism'). Major works that absolve these functions in the literature include The Three Worlds of Welfare Capitalism (Esping-Andersen 1990), Development and Crisis of the Welfare State (Huber and Stephens 2001), Dismantling the Welfare State? Reagan, Thatcher, and the Politics of Retrenchment (Pierson 1994), and Protecting Soldiers and Mothers: The Political Origins of Social Policy in the United States (Skocpol 1992).

III. Disentangling the direction of social policy reforms in the long run: the case of 'selective neoliberalism'

CHA can be employed to disentangle how several reforms might lead to specific outcomes, linking a theoretical concept to the exploration of policy change. This is the case of a study published in New Political Economy that explores how Italy progressively liberalised pension and labour market policies in different steps (Ferragina and Arrigoni 2021); if one analyses reforms in isolation, one cannot correctly observe the comprehensive design of the liberalisation process. This means that an historical analysis might allow us to discern the entire reform process. The study, although only analysing the Italian case, is based on the comparison with other European countries through the framing of the passage from the Fordist to the neoliberal phase of capitalism. More specifically this research illustrates the Italian process of neoliberal institutional adaptation in the main social policy reforms, and suggests that over three decades this process took place selectively. Selective neoliberalism is defined as a modality of institutional adaptation which started from the margins and then expanded to the rest of society.

Selective neoliberalism resulted from a reform process begun in the early 1990s when a neoliberal turn was set in motion (Ferragina et al. 2022). The reform process, with continuity between centre-right and centre-left coalitions, circumvented the resistance of trade unions against an overall social policy liberalisation, hitting first social groups without sufficient power resources to defend their social entitlements and rights. This modality of institutional adaptation can be observed in both labour market and pension reforms.

Through the concept of selective neoliberalism, the initial dualization of social entitlements and rights in the Italian case is interpreted as an intermediary step toward liberalisation (for a discussion see Streeck 2009, Emmenegger 2014). This argument is substantiated with an analysis of the continuity in the social policy reforms, and through insights from comparative historical analysis. Neoliberal ideas, promoted originally by Einaudi in the first part of the twentieth century and kept alive in intellectual circles in the post WWII period, re-emerged like a subterranean river when the international political economy context had turned globally away from Keynesianism. The spread of neoliberal ideas influenced Italian technocratic elites at the Bank of Italy and the Treasury, and also the internal debate of the Socialist (PSI) and Christian Democratic (DC) parties since the 1980s.

The research sequences the 'roll back' of Fordism and the 'roll out' of neoliberalism, and through this historical institutional analysis, it identifies a neoliberal turn in 1992. Different streams of literature have emphasised this year's importance for Italy – which can be regarded as a sliding door on the institutional, economic, and political levels. The notion of critical juncture is used to illustrate how after 1992, the institutional equilibrium was broken; and this gave way to a series of reforms very much at odds with the past. From a methodological perspective 'junctures are "critical" because they place institutional arrangements on paths or trajectories, which are then very difficult to alter' (Pierson 2004: 135). This analytical tool helps to identify a transition from Fordism to neoliberalism as portrayed in the international political economy literature. Then, the concept of selective neoliberalism helps to interpret the labour market and pension reforms holistically. This notion can be applied to other countries and policy contexts, in particular where a strong resistance of veto players is undermined through an incremental reform process that contributes to a neoliberal adaptation.

IV. What are the strengths and limitations of this approach compared to others?

CHA presents advantages and disadvantages in comparison to other methods and approaches. It is unique in helping to address big questions and the analysis of political processes, allowing it to systematically disentangle complex reform processes as we have shown with the example of selective neoliberalism. The application of an historical approach allows one to consider with care the specificity of cases, observe their long term development, proposing in the end contingent generalisations. However, CHA also presents several limits. The approach does not propose a systematic way to approach problems as other methods of analysis. It is difficult to select cases when testing theories, and generalisation, although possible, has to be contingent and limited (because of the small-N). Moreover, this approach can be criticised from a historical point of view, because it is often based on secondary sources rather than archival material.

Other big questions remain open for scholars and students who are willing to employ this approach in the future. How to deal with the tension between structure and agency? Approaching big questions is very important, but CHA does not offer much space to the role of actors and is prevalently concerned with structural change. There are also epistemological questions regarding the tension between the contingent generalisation and the respect of the cases analysed. Almost sixty years ago, Moore (1966: XIV) described this problem with acumen:

> Nevertheless there remains a strong tension between the demands of doing justice to the explanation of a particular case and the search for generalisations, mainly because it is impossible to know just how important a particular problem may be until one has finished examining all of them.

Some bibliographical references to go further

Capoccia, Giovanni. and Kelemen, R. Daniel. 2007. The study of critical junctures: Theory, narrative, and counterfactuals in historical institutionalism. *World politics*, 59(3): 341-369. This article provides a complete analysis of critical junctures. Critical junctures place institutional arrangements on paths or trajectories, which are very difficult to alter.

Mahoney, James. and Thelen, Kathleen. (Eds.). 2015. *Advances in Comparative-Historical Analysis*. Cambridge: Cambridge University Press. This edited book covers multiple uses of comparative historical analysis in political science. It includes contributions from leading authors in the field and discusses the broad agenda of CHA through an analysis of fundamental works, the tools for temporal analysis (such as path dependence and critical junctures), and important methodological developments.

Moore, Barrington. Jr. 1966. *Social Origins of Democracy and Dictatorship: Lord and Peasant in the Making of the Modern World.* Boston: Beacon Press. This seminal book explains the varied political roles played by the landed upper class and the peasantry in the transformation from agrarian societies to modern industrial ones. From a methodological perspective Moore highlights the strong tension between the demands of doing justice to the explanation of a particular case and the search for generalisations. A starting point for all those interested in CHA.

Pierson, Paul. 2004. *Politics in Time: History, Institutions, and Social Analysis*. Princeton: Princeton University Press. The book presents a detailed analysis of the importance of time to understand institutional and social change, providing a methodological backing to the classic statement that history matters. Pierson suggests using comparative historical analysis to move beyond a static view of institutional change.

Skocpol, Theda. 1979. *States and social revolutions: A comparative analysis of France, Russia and China*. Cambridge: Cambridge University Press. According to Skocpol, social revolutions deserve special attention because of their extraordinary significance for the history of nations and their distinctive pattern of socio-political change. What is unique to social revolutions is that basic changes in social and political structure occur together in a mutually reinforcing fashion. To analyse these important historical events Skocpol set a comparative historical analysis of France, Russia and China. This book is a reference for those who want to apply comparative historical analysis to large scale social phenomena.

Cited references

Calhoun, Craig. 1998. Explanation in historical sociology: Narrative, general theory, and historically specific theory. *American journal of sociology*, 104(3): 846-871.

Collier, Ruth Berins. and Collier, David. 1991. *Shaping the political arena: Critical junctures, the labor movement, and regime dynamics in Latin America*. Princeton: Princeton University Press.

Daly, Mary. and Ferragina, Emanuele. 2018. Family policy in high-income countries: Five decades of development. *Journal of European Social Policy*, 28(3): 255-270.

Emmenegger, Patrick. 2014. *The power to dismiss: trade unions and the regulation of job security in Western Europe*. Oxford: Oxford University Press.

Esping-Andersen, Gosta. 1990. *Three Worlds of Welfare Capitalism*. Cambridge: Polity.

Ferragina, Emanuele. 2012. *Social capital in Europe: A comparative regional analysis*. Edward Elgar.

Ferragina, Emanuele. 2013. The socio-economic determinants of social capital and the mediating effect of history: Making Democracy Work revisited. International *Journal of Comparative Sociology*, 54(1): 48-73.

Ferragina, Emanuele. and Arrigoni, Alessandro. 2021. Selective neoliberalism: How Italy went from dualization to liberalisation in labour market and pension reforms. *New Political Economy*, 26(6): 964-984.

Ferragina, Emanuele. and Arrigoni, Alessandro. and Spreckelsen, Thees. 2022. The rising invisible majority: Bringing society back into political economy. *Review of International Political Economy*, 29(1): 114-151.

Goldstone, Jack. 1998. Initial conditions, general laws, path dependence, and explanation in historical sociology. *American journal of sociology*, 104(3): 829-845.

Hacker, Jacob 2004. Privatizing risk without privatizing the welfare state: The hidden politics of social policy retrenchment in the United States. *American Political Science Review*, 98(2): 243-260.

Huber, Evelyne. and Stephens, John. 2001. *Development and crisis of the welfare state. Parties and policies in global markets*. Chicago: Chicago University Press.

Kiser, Edgar. and Hechter, Michael. 1998. The debate on historical sociology: Rational choice theory and its critics. *American Journal of Sociology*, 104(3): 785-816.

Moore, Barrington, Jr. 1966. *Social Origins of Democracy and Dictatorship: Lord and Peasant in the Making of the Modern World*. Boston: Beacon Press.

Pierson, Paul. 1994. *Dismantling the Welfare State? Reagan, Thatcher, and the Politics of Retrenchment*. Cambridge: Cambridge University Press.

Pierson, Paul. 2004. *Politics in time: history, institutions, and social analysis*. Princeton: Princeton University Press.

Skocpol, Theda. 1979. *States and social revolutions. A Comparative analysis of France, Russia, and China*. Cambridge: Cambridge University Press.

Skocpol, Theda. 1992. *Protecting soldiers and mothers: The political origins of social policy in the United States*. Cambridge: Harvard University Press.

Somers, Margaret. 1998. Symposium on Historical Sociology and Rational Choice Theory "We're No Angels": Realism, Rational Choice, and Relationality in Social Science. *American journal of sociology*, 104(3): 722-784.

Streeck, Wolfgang. 2009. *Re-forming capitalism: Institutional change in the German political economy*. Oxford: Oxford University Press.

Streeck, Wolfgang. and Thelen, Kathleen. 2005. *Beyond continuity: Institutional change in advanced political economies*. Oxford: Oxford University Press.

Tocqueville, Alexis De. 1960. *De la démocratie en Amérique*. London: MacMillan & Co Ltd.

Weber, Max. 2001. *The protestant ethic and the spirit of capitalism*. Chicago: Fritzroy Dearborn Publishers.

PARTIE III
MIXED METHODS AND CROSS-CUTTING APPROACHES

16. Mixed methods

PIERRE PLUYE

Abstract

Mixed methods refer to the integration of qualitative and quantitative methods in an evaluation or research project. The approach involves thinking about this integration at all stages of the project, from the formulation of the research questions to the literature review and data analysis. Mixed methods can make a greater descriptive, explanatory or predictive contribution than either qualitative or quantitative methods taken separately.

Keywords: Mixed methods, integration, sequential exploratory design, sequential explanatory design, convergent design, mixed methods literature review

I. What do these methods consist of?

Any programme can be evaluated by combining the power of words (sounds and images) with the power of numbers (Pluye and Hong 2014). For example, you can collect stories from stakeholders and users that illustrate successes or failures from which practical lessons (rooted in stakeholders' experience) can be drawn to improve an intervention; in addition, you can collect available statistics on that intervention, or plan to collect them in a cross-sectional way (e.g., with a survey) or longitudinally (e.g., with routine data collection inserted into daily activities). The integration of stories and statistics is a powerful way to address complex policy challenges and questions. In the following sections, the mixed methods approach is presented along the different stages of the research.

Clearly formulating specific questions

Mixed methods allow you to answer interdependent (e.g., sequential) or complementary (e.g., convergent) qualitative and quantitative evaluation or research questions about a public policy. For example, you may formulate a general mixed methods objective combining exploration and measurement, and then specific qualitative and quantitative questions (see Table 1). Any question should be clearly formulated. It should express a single idea (an interrogative sentence). Evaluation and research questions usually arise from problems and challenges encountered in the creation, development, implementation (e.g., adaptation to the context) and sustainability (e.g., adjustment to changes in the context) of public policies. They are imposed by management or suggested by stakeholders and users.

Conducting a mixed studies review (mixed methods literature review)

Any assessment or research is guided by existing knowledge. This knowledge comes from experts, grey literature (e.g., reports from public organisations that can be identified with Google Scholar or OpenAlex) and publications indexed in bibliographic databases such as Cairn, Érudit, Scopus, etc. The help of a librarian is invaluable. Start by conducting a review of published literature in the form of scientific articles, book chapters or theses. Identify the most relevant documents (those that answer your questions) and plan a knowledge update. Use a document management software to keep track of the process and to make it easier to write the “Introduction” and “Discussion” sections of your report (e.g., the free software Zotero).

To update knowledge, mixed studies reviews combine quantitative, qualitative and/or mixed studies. They are becoming increasingly popular as they allow qualitative and quantitative questions to be answered by taking advantage of the complementarity of knowledge derived from qualitative, quantitative and mixed methods studies. When a public policy and its effects are well known, they can provide a thorough and comprehensive understanding of the policy in several contexts. The vast majority of literature reviews are not systematic (these being expensive and time-consuming), but mixed studies reviews can be systematic, like any other type of review.

Choosing a mixed methods research design

Usually, evaluation and mixed methods research is based on three basic designs: sequential exploratory, sequential explanatory, and convergent design (See Table 2).

The sequential exploratory design [QUAL → QUAN] begins with qualitative data collection and analysis (QUAL). In this design, the results of qualitative phase 1 inform the data collection and analysis of quantitative phase 2 (QUAN). Phase 2 is thus based on a qualitative understanding of the participants’ perspective. This design involves first exploring the phenomenon of interest qualitatively, and then using the qualitative results to guide the sampling and construction of the subsequent quantitative data collection tool (integration).

In the sequential explanatory design [QUAN → QUAL], quantitative data collection and analysis (phase 1) precedes and informs qualitative data collection (phase 2). This design involves an initial quantitative assessment followed by a qualitative exploration of these results, so that the qualitative results contribute to the explanation of unexpected or extreme quantitative results for instance (integration).

The convergent design [QUAN + QUAL] is the most frequently used. It combines qualitative and quantitative methods in an independent and complementary way. In other words, the collection and analysis of qualitative and quantitative data are not dependent on each other. They may or may not be conducted simultaneously. Indeed, it is rare to have enough resources to do everything at once. Convergence (integration) occurs when the qualitative and quantitative results are interpreted. This involves the collection of both qualitative and quantitative data to answer a similar question formulated in a qualitative and a quantitative way.

Data collection and analysis

Data collection and analysis should take into account the available data sources and the specific techniques, qualitative or quantitative, needed to analyse them. Some procedures may be mixed, for example the Delphi technique (combining interviews and questionnaires with a medium-sized sample including experts from around the world). As many statistical and qualitative analysis procedures and techniques can be used, this brief focuses on the integration of qualitative and quantitative methods.

Integration strategies

Plan any relevant combination of strategies to integrate qualitative and quantitative phases (connection), results (comparison) and data (assimilation). Based on a methodological review, we have identified three types of integration and nine operational strategies (three per type of integration) for successfully integrating qualitative and quantitative methods into mixed methods. Furthermore, we identified all possible combinations of these strategies (Pluye et al. 2018). These combinations have been confirmed in the literature on primary care, nursing, and education, environmental and information sciences. To take this further, specific integration techniques are described in a manual (Fetters 2020).

II. How are these methods useful for policy evaluation?

Mixed methods have been developed in several fields since the 1970s. They formalise procedures and techniques to integrate qualitative and quantitative methods in evaluation and research (Pluye et al. 2019). In this way, they provide a greater understanding than the sum of the knowledge obtained separately with qualitative and quantitative methods. For example, they can answer statistical questions about the effects and costs of interventions, and qualitative questions about the processes behind them, and the experiences and perspectives of stakeholders.

III. An example of the use of mixed methods in the health sector

A governmental Health Technology Assessment (HTA) agency produces and disseminates recommendations (e.g., guidelines on the optimal use of medicines and standards on the management of social services) nationally via professional associations, social services and health services. The agency's management implements evaluative research to justify the sustainability of this intervention (accountability). For each recommendation available on the agency's website, a validated questionnaire (Granikov et al. 2020) allows users to assess its relevance, cognitive impact, for example learning, and intention to use it. Over a 2-year period, more than 6000 responses were submitted and analysed (descriptive statistics). In addition, interviews were conducted with 15 users to identify the health-related effects of using the recommendations (thematic analysis). The integration of statistics and themes allows the estimation of the impact of the intervention (use and effects), and the addition of expected types of effects in the questionnaire.

IV. What are the criteria for judging the quality of and reporting mixed methods?

Mixed methods must meet three necessary conditions or essential characteristics: (a) at least one qualitative and one quantitative method are integrated; (b) each method is used in a rigorous manner with respect to the criteria generally accepted in the methodology or research tradition relied upon; and (c) the integration of the methods is accomplished at a minimum through the use of evaluation or research questions, a design, and a strategy for integrating qualitative and quantitative phases, results or data. There are a number of tools that can be used to assess the quality of mixed methods by applying these principles. Their list is updated on the catevaluation.ca website. The most popular validated tool is available free of charge on the Internet (Hong et al. 2018): it includes a checklist, a user manual and answers to frequently asked questions (mixedmethodsappraisaltoolpublic.pbworks.com).

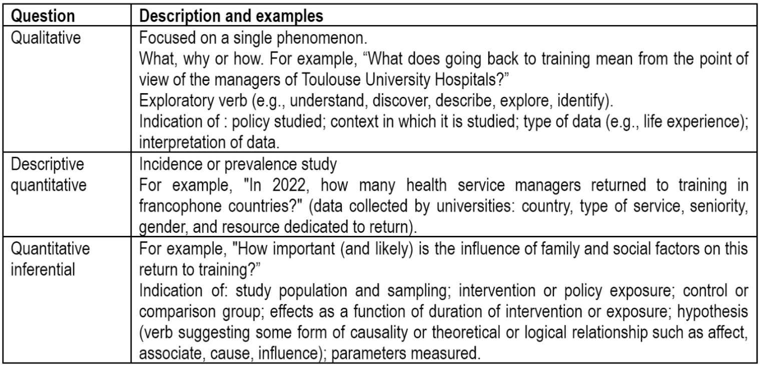
In addition, there are many guides and manuals that facilitate the writing of an evaluation report or scientific publication using mixed methods (Creswell and Plano Clark 2018). Their list is updated on the equator-network.org website. The GRAMMS (“Good Reporting of a Mixed Methods Study”) recommendations list six essential elements to be included in a document based on mixed methods (O’Cathain, Murphy, and Nicholl

2008): (a) justify the use of these methods in relation to the research questions; (b) indicate the design (sequential or convergent) of the use of mixed methods; (c) detail the qualitative and quantitative methods used; (d) specify when, how, and by whom the integration of the methods used was carried out; (e) present the limitations of the methods; and (f) indicate what the different methods contributed, as well as the complementary contribution of their integration.

V. What are the strengths and limitations of mixed methods compared to other methods?

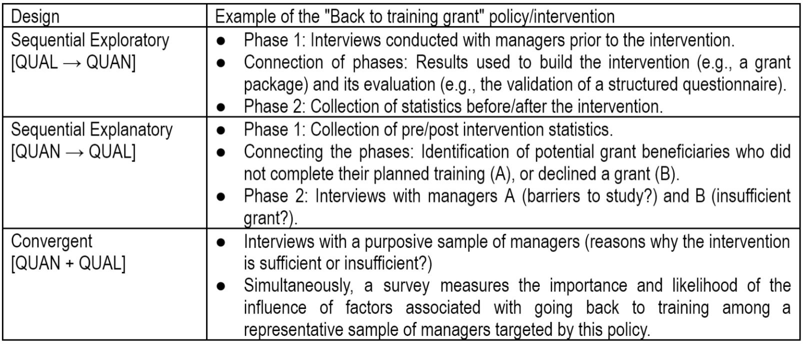
The advantages of mixed methods lie in the synergy between qualitative and quantitative methods. The integration of these methods adds value to the methods taken separately (Fetters and Freshwater 2015). Furthermore, mixed methods entail additional work to collect and analyse both words (sounds and images) and statistics, and to integrate qualitative and quantitative data and results. Their mobilisation can therefore be more time-consuming than a single method, and requires a multidisciplinary team with at least one expert for each of the selected methods. Finally, they require more space in a publication.

Table 1. Qualitative and quantitative questions



Question	Description and examples
Qualitative	Focused on a single phenomenon. What, why or how. For example, “What does going back to training mean from the point of view of the managers of Toulouse University Hospitals?” Exploratory verb (e.g., understand, discover, describe, explore, identify). Indication of : policy studied; context in which it is studied; type of data (e.g., life experience); interpretation of data.
Descriptive quantitative	Incidence or prevalence study For example, "In 2022, how many health service managers returned to training in francophone countries?" (data collected by universities: country, type of service, seniority, gender, and resource dedicated to return).
Quantitative inferential	For example, "How important (and likely) is the influence of family and social factors on this return to training?” Indication of: study population and sampling; intervention or policy exposure; control or comparison group; effects as a function of duration of intervention or exposure; hypothesis (verb suggesting some form of causality or theoretical or logical relationship such as affect, associate, cause, influence); parameters measured.

Table 2. Three basic designs used in mixed methods: Examples



Design	Example of the "Back to training grant" policy/intervention
Sequential Exploratory [QUAL → QUAN]	● Phase 1: Interviews conducted with managers prior to the intervention. ● Connection of phases: Results used to build the intervention (e.g., a grant package) and its evaluation (e.g., the validation of a structured questionnaire). ● Phase 2: Collection of statistics before/after the intervention.
Sequential Explanatory [QUAN → QUAL]	● Phase 1: Collection of pre/post intervention statistics. ● Connecting the phases: Identification of potential grant beneficiaries who did not complete their planned training (A), or declined a grant (B). ● Phase 2: Interviews with managers A (barriers to study?) and B (insufficient grant?).
Convergent [QUAN + QUAL]	● Interviews with a purposive sample of managers (reasons why the intervention is sufficient or insufficient?) ● Simultaneously, a survey measures the importance and likelihood of the influence of factors associated with going back to training among a representative sample of managers targeted by this policy.

Some bibliographical references to go further

Creswell, John. et Vicki, Plano Clark. 2018. *Designing and conducting mixed methods research*. 3rd éd. Thousand Oaks: SAGE.

Fetters, Michael. 2020. *The mixed methods research workbook: Activities for designing, implementing, and publishing projects*. Thousand Oaks: SAGE.

Fetters, Michael. and Freshwater, Dawn. 2015. “The 1+ 1= 3 integration challenge”. *Journal of Mixed Methods Research*, 9(2): 115-17.

Granikov, Vera. and Grad, Roland. and El Sherif, Reem. and Shulha, Michael. and Chaput, Genevieve. and Doray, Genevieve. and Lagarde, François. and Rochette, Annie. and Tang, David Li. and Pluye, Pierre. 2020. “The Information Assessment Method: Over 15 years of research evaluating the value of health information.” *Education for Information*, 36(1): 7-18.

Hong, Quan Nha. and Fàbregues, Sergi. and Bartlett, Gillian. and Boardman, Felicity. and Cargo, Margaret. and Dagenais, Pierre. and Gagnon, Marie-Pierre. and Griffiths, Frances. and Nicolau, Belinda. and O’Cathain, Alicia. 2018. “The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers.” *Education for Information*, 34(4): 285-91.

O’Cathain, Alicia. and Murphy, Elizabeth. and Nicholl, Jon. 2008. “The quality of mixed methods studies in health services research.” *Journal of Health Services Research and Policy*, 13(2): 92-98.

Pluye, Pierre. and Bengoechea, Enrique García. and Granikov, Vera. and Kaur, Navdeep. and Tang, David Li. 2018. “Tout un monde de possibilités en méthodes mixtes: revue des combinaisons des stratégies utilisées pour intégrer les phases, résultats et données qualitatifs et quantitatifs en méthodes mixtes.” In: *Oser les défis des méthodes mixtes en sciences sociales et sciences de la santé*, ed. by Bujold, Mathieu. and Hong, Quan Nha. and Ridde, Valéry. and Bourque, Claude Julie. and Dogba, Maman Joyce. and Vedel, Isabelle. and Pluye, Pierre. 28-48. Montréal: Association francophone pour le savoir.

Pluye, Pierre. and Bengoechea, Enrique García. and Tang, David Li. and Granikov, Vera. 2019. “La pratique de l’intégration en méthodes mixtes.” In: *Évaluation des interventions de santé mondiale: méthodes avancées*, ed. by Ridde, Valéry, and Christian Dagenais, 213-38. Québec: Éditions science et bien commun.

Pluye, Pierre. and Hong, Quan Nha. 2014. “Combining the power of stories and the power of numbers: Mixed methods research and mixed studies reviews.” *Annual Review of Public Health*, 35: 29-45.

17. Mixed methods systematic reviews

QUAN NHA HONG

Abstract

Mixed methods systematic literature reviews, also named mixed studies reviews, consist of reviewing available work (including evaluations already carried out) on a given topic by incorporating studies using qualitative, quantitative and mixed methods. This type of literature review allows for a better understanding of complex interventions and phenomena. By encompassing a diversity of questions, they are particularly useful for informing public decision-making.

Keywords: Mixed methods, qualitative studies, quantitative studies, systematic review, literature review

I. What does this method consist of?

The literature review process consists of summarising, combining, analysing, commenting on or critiquing the literature on a given topic. Mixed methods systematic literature reviews, also named mixed studies reviews, are unique in that they include a variety of types of studies to better understand complex phenomena: quantitative studies (e.g. studies that measure the effects of an intervention or the magnitude of a problem), qualitative studies (e.g. studies that focus on people's experiences), and mixed methods studies (i.e. studies that use both quantitative and qualitative methods).

Mixed methods systematic reviews are part of the larger family of systematic literature reviews, i.e. a type of literature review that aims to answer a research question by following a pre-defined approach to the identification, selection, appraisal and synthesis of relevant studies. It is considered to be one of the most rigorous types of review since it minimises errors and biases that may occur during the review process. These reviews therefore use explicit methods and report them transparently so that they can be replicated. In addition to the systematic review, there are a variety of types of reviews that use a systematic approach, such as scoping review, rapid review and overview of reviews.

In general, conducting a mixed methods systematic review follows eight steps:

1. Formulate a research question(s) – the formulation of questions can be guided by a cursory exploration of the existing literature on the topic of interest;
2. Define eligibility criteria (inclusion and exclusion) for the selection of articles;
3. Identify literature sources to ensure a comprehensive search such as searching bibliographic databases (e.g. PubMed, Health Policy Reference Center, International Political Science Abstracts, Europa World, Worldwide Political Sciences Abstracts, Web of Science, JSTOR, SocINDEX), consulting the table of contents of scientific journals on the topic of interest, searching websites on the topic, consulting the reference list of selected articles or articles that have been cited, and contacting experts;
4. Develop a literature search strategy with the assistance of a specialised librarian who is familiar with literature search techniques in databases and other sources;

5. Selecting relevant documents in two stages: selection based on document titles and abstracts, and selection based on full-text articles;

6. Assessing the quality of the selected documents using critical appraisal tools;

7. Extracting data from the selected documents using a form that specifies all the data to be extracted; and

8. Synthesising the extracted data, i.e. analysing the data extracted in the review into a coherent whole to answer the research question(s). In mixed methods systematic reviews, the synthesis should also consider how the qualitative and quantitative data will be integrated. In general, two main synthesis designs can be used to synthesise qualitative and quantitative data: (a) convergent synthesis designs in which data are synthesised simultaneously (e.g. quantitative and qualitative studies are synthesised separately with different synthesis methods and then the results of the two syntheses are compared) and (b) sequential synthesis designs that involve at least two dependent phases of synthesis (e.g. qualitative studies are synthesised first and then the results of this synthesis inform the synthesis of the quantitative studies).

For more information on each of these steps and on the synthesis designs, you can consult the references at the end of this document as well as this website: http://toolkit4mixedstudiesreviews.pbworks.com.

II. How is this method useful for policy evaluation?

Mixed methods systematic reviews are relevant for policy evaluation, as they allow for a deeper understanding of complex phenomena and interventions. Complex phenomena are often characterised by a multiplicity of actors involved, and a diversity of intervention models and factors influencing their success. Various types of studies can be used to assess these complex phenomena. Thus, the synthesis of these complementary studies will provide a more comprehensive understanding of the state of knowledge on a complex phenomenon.

The mixed methods systematic review provides a broader and more complete picture of the literature on a given topic. It also allows for the combination of complementary questions such as: How effective is a policy? Why is the policy effective or not? How does the policy work? What factors hinder or facilitate the implementation of the policy? To answer these questions, it is necessary to include quantitative studies to answer questions about policy effectiveness and qualitative studies that address the why and how questions. The answers to these complementary questions can lead to better decision-making by policy makers, managers and practitioners.

Other advantages of mixed methods systematic reviews were highlighted such as allowing for a better understanding of the results obtained in quantitative studies from qualitative studies (or vice versa), considering a diversity of perspectives (e.g. perspectives of decision makers and users), corroborating findings obtained from different evidence, and enhancing the credibility and validity of conclusions.

III. An example of the use of this method: the fight against smoking among young people

A mixed methods systematic review was conducted by researchers at the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) in the UK to inform the development of policies to reduce youth smoking rates (Sutcliffe, Twamley, Hinds et al., 2011). In this review, they addressed three main research questions: a) What are the most common sources in which young people aged 11-18 years access retail and non-retail tobacco products and do the sources vary by factors such as age and gender? (b) What are young people's perceptions of access to tobacco products and what do they believe are the barriers and facilitators to accessing tobacco products; and (c) What types of interventions to limit access to tobacco products by young people in non-retail settings have been evaluated and how do these interventions address the barriers and facilitators identified as important by young people in the UK?

For this mixed methods systematic review, a literature search of over 100 sources of information was conducted (e.g. databases, websites, citation searches). This search identified six qualitative studies, seven surveys, and sixteen intervention studies. Two people were independently involved in the selection of studies, quality assessment, data extraction, and synthesis. The syntheses for each type of study included (surveys, qualitative studies, intervention studies) were carried out separately. For integration, the results of these syntheses were then compared in two ways: a) to assess the level of concordance between the results of the surveys and qualitative studies regarding young people's sources of tobacco products and their access patterns by gender, age and smoking status (occasional or regular); and b) to assess the extent to which the interventions addressed the barriers and facilitators identified as important by young people in the qualitative studies.

In light of the findings of this review, three main implications for the development of policies to reduce youth smoking rates were formulated. Firstly, the studies found that tobacco products were easily accessed through friends and peers in schools (social access). Young people described this social access as ubiquitous, organised, and visible. Thus, the development of stricter regulations in schools to reduce social access should be explored. Secondly, the results of the studies indicate that the implementation of retail regulation was variable. It is therefore necessary to explore the reasons for uneven implementation and to identify ways to combat lax implementation of regulation in smaller retail stores. Thirdly, the findings of this review suggest the need to address adult complicity to purchase tobacco products such as by family, friends and strangers (proxy purchase).

In this mixed methods systematic review, the use of data from different types of studies allowed for the identification of different modes of access used, a better understanding of young people's experiences and views on access to tobacco products, and the exploration of potential avenues for intervention. Also, the mixed methods nature of the review, which combines survey data, research on the views of young people in the UK and interventions addressing non-retail access to tobacco products, has provided contextualised evidence for policy development.

IV. What are the criteria for judging the quality of the mobilisation of this method?

To judge the quality of systematic reviews, it is important to understand the potential sources of errors and biases that can influence the results obtained. Four biases are presented in this paper: identification bias, reporting bias, selection bias and interpretation bias.

Identification bias occurs when relevant studies on the topic of interest are not identified. This bias is related to the literature search and the sources of identification. In order to conduct a comprehensive search for all studies relevant to the research question, it is important to diversify literature sources, use different databases and develop rigorous literature search strategies with the collaboration of specialised librarians.

Reporting bias, of which publication bias is the best known, occurs when the nature, direction or strength of a study's results influence its publication. For example, it has been shown that studies with positive effects are more likely to be published and published more quickly in scientific journals than those with negative results. This may lead to an over-representation of studies showing positive effects and may affect the conclusions of the systematic review. To minimise this bias, it is recommended to diversify the sources of data and the types of documents to be included, such as scientific reports from research centres, and master's theses and doctoral dissertations.

Selection bias occurs when the selection of studies is arbitrary or influenced by particular motivations or beliefs. For example, a researcher may believe that an intervention is important and decide to include only studies that have shown that the intervention is effective. This will bias the results of the review. To minimise this bias, it is important to define clear eligibility criteria prior to selection and to involve two people in the selection process.

Interpretation bias is related to the persons' misinterpretation of the studies. This bias can be minimised by involving at least two people in data extraction, quality assessment and data synthesis. Furthermore, in a mixed methods systematic review, it is recommended to have a team with complementary expertise in qualitative and quantitative research to facilitate the synthesis and judgement of studies' quality.

V. What are the strengths and limitations of this method compared to others?

A mixed methods systematic review makes it possible to answer a variety of questions, to take advantage of the complementarity of quantitative and qualitative data, and to gain a thorough and complete understanding of a complex phenomenon. Also, using a rigorous and explicit methodology helps to minimise potential errors and biases that could influence the validity of a review's findings. However, various challenges can arise in its operationalisation.

One important challenge is the time and resources required. The duration of a systematic review can vary from 6 to 24 months. Various factors can influence its duration such as the research questions to be addressed, the number of people involved, the number of documents to be analysed, and the synthesis method(s) to be used. Also, including a variety of study types in a mixed methods systematic review increases the volume of material to be identified, screened, extracted and analysed. It is therefore important to ensure that resources are available and that the choice to conduct a mixed methods systematic review is well justified.

The questions that can be studied in a systematic review depend on the available literature. For example, in the context of a mixed methods systematic review, a research team might be interested in identifying studies on the effects of an intervention and others on the users' perspective on the intervention. However, let us imagine that the literature search only identifies studies on the effects and none on the user perspective of the intervention of interest. In this example, the review is not mixed since only one type of study is synthesised. In order to guide the specific research questions that could be addressed on a topic of interest in a mixed methods systematic review, it may be useful to conduct a preliminary, cursory exploration of the existing literature.

Another challenge is the integration of data, i.e. how the qualitative and quantitative components are combined. This is a key feature of the mixed methods systematic review, which allows the full range of results from the various types of studies selected to be integrated in order to provide a deeper understanding of the topic of interest and to make recommendations that reflect the full body of literature covered. A review that does not have integration could be considered to include several independent reviews rather than a mixed methods review. It is therefore essential that the way in which the data is integrated is well described and that the added value of this integration and its limitations are well reflected.

Some bibliographical references to go further

Granikov, Vera. and Hong, Quan Nha. and Pluye, Pierre. 2022. “Mixing Qualitative and Quantitative Evidence in a Systematic Review: Methodological Guidance with a Worked Example of Collaborative Information Monitoring.” In Handbook of Research on Mixed Methods Research in Information Science, edited by Ngulube, Patrick. 125-146.

Heyvaert, Mieke. and Hannes, Karin. and Onghena, Patrick. 2016. Using Mixed Methods Research Synthesis For Literature Reviews: The Mixed Methods Research Synthesis Approach. Thousand Oaks, CA: SAGE Publications.

Hong, Quan Nha, and Pluye, Pierre. and Bujold, Mathieu. and Wassef, Maggy. 2017. “Convergent and sequential synthesis designs: Implications for conducting and reporting systematic reviews of qualitative and quantitative evidence.” *Systematic Reviews*, 6(61): 1-14. https://doi.org/10.1186/s13643-017-0454-2.

Hong, Quan Nha, Rees, Rebecca. and Sutcliffe, Katy. and Thomas, James. 2020. “Variations of mixed methods reviews approaches: A case study.” *Research Synthesis Methods*, 11 (6): 795-811. https://doi.org/10.1002/jrsm.1437.

Lizarondo, Lucylynn. and Stern, Cindy. and Carrier, Judith. and Godfrey, Christina. and Rieger, Kendra. and Salmond, Susan. and Apostolo, Joao. and Kirkpatrick, Pamela. and Loveday, Heather. 2020. “Chapter 8: Mixed methods systematic reviews.” In *JBI Manual for Evidence Synthesis*, ed. by E. Aromataris and Z. Munn. Adelaide: JBI https://jbi-global-wiki.refined.site/space/MANUAL.

Sutcliffe, Katy. and Twamley, Katherine. and Hinds, Kate. and O’Mara, Alison. and Thomas, James. and Brunton, Ginny. 2011. *Young people’s access to tobacco: A mixed-method systematic review*. EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London (London). https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3301.

18. Macro comparisons

EMANUELE FERRAGINA

Abstract

Macro comparisons is an approach that exploits variation and similarities across large macro-social units of analysis (e.g. states, regions, provinces) to investigate different social phenomena. Studies can be undertaken at different scales and for various purposes, for example describing macro differences among different states, or evaluating the influence of a different welfare state structure on individual outcomes (such us levels of unemployment, life expectancy etc…).

Keywords: Mixed methods, macro-social units, variation, similarities, welfare state

I. What does this approach consist of?

All scientific inquiry is inherently comparative, and this is clearly observable when considering the logic applied to the most common methods in social sciences. To provide some examples: experiments are comparative because they need a control group to measure against a null case the effect of a treatment; regression analyses control for the effect of several variables comparing their effect on a range of cases. Hence, while all research methods are comparative in a broad sense, in the social sciences the idea of comparative inquiry often refers to research involving the use of large macro-social units of analysis (Ragin, 2014). Research in this sense is comparative when it exploits the variation or similarity of macro social units of analysis, e.g. a state, a region, a province[1]. This can give way then to studies that are based upon different levels and scales, but all include the use of macro-units of analysis. The goal of these macro comparisons is to understand causal complexity and describe the relation between macro and micro units of analysis and between macro units of analysis among each other. The literature provides different examples, e.g. the comparisons between different social security models, or the evaluation of how a specific configuration of family policy impacts on female employment and fertility rates. The analysis of macrosocial units is a 'meta-theoretical category', which basically distinguishes comparative social scientists from the others, because they use 'macrosocial units in explanatory (and descriptive) statements' (Ragin, 2014: 5). Indeed, the vast majority of scholars working in the field (including the author of this chapter!), often do not define the nature and the role of the macrosocial units, but rather use them implicitly as 'observations' and/or 'explanatory' units of analysis (Ragin, 2014: 8).

Accordingly, the use of macro comparisons is more a way of thinking than a method stricto sensu. Macro comparisons can be set employing different techniques at the quantitative, qualitative and historical level, e.g. descriptive statistics, case studies and comparative historical analysis (CHA), qualitative comparative analysis (QCA)/fuzzy-sets, regression techniques, structural equation modelling (SEM) and factorial analyses,

[1] So for example multi-level modelling is included within this definition. However, Ragin's definition of comparative research as grounded in macro-social units of analysis is not universally accepted. Other scholars have proposed different boundaries to delimit the domain of comparative inquiry. On the one hand, those more geared towards the use of quantitative and multivariate techniques have defined the comparative method simply by considering studies that include comparative data from different societies (Andreski, 1965; Armer, 1973) or works based on multilevel analysis (Rokkan, 1966; Przeworski and Teune, 1970). On the other hand, scholars more versed in qualitative/historical analysis such as Moore (1966) and Skocpol (1979) tend to distinguish between case-based and variable-orientated comparative methods (the lineage is of course traced to the founding fathers of sociology and political science, e.g. Tocqueville, Durkheim and Weber). We suggest that these views are too restrictive for our purposes, and for this reason, together with Ragin (2014), we define the comparative method and macro comparisons on the basis of their main goal.

and cluster analysis. Other techniques used less frequently are diagonal reference models, sequence analysis, scale construction, thematic analysis, propensity score matching (PSM), optimal matching, Krippendorff's alpha (KA) and event history analysis (for a systematic review of methods used in macro comparative research see Ferragina and Deeming 2022). This means that macro comparisons are not bounded to specific techniques, but rather need to be viewed as structuring 'thinking about thinking' (Sartori 1970) in order to increase the inference (the broader conclusions that may be drawn) we gauge from the study of specific cases.

II. How is this approach useful for policy evaluation?

Macro comparisons are extremely useful for the evaluation of public policy both ex ante and ex post. In particular macro comparisons have an important role in helping to contextualise the evidence provided by specific case studies or experimental evaluations of public policies. Key to advancing the debate about the relation between specific policies and their effects is the ability of comparative macro comparisons and national case studies to learn from each other (Ferragina 2020). National case studies – e.g. the evaluation of a specific policy within a country – are often plagued by a lack of external validity (the capacity to generalise the conclusions beyond the case under study). On the other hand, when using experiments scholars are able to test the effect of incremental reforms, but not the overall effect of a policy component on a specific outcome. So for example, in the field of family policy, macro comparisons can help to disentangle how the joint effect of explicit family policies differently (i.e. childcare, leave and child income support) impact on female employment across countries, while experiment can allow to disentangle the specific effect of an increase in the number of childcare facility on women's' employment elasticity in a specific case. For this reason, we need more studies that interact systematically with policy measures and the context in which they are implemented. In this sense macro comparisons can not only offer interesting insights about the effects of different policies cross-nationally or cross-regionally, but also allow us to critically evaluate the results from specific evaluations. Moreover, from an explanatory point of view, the existence of consolidated macro comparative evidence can help to interpret the results from studies run at the national level. This is the case of one of the most famous macro comparative works ever published, namely The Three Worlds of Welfare Capitalism by Gøsta Esping-Andersen (1990).

III. The three worlds of welfare capitalism: A famous example of how the comparative method can inform different types of policy evaluations

The Three Worlds of Welfare Capitalism is part of a long-standing academic tradition in sociology and political science rooted in deductive reasoning[2] and the use of ideal types[3]. As Max Weber (1904: 87) highlighted, 'the construction of a system of abstract and therefore purely formal propositions …, is the only means of analysing and intellectually mastering the complexity of social life'. In this vein, Esping-Andersen (1990) constructed the welfare regime typology acknowledging the ideational importance and power of the three dominant political movements of the long 20th century in Western Europe and North America, that is, social democracy, Christian democracy (conservatism) and liberalism.

[2] Deductive reasoning is a form of logical thinking that starts with a general idea and reaches a specific conclusion. It is a top-down thinking that moves from the general to the specific.

[3] An ideal type is an analytical construct derived from observable reality although not conforming to it in detail because of deliberate simplification. It is "ideal" because it is used to approximate reality by selecting and accentuating certain elements.

The ideal social-democratic welfare state is based on the principle of universalism, granting access to benefits and services based on citizenship. Such a welfare state is said to provide a relatively high degree of autonomy, limiting the reliance on family and market. In order to achieve autonomy, social-democratic welfare states are characterised by a high level of decommodification and a low degree of stratification. Social policies are perceived as 'politics against the market' (Esping-Andersen, 1985). Christian-democratic welfare states are based on the principle of subsidiarity and the dominance of social insurance schemes, offering a medium level of decommodification[4] and a high degree of social stratification. The liberal regime is based on the notion of market dominance and private provision; ideally, the state only interferes to ameliorate poverty and provide for basic needs, largely on a means-tested basis. Hence, the decommodification potential of state benefits is low and social stratification high. However, these models are not pure and in each real national case different features are mixed. In this sense Esping-Andersen clearly shows how the comparative device is a way to classify and understand differences and clusters of countries, but needs to be considered with caution:

> We show that welfare state clusters, but we must recognise that there is no single pure case. The Scandinavian countries may be predominantly social democratic, but they are not free of crucial Liberal elements. Neither are the Liberal regimes pure types. The American social-security system is redistributive, compulsory and far from actuarial. At least in its early formulation, the New Deal was as social democratic as was contemporary Scandinavian social democracy. And European conservative regimes have

[4] Decommodification refers to the degree of to which individuals, or families, can uphold a socially accepted standard of living independently of market participation (as defined by Esping-Andersen in the Three Worlds of Welfare Capitalism).

> incorporated both Liberal and social democratic impulses. Over the decades, they have become less corporatist and less authoritarian (Esping-Andersen, 1990: 28-29).

Various contributions have confirmed his typology, while others have challenged, and expanded it, from substantive and methodological perspectives (see Ferragina and Seeleib-Kaiser 2011; Ferragina and Filetti 2022 for a discussion). However, despite this lengthy debate and important controversies in the literature, one cannot deny the fundamental role this work has assumed in the structuration and understanding of an important segment of public policy, namely social policy. In particular, The Three Worlds of Welfare Capitalism offers a plastic representation of the utility of macro comparisons and the framework developed by Esping-Andersen has been used as a departure point for thousands of studies (at the 26th of October 2022 the book has been cited 44086 times!).

Concerning public policy evaluation Esping-Andersen's work has been used:

- To select different studies for analysis. The selection of at least one social democratic case, one Christian democratic case and one liberal case has allowed scholars to draw more insights from the study of a few countries.
- As a heuristic device to interpret the effects of different policies across countries.
- To understand and describe the different trajectories of countries over time.
- To contextualise the results obtained when comparing different countries.

IV. What are the strengths and limitations of this approach compared to others?

Macro comparisons are used to test hypotheses, infer causation, illustrate and gain in depth understanding of specific patterns, and interpret social change. They allow greater interpretative power in comparison to single case studies. This implies a strong heuristic power. It is not a random coincidence that highly cited works like the Three Worlds of Welfare Capitalism provided researchers in public policy with important insights on a large number of developed countries, which remain still valid more than 30 years after the publication of Esping-Andersen's work. Macro comparisons allow us to pay attention to the context and the potential effects that this context might exert on specific outcomes. However, the fact of looking at 'the forest' instead of 'the trees' impose on the one hand high costs for the researcher (in terms of expertise about multiple cases), and on the other it requires a simplification of the analysis to accommodate the comparisons between different macro units of analysis. This can generate several issues, such as misclassification (the creation of pseudo-classes that incorrectly simplify the universe of cases analysed) and 'conceptual stretching', that is the erroneous application of theories and concepts to cases other than the ones that have been analysed.

Often scholars tend to include a lot of countries in their comparison by broadening the categories they have developed on the basis of direct knowledge acquired through few cases. However, this broadening can be problematic in many respects. On the one hand it is useful to have more countries in order to provide a better test of a series of hypotheses, but on the other, with fewer cases one can be more precise in the definition of concepts. This trade-off is not always considered in modern social sciences, with comparisons that end up over-stretching concepts. Therefore, concepts and insights extracted from macro comparisons need to be used with a grain of salt. As an approach more than a method, macro comparisons allow a critical approach to social sciences and historically raised important questions on the results obtained from researchers. In conclusion, macro comparisons are a double edge sword, they can inform in a meaningful way public policy evaluation, but they have also to be considered with caution.

Cited references

Andreski, Stanislav. 1965. *The Uses of Comparative Sociology*. Berkeley: University of California Press.

Armer, Michael. 1973. “Methodological problems and possibilities in comparative research”. In: Armer, Michael. and Grinmshaw, Qllen. (eds). *Comparative Social Research*. New York: Wiley, 49–79.

Esping-Andersen, Gosta. 1985. *Politics against markets*. Princeton: Princeton University Press.

Esping-Andersen, Gosta. 1990. *The Three Worlds of Welfare Capitalism*. Princeton: Princeton University Press.

Ferragina, Emanuele. 2020. Family policy and women’s employment outcomes in 45 high-income countries: A systematic qualitative review of 238 comparative and national studies. *Social Policy & Administration*, 54(7): 1016-1066.

Ferragina, Emanuele. and Deeming, Christopher. 2022. “Methodologies for comparative social policy analysis”. In: Yerkes Mara A. and Nelson, Keneth. and Nieuwenhuis, Rense (eds). *Changing European Societies: The Role for Social Policy Research*. Cheltenham: Edward Elgar, 218–235.

Ferragina, Emanuele. and Filetti, Federico Danilo. 2022. Labour market protection across space and time: a revised typology and a taxonomy of countries’ trajectories of change. *Journal of European Social Policy*, 32(2): 148–165.

Ferragina, Emanuele. and Seeleib-Kaiser, Martin. 2011. Welfare regime debate: past, present, futures? *Policy & Politics*, 39(4): 583–611.

Moore, Barrington. Jr. 1966. *Social Origins of Dictatorship and Democracy*. London: Penguin.

Przeworski, Adam. and Teune, Henry. 1970. *The Logic of Comparative Social Enquiry*. New York: Wiley.

Ragin, Charles. 2014. *The Comparative Method. Moving beyond Qualitative and Quantitative Strategies*. Oakland: University of California Press.

Rokkan, Stein. 1966. "Comparative cross-national research". In: Merritt, Richard. and Rokkan, Stein. (eds). *Comparing Nations*. New Haven: Yale University Press, 3–26.

Sartori, Giovanni. 1970. Concept misformation in comparative politics. *American Political Science Review*, 64(4): 1033–1053.

Skocpol, Theda. 1979. *States and social revolutions*. Cambridge: Cambridge University Press.

Skocpol, Theda. 1992. *Protecting soldiers and mothers: The political origins of social policy in the United States*. Harvard: Harvard University Press.

Weber, Max. 1904. *On the methodology of the social sciences*. Glencoe: The Free Press.

Some bibliographical references to go further

Ferragina, Emanuele. and Deeming, Christopher. (Forthcoming). 'Comparative mainstreaming? Mapping the uses of the comparative method in social policy, sociology and political science since the 1970s'. *Journal of European Social Policy*. An analysis of 50 years of comparative research based on a database including thousands of comparative articles from top journals in sociology, political science and sociology. The quantitative analysis of the main trends in the use of the comparative method is complemented with a qualitative analysis of the most cited articles in the comparative field.

Kohn, Melvin L.. 1987. ‘Cross-national research as an analytic strategy. American Sociological Association, 1987 presidential address.’ *American sociological review*, 52(6): 713-731. This presidential address of the American Sociological Association suggests that cross-national comparative research is an essential tool to generate, test, and develop sociological theory. The comparative method is costly and hard to apply, and it can also generate some interpretative problem. However, despite its limitations it is a fundamental tool of social science research.

Lijphart, Arend. 1971. Comparative politics and the comparative method. *American Political Science Review*, 65(3): 682-693. The article offers a systematic analysis of the comparative method. Its emphasis is on both the limitations of the method and the ways in which, despite these limitations, it can be used to maximum advantage. Lijphart focuses on the role of case studies (in their different forms) as the main way to undertake macro comparisons. In the article, he contrasts case-based comparisons to experimental and statistical methods.

Przeworski, Adam. and Teune, Henry. 1970. *The Logic of Comparative Social Inquiry*. New York: John Wiley and Sons. In this ground-breaking book the authors proposed insights and views about comparative research that have profoundly shaped political science research. The book focuses mostly on quantitative analysis. A must read for all students interested in the comparative method and what can be done with it.

Ragin, Charles. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press. This book offers considerable insights for the understanding and use of comparative analysis. Originally written to present the utility of the

Qualitative Comparative Analysis (QCA) in comparison to qualitative and quantitative techniques, it also provides theoretical and substantive reasons for the use of the comparative method and macro comparisons in the public policy field.

19. Qualitative Comparative Analysis

VALÉRIE PATTYN

Abstract

Qualitative Comparative Analysis (QCA) is a mixed method which translates qualitative data into a numerical format in order to systematically analyse which configurations of factors produce a given outcome. QCA indeed relies on a configurational conception of causality, according to which outcomes derive from combinations of conditions. It is very useful for ex post impact evaluation, more specifically to understand why the same policy may lead to certain changes in some circumstances and not in others.

Keywords: Configurations, combinations of conditions, causal complexity, systematic identification of cross-case patterns, equifinality, conjunctural causation, asymmetrical causality

I. What does this method consist of?

Why is it that the same policy leads to certain changes in some circumstances and not in others? Take, for example, a subsidy programme supporting firms to provide in-company training on leadership skills. Why is it that such training is effective for some employees, and for others not? Or put differently: under what conditions does, or does not, successful ‘training transfer effectiveness’ occur? Qualitative Comparative Analysis (QCA) is a method to answer such a question.

QCA assumes that configurations – i.e. combinations of conditions – are necessary and/or sufficient to achieve a given outcome. Conditions can be conceived as causal variables, determinants, or factors (Rihoux, and Ragin, 2009: xix). An outcome, in an evaluation context, is usually a well-defined intended or unintended policy effect that may be present or absent. In the above example, the outcome is the occurrence or non-occurrence of ‘training transfer effectiveness’.

Different from other case-based methods (See separate chapter on case studies), QCA enables comparison of case-based information systematically, and as such allows for modest generalisation. At the same time, different from statistical methods, it enables us to keep rich contextual information and some complexity. Because of this twofold potential, the method is often portrayed as a bridge builder between qualitative and quantitative methods. The method was originally developed for researchers confronted with an intermediate number of cases (between 10 and 50), but is increasingly also applied in settings with a large number of cases (see Thomann, and Maggetti 2020).

Importantly, QCA is not only an analytical technique, but also comes with a specific approach to causality, labeled as multiple conjunctural causation, which is very compatible with assumptions underpinning realist evaluation (see separate chapter on realist evaluation). In particular, this approach entails that:

- Policy effects are often the result of combinations of conditions rather than the result of a single condition (‘conjunctural causation’)

- Different possible configurations can lead to the same observed effects or outcomes: this is what QCA refers to as ‘equifinality’.

- Causality is understood asymmetrically: if in a given case a certain condition is relevant for the outcome, its absence does not necessarily entail the absence of the outcome.

QCA belongs to the family of set theoretic methods. A case may be part of one or more sets. Sets articulate characteristics that certain cases may have in common. Building on our example, an employee attending an in-company training may be part of the set of cases of 'employees with autonomy in their work and decision making' and/or of the set of 'employees who received support from their supervisors in following the training'. Through identifying the extent to which a case is part of certain set and systematically comparing this with other cases with variation in the occurrence of a certain outcome (i.e. training transfer effectiveness), one can find out which (combinations of) factors are necessary and/or sufficient for this outcome:

- A (combination) of condition(s) found as necessary implies that it will always be present/absent whenever the outcome is present/absent. Or to put it in terms of set theory, X is a necessary condition for Y, if Y is a subset of X (X ← Y). For example, if we would find that all in-company training leading to training transfer effectiveness were lectured by instructors with a lot of teaching experience, the latter can be qualified as a necessary condition.

- In order for a condition (or a combination of conditions, i.e., configuration) to be sufficient, the outcome should appear whenever the condition is present. In set theory, a condition (X) is classified as sufficient if it constitutes a subset of the outcome (X → Y). For instance, a training attended by an employee with high autonomy in their work, and who is strongly motivated can constitute a sufficient path to training transfer effectiveness.

Cases can take different forms in QCA. In an evaluation setting, cases are commonly contexts in which an intervention has been applied. In the example mentioned earlier, cases relate to employees who attended a subsidised training. Cases can also be organisations or firms, or be situated at the macro-level (i.e. countries).

How then to compare such cases systematically? One can hereto resort to different QCA techniques. In the crisp set QCA (csQCA), the original version of QCA, conditions and outcome need to be translated in binary terms, 1 or 0. This is called calibration. Conditions or outcomes assigned a score of 1 should be read as present (or high, or large), while those with a score of 0 are regarded as absent (or low, or small). The binary scores express qualitative differences in kind. In the fuzzy set variant of QCA (fsQCA), cases can have partial membership in a set and any score between 0 and 1, which takes account of the fact that many social phenomena are dichotomous 'in principle' but that empirical manifestations of these phenomena in practice often differ in degree (Schneider, and Wagemann 2012, 14).

Irrespective the technique that is used, the QCA research cycle runs through similar stages:

First, with the data being calibrated, a data matrix can be constructed which basically presents the empirically observed data as a list of configurations.

Second, the calibrated data matrix can, in a subsequent stage, be transformed into a so-called truth table, which lists all possible configurations leading to a particular outcome. As a single configuration possibly corresponds with various empirical cases, the truth table thus summarises the empirical data table. The total number of theoretically possible configurations in the truth table is determined by the number of conditions included in the research. Researchers should strive for a good balance between the number of cases and conditions. The configurations not covered by empirical observations can be considered logical remainders, that is, they are logically possible, yet not observed. QCA provides the interesting opportunity to include plausible assumptions about the outcome of (a selection of) logical remainders for drawing more parsimonious inference (Schneider, and Wagemann 2012).

Third, the truth table paves the way to proceed to the QCA analytical moment, coined as Boolean minimisation. In this process, the researcher can rely on different software packages. The minimisation is built upon the assumption that if two combinations differ on only one condition, but show the same outcome, this particular condition is redundant. Thus, it can be eliminated to obtain a simpler representation of the case (or group of cases). Applying this rule iteratively on all possible pairs of combinations until no further simplification is possible results in a series of sufficient paths to the outcome. The type of findings (i.e. solution formulas) typically resulting from the QCA analysis will be expressions about 'the (combination of) conditions that are necessary and/or sufficient for the occurrence or non-occurrence of a particular outcome'.

Fourth, and most crucially, a QCA study does not stop after the application of the software. It is essential that the researcher spells out the causal link in a narrative fashion (Schneider, and Wagemann 2010), by returning to the individual cases and by relating the findings with broader theoretical and conceptual knowledge. Common to the method is its iterative nature: researchers can go back and forth between preliminary data analysis and the dataset or the theory of change. This process is also a useful vehicle to get to know the cases in more depth.

II. How is this method useful for policy evaluation?

QCA can be used for explanatory purposes which enables one to test program theories in a systematic way, or for exploratory purposes to develop theories from case-based knowledge. As the outcome (effects) should be known before the evaluation, the method can in principle only be applied for ex post and in itinere evaluations, in which an intervention has led to variation in success.

By its focus, QCA is primarily suitable for learning rather than accountability-oriented evaluations. In particular, it is often said that it can contribute to both double loop and triple loop learning: it not only sheds light on the conditions under which policy interventions work; yet it also provides potential to actively involve stakeholders in the process of selecting outcomes, conditions, and in calibrating these. As a result, stakeholders and commissioners can get a better understanding of what 'successful change' means in the context of the intervention, and of what makes a difference.

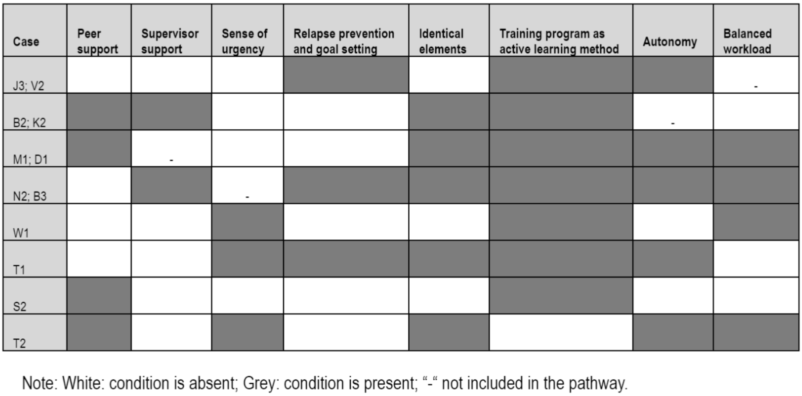
III. An example of the use of this method in training policy

We already hinted at an evaluation about the effectiveness of soft skills training (such as leadership skills). This evaluation was conducted in Flemish (Belgian) firms, commissioned by the Flemish European Social Fund (ESF) agency, which also subsidised the training. Although previous counterfactual research demonstrated the positive impact of training subsidies, it also revealed that there is not always transfer of what is learned to the working environment. This observation constituted the main rationale for the evaluation and triggered the commissioner to switch focus from 'whether the subsidised training works' to the conditions under which training programmes work. The study included 50 cases, of which 15 successful cases in which social skills were transferred and 35 failed cases where training transfer effectiveness was not achieved at the time of evaluation.

Based on relevant educational literature, 8 conditions were identified with potential explanatory power and included in our QCA model: (1) peer support; (2) supervisor support; (3) sense of urgency; (4) relapse prevention and goal setting; and the following contextual conditions: (5) identical elements, (6) training program as active learning method; (7) autonomy, and (8) balanced workload. Of these conditions, no single condition proved necessary for successful training transfer. However, we identified several pathways consisting of combinations of conditions that were sufficient to success: whenever these pathways were present, the training content was successfully retained and applied to the workplace. Table 1 below visualises the eight pathways found.

The QCA analysis proved useful to know which conditions to monitor in future subsidised programmes. The evaluation was a component of a multimethod evaluation: it was followed by Process Tracing (see separate chapter on process tracing) which focused on a selection of cases to identify the mechanisms through which training programmes make a difference in the working environment. The QCA analysis also helped systematically identify the cases for which further in-depth within-case analysis was most relevant.

Table 1: Pathways to successful training transfer



Case	Peer support	Supervisor support	Sense of urgency	Relapse prevention and goal setting	Identical elements	Training program as active learning method	Autonomy	Balanced workload
J3; V2	□	□	□	■	□	■	■	-
B2; K2	■	■	□	□	■	■	-	□
M1; D1	■	-	□	□	■	■	■	■
N2; B3	□	■	-	■	■	■	■	■
W1	□	□	■	□	□	■	□	■
T1	□	□	■	■	■	■	■	□
S2	■	□	□	□	□	■	□	□
T2	■	□	■	□	■	□	■	■

Note: White: condition is absent; Grey: condition is present; "-" not included in the pathway.

Source: Álamos-Concha, Prischilla. Cambré, Bart. Foubert, Josephine. Pattyn Valérie. Rihoux, Benoit. Schalembier Benjamin. 2020. Impactevaluatie ESF-interventie Opleidingen in bedrijven. What drives training transfer effectiveness and how does this transfer work? Commissioned by Departement Werk en Sociale Economie. Vlaamse Overheid: p. 11. Full report: https://www.vlaanderen.be/publicaties/impactevaluatie-esf-interventie-opleidingen-in-bedrijven-what-drives-training-transfer-effectiveness-and-how-does-this-transfer-work

IV. What are the criteria for judging the quality of the mobilisation of this method?

Several checklists circulate with overviews of what QCA good practice involves (Schneider, and Wagemann 2010; Befani 2016: 183-185), and it would exceed the scope of this methodological sheet to elaborate on all quality criteria. It is imperative that QCA as a data analysis technique is applied consistently with the 'spirit' of QCA as a research approach, which implies that QCA should not be reduced to a mechanistic 'push button process'. Besides, evaluators should be transparent about all choices made in the research process, and ideally resort to robustness tests for all decisions made. The latter is particularly important, given the strong case-sensitivity of the method.

The QCA analysis will generate different parameters of fit which help evaluating the analyses of necessity and sufficiency. Simply put, consistency describes the extent to which an empirical relationship between a (combination of) condition(s) and the outcome approximates set-theoretic necessity and/or sufficiency. Coverage describes the empirical importance or the relevance of a (combination) of condition(s). For necessary conditions, consistency is typically set very high, at 0.9; whereas for sufficient conditions, lower consistency values (e.g. 0.75) are relatively common. Coverage values should usually be 0.60 or higher. Importantly, however, the thresholds for what is deemed 'good' can vary with the research design and aim of the research (Schneider, and Wagemann 2010).

V. What are the strengths and limitations of this method compared to others?

QCA has the unique advantage to account for causal complexity, while also allowing for modest generalisation by the systematic identification of cross-case patterns. The rigorous procedures it relies on also makes the findings perfectly replicable. An additional advantage is that it does not need a large number of cases to be applied.

Strictly speaking, however, QCA will only unravel 'associations' between a condition and an outcome. The actual causal interpretation is up to the evaluators themselves. A similar limitation applies to the time element. Although various ways of including 'time' in a QCA analysis are being worked on (see Verweij, and Vis 2021), the type of findings is static in nature rather than dynamic.

Just because of these reasons, it is advisable to combine QCA with other within-case methods that have the ability to open the causal black box. In particular, the combination of QCA and Process Tracing is increasingly used for this purpose. The evaluation referred to in this methodological sheet is an example of this.

Some bibliographical references to go further

Befani, Barbara. 2016. "Pathways to change: Evaluating development interventions with Qualitative Comparative Analysis (QCA)." Sztokholm: Expertgruppen för biståndsanalys (the Expert Group for Aid Studies). Pobrane: http://eba.se/en/pathways-to-change-evaluating-development-interventions-with-qualitative-comparative-analysis-qca

Rihoux, Benoît. and Charles C. Ragin. 2008. *Configurational comparative methods: Qualitative comparative analysis* (QCA) *and related techniques*. Sage Publications.

Schneider, Carsten. and Claudius Wagemann. 2010. "Standards of good practice in qualitative comparative analysis (QCA) and fuzzy-sets." *Comparative sociology*, 9, no.3: 397-418.

Schneider, Carsten. and Claudius Wagemann. 2012. *Set-theoretic methods for the social sciences: A guide to qualitative comparative analysis*. Cambridge University Press.

Thomann, Eva. and Martino Maggetti. 2020. "Designing research with qualitative comparative analysis (QCA): Approaches, challenges, and tools." *Sociological Methods & Research*, 49, no.2: 356-386.

Verweij, Stefan. and Barbara Vis. 2021. "Three strategies to track configurations over time with Qualitative Comparative Analysis." *European Political Science Review*, 13, no.1: 95-111.

An excellent source is also www.compasss.org, which includes an extensive bibliography, an overview of software, tutorials and guidelines on QCA.

20. Theory-based Evaluation

AGATHE DEVAUX-SPATARAKIS

Abstract

Theory-based evaluation was developed in response to the limitations of experimental and quasi-experimental approaches, which do not capture the mechanisms by which an intervention produces its impacts. This approach consists of opening the "black box" of public policy by breaking down the different stages of the causal chain linking the intervention to its final impacts. The hypotheses thus formulated on the mechanisms at play can then be tested empirically.

Keywords: Qualitative methods, theory, causal chain, intervention logic, theory of change, impact pathway

I. What does this approach consist of?

Formalised in the 1970s and 1980s in the United States, theory-based evaluation (TBE) is not a method, but rather a logic of evaluative research, an analytical approach that can mobilise several evaluation methods or tools. The term "theory" is to be understood here as the decomposition and explanation of a causal chain linking the public intervention to its expected results and impacts on the targeted and beneficiary publics. This takes the form of a graphic representation which then becomes the "scaffolding" underlying the evaluation investigations. Depending on its use, this tool may be called an "intervention logic", "theory of change" or “impact pathway”. This approach is widely used in evaluation practice, although it covers a diversity of practices of varying quality and using more or less rigorous approaches.

Theory-based evaluation, contrary to what this name might suggest, seeks to be as close as possible to the reality on the ground and the context of the intervention. The term theory takes on a more or less scientific meaning depending on usage. At the very least, a theory-based approach is concerned with the theory of implementation of an intervention, i.e. the actions that are put in place for the proper implementation of the intervention in order to achieve the first results (intervention logic). In most cases, it tends to go a step further and make explicit the hypothetical causal links between an intervention and its intended outcomes up to and including final impacts (theory of change). This analysis is generally structured in several categories:

- Outputs represent the implementation of actions by public authorities,
- The results are the first immediate effects, directly linked to the achievements and which can be observed on the public participating in the actions,
- Intermediate and final impacts describe the expected impacts of the outputs on the final beneficiaries whose situation is to be improved.

This process systematically begins with the development of the ‘theory’. This is generally based on a literature review and discussions with the stakeholders of the intervention in order to identify the key components of the intervention, the target audiences, the main expected results and the final impacts targeted by the intervention. The theory is thus often co-constructed between evaluation teams and stakeholders and presents the main steps from outputs to final outcomes and impacts by connecting them and making explicit the assumptions for moving from one step to another. These hypotheses can cover the conditions for success, the risks linked to the context, but also, in the framework of a realist evaluation, the psycho-social mechanisms at work or, in the framework of a contribution analysis, specify the intervention's contribution hypotheses. At this stage, the work can be consolidated by an analysis of the literature and thus a mobilisation of academic theories to support certain causal links (for example, in an intervention targeting a return to employment, drawing on what the scientific literature says about the causal relationship between a level of training and employment).

This theory becomes the scaffolding for the evaluation. Two types of empirical investigation are then conducted:

- Firstly, an analysis of change: are the steps identified or desired when the theory was conceived verified in the field? For whom and in what contexts?
- Then a causal analysis: How can we explain the passage or not of one stage to another? To what extent does the intervention contribute to generating these results? Under what conditions ? What psychosociological mechanisms can explain it?

The evaluation questions are therefore structured around theory testing and explanation. In this task, the evaluation team can mobilise a variety of evaluation methods or tools. For example, quantitative methods for estimating outcomes or impacts can be used, and these can be combined with qualitative methods to deepen the causal analysis.

Ultimately, the results of this approach take the form of a new 'tested' theory of the intervention showing the causal processes actually at work, those that do not work as expected and explaining the reasons why. This type of result then makes it possible to identify at which stage, in which context and with which audiences the intervention encountered difficulties and to propose operational and strategic improvements for future programming.

II. How is this approach useful for policy evaluation?

Mobilising a theory of change in an evaluation can serve different purposes. Breaking down the stages and causal links allows investigations to identify which stages are functional in the theory and which are more problematic. A public intervention is never a complete failure or success. For example, if it is found that the intervention did not achieve the desired results, the theory of change help identify whether the difficulties are related to implementation problems (actions that were not properly organised or difficulty in mobilising the target audiences) or to the capacity of the intervention to have the desired effects (actions were well organised and reached the target audiences, but failed to generate the desired effects on all or certain categories of audience). It is then possible to deepen the analysis of causality to try to understand whether these difficulties are linked to the programme itself or to external contextual elements that were not sufficiently taken into account during the design or implementation of the programme (competition with other policies, constraints that hindered the participation of the public, insufficient analysis of needs, or changes in the economic situation, for example).

This approach can be used to answer questions such as: “Where did intervention Y contribute to generating outcome X and why?” More generally, it can be used to answer all types of evaluative questions on effectiveness, impact, relevance, internal or external coherence and efficiency. Since each of these areas can be investigated to explain success or failure in moving from one stage of the theory to another.

It can be used at the design stage of an intervention to bring together stakeholders, co-construct the theory and anticipate risks before implementation, and modify programming to increase the chances of success. In this case, it is also useful for the appropriation of a common and shared vision of the intervention by the stakeholders.

It is also particularly relevant in in itinere evaluations (conducted as the intervention unfolds) in order to verify during the implementation, the achievement of the different stages and to adjust the implementation accordingly. Finally, in the case of an ex-post evaluation, it allows for a comparison between the initial theory and the theory as reformulated as a result of the empirical investigation phase of the evaluation, and to understand the reasons for their differences.

III. An example of the use of this approach in the evaluation of a nutritional programme

Theory-based evaluation approaches are mobilised in a variety of contexts and are equally suited to simple projects and complex public policies. An example of its use in a simple project can briefly illustrate the added value of this approach.

In a paper for the International Initiative for Impact Evaluation, Howard White (2009) presents the case of the Bangladesh Integrated Nutrition Project (BINP) evaluation. This project identifies malnourished children and provides supplementary feeding and nutrition counselling to the mothers of these children. The ultimate goal is to increase the growth of the children.

An initial comparison group evaluation by propensity score matching found no impact of the project on the nutritional status of children, but a positive impact only on the most malnourished children. However, this result by itself does not provide any lessons on how the project is not working or what can be done to improve it.

A complementary theory-based evaluation has enriched these results and proposed orientations for action. This approach focused first on reconstructing the theory of the project and then on clarifying and investigating the causal assumptions underlying the action.

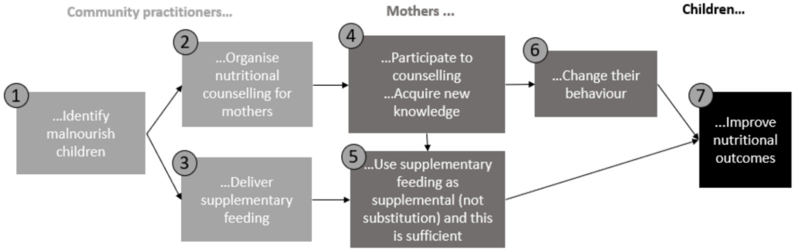

Community practitioners...
Mothers ...
Children...
1
...Identify malnourish children
2
...Organise nutritional counselling for mothers
3
...Deliver supplementary feeding
4
...Participate to counselling ...Acquire new knowledge
5
...Use supplementary feeding as supplemental (not substitution) and this is sufficient
6
...Change their behaviour
7
...Improve nutritional outcomes

Figure 1: Integrated nutrition project theory (adapted by the author of this sheet)

A first step is the correct identification of malnourished children (1). This step is based on the assumption that parents bring their children to the prevention centres and that malnourished children are well identified. Here it seems that the programme was able to reach the target audience (90% of eligible women brought their children), nevertheless the selection of children by community practitioners showed several type 1 errors (malnourished children not selected) and type 2 errors (children selected when they are not malnourished).

The 2-4-6-7 causal chain is based on the assumption that the right target audiences have been identified and that the modes of action are relevant. However, an anthropological study and focus groups revealed that the mothers targeted by this public policy device had little influence on their child's nutrition, as the men were responsible for food shopping and most mothers-in-law were in charge of preparing meals. Thus, although they participated in the training, and learned new knowledge, the mothers were only in a position to apply it when they became mothers-in-law in turn.

The 3-5-7 causal chain is based on the assumption that malnourished children were indeed selected, and that the food supplement is used as a complement and not as a substitute or shared among the children and that it is sufficient to improve the nutritional outcomes of the effects. However, the surveys conducted showed that these two assumptions were not verified in many cases.

This evaluation identified very operational recommendations to improve the effects of the project, including the participation of mothers-in-law and husbands in nutrition counselling sessions, as well as better training of community practitioners in charge of selecting children and targeting the most malnourished children.

IV. What are the criteria for judging the quality of the mobilisation of this approach?

The quality of the mobilisation of this approach can be analysed at two points in the evaluation.

First, the theory development phase should result in the most plausible theory before the field investigation. Ideally, a good theory is co-constructed between the evaluation team and the different stakeholders and also mobilises external knowledge that can inform the plausibility of the hypotheses of causal links between the different stages of this theory. Plausibility means that it reflects a reasonable ambition for the intervention in terms of the changes it is likely to generate. Finally, a good theory is synthetic, simple to understand and clearly indicates through a graphical representation who is acting and who is likely to change their behaviour or see their situation improve as a result of the intervention.

Secondly, the quality criteria for the theory-testing phase are in line with the general standards of data collection and analysis. Collection tools should be diversified to reflect a plurality of data sources (quantitative and qualitative) and points of view in order to consolidate the results as much as possible. Data collection and analysis are ideally conducted in an iterative manner (in successive phases) in order to refine the theory of change and its assumptions throughout the evaluation.

The results of the evaluation should then be presented in terms of testing the theory and showing what worked and what did not work for which audiences, why and under what conditions.

V. What are the strengths and limitations of this approach compared to others?

The main advantage of this approach is to ‘open the black box’ of public policy. Unlike experimental or quasi-experimental approaches, which make it possible to identify impacts without analysing how these were produced, theory-based evaluation makes it possible to identify and break down the mechanisms leading to the production of the impact. To do this, it applies an analytical approach that breaks down the intervention into stages and causal links and distinguishes between what is contextual and what is interventional. Explaining the assumptions underlying the action requires the evaluation team to look at the contexts of implementation, the characteristics of the audiences and the “bets” of the intervention, i.e. how it is likely to be a lever for change. The evaluation results are therefore particularly contextualised and generate lessons even when the intervention is not finalised, since each stage of implementation can be investigated.

However, this approach can be destabilising, as unlike other methods, there is not one distinct way and protocol to be followed in implementing it. Thus, many evaluation practitioners may claim to be part of this approach without meeting the quality standards of this exercise.

Furthermore, this method has been criticised for paying too much attention to the analysis of the expected theory and for obscuring the identification of alternative explanations for the observed processes or unanticipated results. This is indeed a pitfall to which the evaluation team must be attentive by attempting both in the design of the theory and in its testing to be open to the formulation and identification of such alternative explanations and identification of unexpected impacts. The use of the contribution analysis method makes it possible to focus the analysis on the role of the intervention in a more complex causal package and to investigate alternative explanations.

Some bibliographical references to go further

Birckmayer, Johanna D.. and Weiss, Carol H. 2000. "Theory-Based Evaluation in Practice, What Do We Learn?" *Evaluation Review*, 24(4): 407-431.

Funnell, Sue., and Rogers, Patricia. 2011. *Purposeful Program Theory. Effective use of Theories of Change and Logic Models*. Jossey-Bass, Chichester.

Rogers, Patricia., and Petrosino, Anthony. et al. 2000. "Program Theory Evaluation: Practice, Promise, and Problems." N*ew directions for evaluation*, 87: 5-13.

Van Es, Marjan, and Guijt, Irene, and Vogel, Isabel. 2015. *Theory of change thinking in practice*: A *stepwise approach*, Hivos ToC Guidelines, Hivos people unlimited.

Weiss, Carol. 1997. "Theory-based Evaluation: Past, Present, and Future." *New directions for evaluation*, 76: 41-55.

White, Howard. 2009. “Theory-Based Impact Evaluation: Principles and Practice.” *Working paper* 3, International initiative for Impact Evaluation.

21. Realist Evaluation

SARAH LOUART, HABIBATA BALDÉ, ÉMILIE ROBERT, AND VALÉRY RIDDE

Abstract

Realist evaluation is based on a conception of public policies as interventions that produce their outcomes through mechanisms that are only triggered in specific contexts. The analysis of these links between contexts, mechanisms and outcomes is therefore at the heart of this approach. This approach can be based on a variety of methods, but will in all cases use qualitative methods to investigate the mechanisms involved. Belonging to the family of theory-based evaluations, realist evaluation aspires to produce middle range theories that will facilitate the transfer of the knowledge produced on the intervention under study to other contexts or other interventions of the same type.

Keywords: Qualitative methods, theory-based evaluation, context-mechanism-outcome (CMO) configurations, middle range theory, critical realism

I. What does this approach consist of?

Critical realism, a school of thought led in particular by Roy Bhaskar (1975), paved the way for the development of the realist approach to evaluation. In his book A Realist Theory of Science, he asks: what must the world be like for science to be possible? The idea is to question the nature of reality, since this will then determine how it can be explored and understood. Bhaskar argues that there are structures, powers, mechanisms (e.g. gravity) that exist and can produce effects, even if we do not know them. A leaf can fall from the tree and reach the ground under the effect of gravity, even if we have not observed it. The aim is then to try to produce knowledge about the mechanisms, powers, structures that exist and the ways in which they act, including the conditions that favour their triggering and the effects they can produce. These theories are therefore 'statements about the way things act in the world' (Bhaskar 1975). The aim of researchers is then to produce theories that will try to elucidate the existence of mechanisms and the way they operate. However, these theories will always be perfectible and evolving, as reality and knowledge evolve. For 'realist' researchers, there is a reality, but there are no general truths that are valid at all times and in all places.

The realist approach to policy evaluation is based on these principles (Pawson and Tilley 1997; Westhorp et al. 2011; Westhorp 2014; Robert and Ridde 2020). It was introduced by Pawson and Tilley (1997). The presupposition is that policies have real effects, but do not cause them directly. They may or may not trigger mechanisms, which in turn produce outcomes. Mechanisms refer to the ways people react to the resources, sanctions or opportunities (depending on the type of intervention) made available to them within the policy framework (Lacouture et al. 2015). People are therefore the drivers of change; it is their reactions that will produce outcomes, whether positive, negative, expected or unexpected. It is therefore no longer enough to answer the question: does the policy work (or not), but rather it becomes necessary to investigate the mechanisms that, triggered by the policy in specific contexts, produce outcomes. This helps answer the question: how does the policy work or not, why, for whom, and in what contexts? The objective of realist evaluation is therefore to make links between the triggering of mechanisms and contextual factors, and between the action of these mechanisms and the occurrence of outcomes.

To answer all these evaluation questions, realistic evaluation mobilises a theory-based evaluation approach (see separate chapter on theory-based evaluation). This involves starting with the intervention theory of the policy under study (the way in which the policy is expected to produce its effects), and developing it in the light of existing knowledge and field data. The aim is to arrive at what is known as a middle range theory (Merton 1968). This is a theory that lies between the intervention theory and a theory that would like to be general. It is an explanatory theory of a regularity (observed trend) that is also contextualised (linked to particular contexts). To achieve this, different stages of the evaluation must be organised.

I.I. Reconstructing intervention theory

The first step is to understand the policy being implemented. There is a set of beliefs or assumptions that underpin the activities of the policy. Indeed, all policy is based on a theory, i.e. an idea that the activities implemented can produce change. It is about answering questions such as: what resources does the policy make available and why? What changes could the policy generate and how? What elements of the context might influence the policy? This intervention theory is often not explicit at the outset, and needs to be investigated. This may involve discussions with the implementation team and a review of documents. Often used interchangeably, the concepts of 'intervention theory' and 'theory of change' do have differences. Intervention theory is a detailed, intervention-centred theory describing all the components of an intervention and their logical relationship to their desired impact. From a realist perspective, it incorporates the concepts of mechanism and context. The theory of change, on the other hand, focuses on the objective of the intervention and the social change it aims to achieve: it describes the part of the logical reasoning between the expected results and the desired impacts of an intervention.

I.II. Formulate theoretical proposals on how policy can produce effects

At this stage, it is necessary to draw on both the intervention theory reconstructed in the previous stage, and on scientific knowledge. The aim is to formulate theoretical proposals on how policy activities in specific contexts might trigger mechanisms, which in turn might produce outcomes, and which ones. The aim is to work on the interactions between a particular context, the possibility of triggering a mechanism via the intervention, and the production of an outcome. The context is the set of factors external to the policy which have an influence on the triggering of a mechanism (e.g. socio-economic characteristics of the participants, social norms, interpersonal relations, political environment, etc.). Outcomes are the results produced by the triggering of these mechanisms. These interactions are therefore theoretical propositions about how the policy is expected to work and why.

I.III. Empirical testing of the theoretical propositions

On the basis of these a priori theoretical propositions, the objective is then to test them empirically, in order to confirm, revise or clarify them. This allows the theoretical propositions and therefore the intervention theory to evolve. This empirical testing may involve both quantitative and qualitative data collection methods. Indeed, realist evaluation does not necessarily rely on one type of method or data. The aim is to use a range of methods according to their relevance to test theoretical propositions. It is therefore not really a research method but rather a 'logic of inquiry' (Pawson and Tilley 2004). Nevertheless, in order to investigate the mechanisms and therefore the reasoning of the actors involved in the policy, qualitative methods will necessarily play a role at some point to understand how actors perceive and react to the policy. The aim of this step is to identify links between contexts, mechanisms and outcomes, which occur on a regular basis in the field.

I.IV. Specify the middle range theory

Based on the empirical data, the theoretical proposals formulated are therefore tested and refined. The proposals initially constructed evolve and are completed with the new data. This allows us to have a consolidated theory, which can be formalised by a set of ‘context, mechanisms, outcomes’ (CMO) configurations, to explain how, why, for whom and in which contexts this type of policy may or may not work. As the mechanisms are triggered by the policy implemented (intervention), and are necessarily linked to a person or group of people, it is possible to formulate the theories as more complete configurations: ICAMO (intervention – context – actor – mechanism – outcome). This theory is a middle range theory because it has a broader validity than the intervention theory, which is very specific to a given intervention. The middle range theory is more abstract and can be used as a basis for analysing and evaluating other policies of the same type.

II. How is this approach useful for policy evaluation?

The purpose of policy evaluation is to guide public policy by identifying the most relevant activities to be undertaken in view of the desired objectives. To do this, it is necessary to draw on the lessons learned from the implementation of past or current policies. These lessons should not be limited to assessing whether or not the objectives have been achieved. They must also draw on a critical, empirically-based assessment of the assumptions and preconceptions on which the policy was based, as well as the action strategies of those responsible for implementing it. The realist approach thus distinguishes itself from more traditional approaches to evaluation, which often aim to assess only the effectiveness of a policy, using rather quantitative indicators. These methods alone are often insufficient to draw relevant lessons from the implementation of complex policies. Most policies are complex because they operate at different levels, involve many people, change as they are implemented, are influenced by context and a myriad of factors. It is therefore useful to turn to other approaches, such as realist evaluation, which allow for the complexity of policies. It allows us to understand how a policy may, or may not, bring about change. It is an explanatory type of evaluative research.

In the critical literature on public policy, it is often pointed out that the same type of policy is disseminated without being adapted to different contexts (Olivier de Sardan 2021). This diffusion of standardised policies often does not produce the same results elsewhere. The realist approach helps to explain this and can help to avoid such pitfalls. It helps to understand what does or does not trigger the mechanisms that produce positive effects, and to understand how and why these triggers could potentially occur elsewhere. Understanding why and how public policies work, with which beneficiaries and in which contexts, provides guidance for decision-making. Asking the question 'for whom' the policy works is also a key question in policy evaluation. This is necessary to take into account the different impacts of the policy on different sub-groups, particularly the most marginalised, among whom differential, counter-intuitive or even undesirable effects may be observed. All of these questions, which are found in the realist approach, can provide guidance on whether a policy should be implemented in a different context, or how the policy can be adapted or changed to maximise its potential to produce the desired effects. It is therefore a particularly appropriate approach when the policy is intended to be scaled up and extended to other populations in other contexts.

A realist rationale, based on the results of previous evaluations or a review of the scientific literature using this approach, can be used to guide policy formulation prior to implementation. However, realist evaluation cannot be carried out only *ex ante* if no outcome data are available, since in order to develop CMO configurations, outcome data are needed.

III. An example of the use of this approach to evaluate the implementation of universal health coverage

Universal health coverage (UHC) promotes the access to health services for all people who need them, without exposing them to financial hardship. To achieve this goal, the World Health Organization (WHO) has established a partnership to support UHC in several countries. This partnership aims to support collaborative policy dialogue as a governance tool in countries that aim to implement actions for UHC. This intervention consists of providing resources and expertise (e.g. technical assistant, training for ministry officials, etc.) according to the needs of the Ministries of Health (Robert and Ridde 2020). This type of intervention is complex and takes place in very different contexts and in varying forms. It cannot be evaluated using only quantitative data and indicators. Drawing on realist evaluation has allowed for a better understanding of how this partnership can work, and of the potential differences in results depending on the context of implementation. The overall objective of the study was to understand how, in which contexts and through which mechanisms the partnership can support policy dialogue.

The objective was to investigate: 1) how and in which contexts the partnership can initiate and nurture policy dialogue; 2) how the collaborative dynamics unfold within the policy dialogue supported by the partnership (Robert et al. 2022). A multiple case study was conducted in six countries. Theoretical propositions on how policy could work were drawn from the project documents but also from existing theories in the scientific literature, for example theories on partnership relations and collaborative governance. An example of a theoretical proposition is that capacity building (through training, technical expertise and continued support from WHO) would empower the MoH (M) while triggering a shared understanding of governance and policy dialogue (M); this should lead the MoH to conduct inclusive and participatory policy dialogues (O). The triggering of these mechanisms could be facilitated by contextual factors, such as the fact that WHO and the MoH have an enduring relationship (C) or that the human resources of the two institutions involved in the partnership are stable (C).

A collaborative approach was adopted, involving stakeholders at key stages of the evaluation: development of the protocol, development of the intervention theory, interpretation of the results, etc. By drawing on theories to increase abstraction, the intervention theory as well as field data to consolidate or refute the initial theoretical propositions, several CMO configurations were formulated. For example: partnership facilitates the initiation of policy dialogue (O) by generating stakeholder interest in multi-sectoral collaboration (M), provided that stakeholders recognise their interdependence and the uncertainty of managing critical health issues (C). It can be seen that one of the outcomes that WHO expects from the establishment of the partnership will only be realised in a particular context that will allow a specific response mechanism to be triggered by stakeholders. This type of result could support the implementation of similar actions, help to adapt it, or help identify the contexts where this type of intervention is most likely to respond positively to the expected outcomes.

IV. What are the criteria for judging the quality of the mobilisation of this approach?

Realistic evaluation is more an approach to evaluation than a method. It belongs to the category of "evaluative research". In order to judge the quality of a realist evaluation, it is therefore more important to ensure that the evaluation meets certain basic criteria of the approach. For example, the evaluation should focus on discovering the mechanisms at work, and the concept of mechanism should be properly understood and applied. The evaluation should uncover configurations of contexts, mechanisms, outcomes and actors. It must allow for a greater abstraction from the intervention theory. Other elements can also favour the quality of a realist evaluation: carrying out a review of scientific literature to investigate existing theories and support the formulation of theoretical proposals to be tested; triangulating data sources (qualitative and quantitative); involving different stakeholders at different stages of the evaluation, etc. There are guides, such as the 'Quality Standards for Realist Evaluation' (Wong et al. 2016) that provide guidance at each stage of the evaluation, in order to carry out a quality realist evaluation.

V. What are the strengths and limitations of this approach compared to others?

Realistic evaluation has many advantages. It makes it possible to take into account the complexity of public policies, as well as that of the social, political and economic world in which they take place. It is based on a collaborative approach and encourages the involvement of all stakeholders (at the institutional, operational and policy recipient levels). In particular, this approach allows the "beneficiaries" and front-line people to be placed at the centre of the evaluation, by considering them as experts. It is their reactions that we try to understand so they are the ones who are best able to inform us.

It helps to explain multiple processes and outcomes, to highlight unexpected results of policies, and to answer evaluative questions that are often overlooked (understanding the how rather than just the outcome). Seeking to understand in depth how policies work provides knowledge that is more likely to be mobilised in other contexts. Its attention to context is fundamental because context is too often forgotten in standard evaluation approaches. Moreover, its grounding in theory allows for the use and accumulation of available knowledge. It allows scientific knowledge to be mobilised in a concrete way, whereas it is often not used enough in the field. The fact that it is based on both scientific literature and data from the field makes it possible to ensure a certain transferability of the results produced and to provide appropriate recommendations to political decision-makers (whether or not it is appropriate to implement this type of policy in certain contexts, how to adapt a policy to a specific context, etc.). Finally, it allows for collaboration between teams with different expertise and research areas, as well as the mobilisation of very different research methods.

Nevertheless, mobilising this approach involves some challenges. First, it is time-consuming and can be difficult to master. The concepts of context and mechanism can be difficult to grasp and operationalise. There is still a lack of dedicated courses in advanced evaluation practice and evaluative research in which realistic evaluation is taught. Moreover, many evaluation stakeholders (donors, operational partners, ethics committees, etc.) are not familiar with this approach, which can cause problems in understanding what is and what is not possible, and thus in meeting their expectations. Secondly, evaluation is still very much marked by the search for impact results measured by indicators, whereas realist evaluation offers commissioners and interested stakeholders a completely different format of results. Finally, this approach to evaluation is not straightforward, and does not produce linear results: several mechanisms can act at the same time, having opposite influences on the outcomes; an effect in one configuration can become a context in another. CMO configurations can therefore sometimes be difficult to construct.

Some bibliographical references to go further

Bhaskar, Roy. 1975. A *Realist Theory of Science*.

Lacouture, Anthony. and Breton, Eric. and Guichard, Anne. and Ridde, Valéry. 2015. “The concept of mechanism from a realist approach: a scoping review to facilitate its operationalization in public health program evaluation.” *Implementation Science*, 10(1): 153. https://doi.org/10.1186/s13012-015-0345-7.

Merton, Robert C. 1968. *Social Theory and Social Structure*. Simon and Schuster.

Olivier de Sardan, Jean-Pierre. 2021. *La revanche des contextes: Des mésaventures de l'ingénierie sociale en Afrique et au-delà*. KARTHALA Editions.

Pawson, Ray. and Tilley, Nicholas. 1997. *Realistic Evaluation*. London: Sage.

Pawson, Ray. and Tilley, Nicholas. 2004. *Realist Evaluation*. DPRN Thematic Meeting 2006 Report on Evaluation.

Robert, Emilie. and Ridde, Valéry. 2020. *Dealing With Complexity and Heterogeneity in a Collaborative Realist Multiple Case Study in Low- and Middle-Income Countries*. SAGE Research Methods Cases. SAGE Publications Ltd. https://doi.org/10.4135/9781529732306.

Robert, Emilie. and Ridde, Valéry. and Rajan, Dheepa. and Sam, Omar. and Dravé, Mamadou. and Porignon, Denis. 2019. “Realist Evaluation of the Role of the Universal Health Coverage partnership in Strengthening Policy Dialogue for Health Planning and Financing: A Protocol.” *BMJ Open*, 9(1): e022345. https://doi.org/10.1136/bmjopen-2018-022345.

Robert, Emilie. and Zongo, Sylvie. and Rajan, Dheepa. and Ridde, Valéry. 2022. “Contributing to Collaborative Health Governance in Africa: A Realist Evaluation of the Universal Health Coverage partnership.” *BMC Health Services Research*, 22(1): 753. https://doi.org/10.1186/s12913-022-08120-0.

Westhorp, Gill. 2014. “Realist Impact Evaluation: An Introduction.” Methods Lab.

Westhorp, Gill. and Prins, Ester. and Kusters, Cecile. and Hultink, Mirte. and Guijt, Irene. and Brouwers, Jan. 2011. “Realist Evaluation: An Overview.” Seminar report.

Wong, Geoff. and Westhorp, Gill. and Manzano, Ana. and Greenhalgh, Joanne. and Jagosh, Justic. and Greenhalgh, Trish. 2016. “RAMESES II reporting standards for realist evaluations”. *BMC Medicine*, 14(1): 96. https://doi.org/10.1186/s12916-016-0643-1.

22. Contribution analysis

THOMAS DELAHAIS

Abstract

Contribution Analysis is a theory-based evaluative approach particularly suited to the evaluation of complex interventions. It consists of progressively formulating "contribution claims" in a process involving policy stakeholders, and then testing these hypotheses systematically using a variety of methods (which may be qualitative or mixed).

Keywords: Mixed methods, complex interventions, contribution claims, abductive approach, context, causal pathways, causal packages, narrative approach

I. What does this approach consist of?

Contribution analysis is a so-called theory-based evaluation approach[1] (TBE): it is organised around a process of 1) developing a set of hypotheses about the effects of an intervention being evaluated (how these effects are achieved, in which cases, why…) – known as the 'theory of change'; 2) testing these hypotheses through the collection and analysis of empirical information; and 3) updating the original theory by indicating which hypotheses are verified.

[1] Many thanks to Kevin Williams for his help with the translation in English of this brief initially written in French.

Like Realist Evaluation or Process Tracing, for example, Contribution Analysis is part of the new generation of TBEs that emerged at the start of the 2000s (sometimes referred to as theory-based impact evaluations – TBIE). It considers the interventions being evaluated as complex objects in complex environments. Central to Contribution Analysis is the postulate that interventions do not intrinsically 'work'; their success or failure always depends on a diversity of drivers and contexts, which the evaluation needs to document. This is in contrast with Counterfactual approaches, for instance, which aim at identifying "what works" in isolation from their context. But what distinguishes Contribution Analysis from other approaches is that it also rejects the idea that the role of evaluation is to establish impact irrefutably: in a complex context, its aim is not to prove the effects of interventions, but to reduce uncertainty about their contribution to any changes that have occurred. It is in fact uncertainty that can be considered to be detrimental to decision-making and policymaking more generally.

I.I. Theory-building

The whole process of Contribution Analysis thus consists of gradually reducing uncertainty about the effects of the intervention being evaluated. As with all TBEs, the first phase, theory-building, consists of asking a question about the cause-and-effect relationships that are to be investigated and developing causal assumptions in response to this question. The latter usually concerns the contributions of the intervention to the desired changes. Let us imagine a governmental plan to prevent or deal with sexual violence in higher education institutions. The question asked might be: "How has the plan contributed to the effective reduction of sexual violence and better management of its consequences?"

The level of violence and the responses provided in institutions, however, are societal changes that are only partially dependent on any ministerial plan. Indeed, Contribution Analysis does not assume that these changes are due to the intervention. Rather, it assumes that any change is the result of a multitude of intertwined causes, including (perhaps) the intervention. Thus, Contribution Analysis starts from the change (*in this case, the evolution of sexual violence*) to look for contributions, rather than from the intervention being evaluated (*the governmental plan*).

The focus of Contribution Analysis in this initial phase is therefore to make explicit what the contribution of the intervention might be (among other factors) and to ensure that such a contribution is plausible. Plausible means that the contribution, while not verified, is nevertheless likely: it could occur in the context of the intervention being evaluated.

The more complex the setting of the intervention, the longer this initial investigation may take. The plausibility of an assumption is not judged in abstracto: it is assessed on the basis of the convergence between the observations, experiences and informed opinion of the stakeholders, its proximity to assumptions validated in other settings presenting similarities to the intervention being evaluated, the possible significance of the intervention in relation to other factors, initial indications of a possible effect, etc.

This phase is usually based on an initial collection of empirical data (exchanges with stakeholders, a literature review or a document analysis) that leads to "contribution claims" and alternative explanations (i.e. claims about other factors that could plausibly explain the observed changes). In our case, an evaluation would look at changes in sexual violence and institutional practices over the past few years to identify possible contribution claims. If a number of institutions have drastically changed their practices in this area, it may be because the plan included an obligation to put in place strategies to combat sexual violence and to report on progress annually; or it may be that actors already in favour of active approaches to sexual violence in the administration have used the plan to support their internal agenda; or it may be that student groups have used the plan to bring reluctant administrations to act. Each of these three assumptions, if supported by examples, a convincing theoretical framework, etc., can become a contribution claim.

At this stage, the level of uncertainty regarding the effects of the intervention has already been reduced compared to the initial situation: some claims have been rejected, others appear more or less plausible at the current phase of the evaluation. Those that are retained are studied in the next step.

I.II. Theory-testing

Only those claims that are sufficiently plausible (or those considered particularly important to stakeholders) are tested in depth. In Contribution Analysis, a very wide range of tools or methods, both qualitative and quantitative, can be used to estimate changes and to collect evidence in favour of or against the contribution claims, in combination with other factors. In this process, contribution claims are not validated or discarded. Rather, they are progressively fleshed out, for example from "the intervention contributes in such and such a way" to "when conditions x and y are met, the intervention contributes in such and such a way, unless event z occurs", leading to "causal packages" that bring together several factors associated with observed changes. Contribution Analysis can also focus on identifying the impact pathways and underlying mechanisms that explain these contributions. For example, in our case, perhaps the testing phase would show that putting the issue of sexual violence into the framework of the accountability relationship between the ministry for Education and the educational institutions had direct consequences in terms of setting up a helpline for reporting violence; but that not all ministries really grasped this issue in their accountability relationship with institutions within their remit.

Ideally, the next step in the data collection process would be to check whether such helplines for reporting violence exist in the institutions supervised by other ministries, and why.

Contribution Analysis does not impose any particular approach to infer causality. One possible way is to identify a series of empirical tests, as in the Process tracing approach. These tests each define a condition that must be satisfied in order to conclude that the intervention contributes to the observed changes. Tests may also be conducted for other factors that could plausibly explain the changes. All the tools of evaluation and, more broadly, of the social sciences, whether qualitative or quantitative, can be used to conduct these tests: interviews, case studies, documentary analyses, as well as surveys, statistical analyses, etc. can be used. The combination of different tools makes it possible, through triangulation, to strengthen (or reduce) the degree of confidence in the contribution and to arrive at the findings and conclusions of the evaluation. Realist Evaluation can also be used here to identify mechanisms underlying causal relationships.

A final specificity of Contribution Analysis is that it ultimately generates contribution stories. The contribution story initially brings together the contribution claims, which are gradually reinforced through collection and analysis. It is intended to consolidate, complement or challenge the dominant narratives underlying the intervention being evaluated. Unlike a counterfactual evaluation, for example, which seeks to be convincing through quantification, Contribution Analysis relies on narratives supported by evidence, which can then be used in the making of public policy. In our case, perhaps the contribution narrative would show how stakeholders already involved in the fight against sexual violence have seized on the governmental plan to tip the balance in their favour in the internal governance of institutions, to the detriment of a national narrative based on state control of the practices of educational institutions.

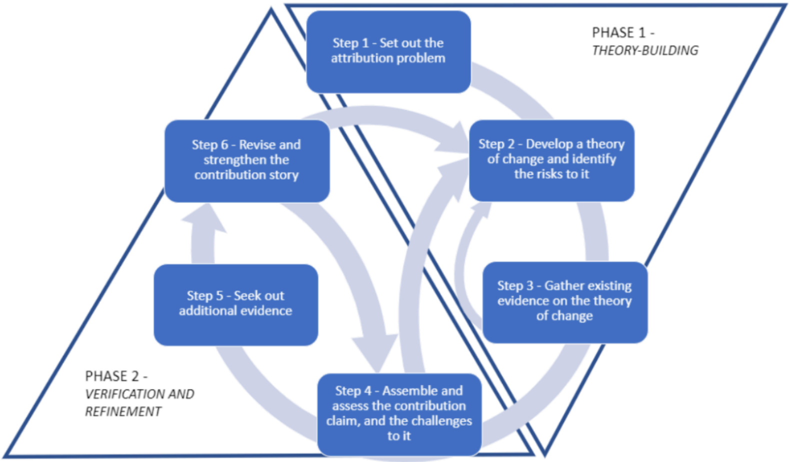

Step 1 - Set out the attribution problem
PHASE 1 -
THEORY-BUILDING
Step 6 - Revise and strengthen the contribution story
Step 2 - Develop a theory of change and identify the risks to it
Step 5 - Seek out additional evidence
Step 3 - Gather existing evidence on the theory of change
PHASE 2 -
VERIFICATION AND
REFINEMENT
Step 4 - Assemble and assess the contribution claim, and the challenges to it

Figure 1: A two-phase process (Ton, Giel. 2021. "Development Policy and Impact Evaluation: Learning and Accountability in Private Sector Development". In Handbook of Development Policy, by Habib Zafarullah and Ahmed Huque, 378-90. Edward Elgar Publishing. P. 380.

II. How is this approach useful for policy evaluation?

Contribution Analysis is mainly used ex-post, although there are attempts to use it in support of policy implementation. It is particularly useful in cases where the contribution of an intervention to the expected changes is very uncertain, or seems unlikely, but the contribution is of strategic interest to stakeholders: for example, because expectations are very high for the contribution, or because the continuation of the intervention depends on it.

The work done in the theory-building phase, because it allows for the formulation of plausible contributions that often deviate from stated objectives, is particularly useful for strategic management or redesign of interventions.

Contribution Analysis lends itself particularly well to collaborative or participatory approaches, allowing stakeholders to discuss contribution claims and the conditions under which they are likely to be verified or not. The contribution stories it produces, if discussed and owned by stakeholders, provide a useful basis for strategic reorientations. In their final form, the contribution claims, because they are explanatory and contextualised, are also useful to improve the intervention or to modify the practices of the actors involved.

III. The example of a Foundation's contribution to research in the life sciences

A Foundation[2] provides long-term support (funding, guidance) to high-level research teams and institutions in the field of life sciences. The Foundation's management is aware that the results of the work funded cannot be solely attributed to their support: the research teams are in fact the main driving force behind the results obtained; they generally rely on a plurality of funding sources; they are part of research trends, build on past research and work in conjunction with other teams around the world. Finally, the contributions that the Foundation can make are inseparable from the research context (underfunding of research in France, international competition, etc.). Nevertheless, its managers believe that its contribution can be significant, and they wish to explore it.

Given the diversity of the projects supported (individual or collective support, research work, equipment, multidisciplinary approaches, etc.), several theories of change are initially designed and fed by an exploratory data collection exercise (documentary analysis, interviews). This initial

[2] This example from an actual evaluation has been simplified for teaching purposes.

phase leads to a first draft of a ‘meta-theory’ of change (bringing together the different theories developed), in which a certain number of contribution claims are proposed. These differ in particular according to the maturity of the supported project, and the type of support. For each of these contribution claims empirical tests are developed, with a view to estimate the degree of confidence that can be placed in the veracity of these contributions. These claims are then scrutinised through related tests in a series of case studies of projects supported by the Foundation.

The cross-analysis of the case studies allows the Foundation’s contributions to the projects it supports to be refined and detailed. In total, eight main contributions are identified, through different channels: for example, funding from the Foundation can contribute to the sustainability of a project through its long-term commitment, but also because it brings credibility to the project, which can then attract other funding. The Foundation does not always activate these eight contributions, but its value added is more important when several are observed on the same project. The contribution story emphasises that these contributions draw on a set of common explanatory factors: for example, the relevant choice of researchers who know how to use additional funding to go further, or to test what they would not otherwise have been able to test; or the relationship of trust that has been established, with a great deal of freedom given to the research teams (which is reflected in particular in minimal expectations in terms of reporting on the funding). This human dimension is also what explains why its contributions are more significant in supporting research teams than institutions. The evaluation thus feeds into to the Foundation’s strategic development by identifying the situations in which its contribution can be most important and the choices that a reorientation would imply in terms of human and financial resources.

IV. What are the criteria for judging the quality of the mobilisation of this approach?

The quality of Contribution Analysis is essentially judged by the ability to work along a continuum of plausibility, which means being able to start with a number of assumptions about the factors underlying the observed changes, including the intervention, to review them, identify the most plausible ones, and then test and build on them.

In recent years, the term Contribution Analysis is sometimes used as a seemingly flattering synonym for theory-based evaluation. The main criteria to differentiate them include:

1. the iterative (so-called abductive) approach of Contribution Analysis (assumptions are constantly revised throughout the evaluation);
2. the fact that the search for contributions starts with the expected changes and works backwards to the intervention, rather than the other way around;
3. the collection of information to progressively contextualise and explain the contribution claims;
4. the care taken to test alternative explanations;
5. the narrative dimension of the results, in the form of a contribution story.

V. What are the strengths and limitations of this approach compared to others?

Contribution Analysis provides credible and useful findings for the design of public policy in very specific situations, which initially seem very complicated to evaluate. It owes its credibility to it being an iterative process, which can be made transparent in a participatory approach. Due to stakeholders being involved at each stage and the traceability of the tests carried out, as well as the humility of the approach, it gives rise to a high degree of trust, which is a precondition to the use of the results.

Nevertheless, it should be borne in mind that the process of Contribution Analysis is itself uncertain: it is not known at the outset which contribution claims will be tested and how. It is usually necessary to keep an open mind at the beginning of the evaluation to understand the context in which the intervention is situated and which interventions or factors explain the observed changes. This initial phase, which consists of describing the changes observed, is what makes Contribution Analysis so interesting compared to other approaches that tend to examine interventions out of their contexts. However, this phase can be extremely time-consuming, especially as it relies heavily on secondary sources, external to the evaluation, and the right degree of descriptive thickness must be found, given that evaluation does not usually aim at being comprehensive.

As with any TBE, there is a risk of overestimating contributions, although starting with the changes rather than the intervention itself reduces this risk. One solution is the systematic application of empirical tests to the intervention and to alternative explanations. However, this can be cumbersome and confusing, especially when the tests are too numerous or poorly calibrated (i.e. they do not allow for sufficient variation in the degree of confidence in a contribution claim).

It should be noted that whereas realist evaluation and process tracing are used more at the project level or to test a single impact pathway, Contribution Analysis is used more at the programme or policy level, when there are many actors involved and many impact pathways. This broader focus is what makes Contribution Analysis interesting, but it reinforces the uncertainties described above.

Some bibliographical references to go further

Contribution Analysis was first developed by John Mayne in the late 1990s. The following two articles can be read, the first one marking the beginning of the consideration of complexity by Contribution Analysis and the second one presenting a state of the debates and developments of Contribution Analysis in 2019:

Mayne, John. 2012. Contribution Analysis: Coming of age?. *Evaluation*, 18(3): 270-80. https://doi.org/10.1177%2F1356389012451663)

Mayne, John. 2019. Revisiting Contribution Analysis. *Canadian Journal of Program Evaluation*, 34(2). https://doi.org/10.3138/cjpe.68004

The following articles reflect the progressive operationalisation of the approach in the 2010s. The first reports on a number of practical obstacles and ways in which practitioners can overcome them; the second is an emblematic example of a situation in which the intervention being evaluated is clearly not the main driver of the expected changes; the third gives an example of the use of contribution analysis in private sector development:

Delahais, Thomas. and Toulemonde, Jacques. 2012. Applying Contribution Analysis: Lessons from Five Years of Practice. *Evaluation*, 18(3): 281-93. https://doi.org/10.1177/1356389012450810

Delahais, Thomas, and Toulemonde, Jacques. 2017. Making Rigorous Causal Claims in a Real-Life Context: Has Research Contributed to Sustainable Forest Management?. *Evaluation*, 23(4): 370-88. https://doi.org/10.1177/1356389017733211

Ton, Giel. 2021. Development Policy and Impact Evaluation: Learning and Accountability in Private Sector Development. In *Handbook of Development Policy* by Zafarullah, Habib. and Huque, Ahmed. 378-90. Edward Elgar Publishing. https://doi.org/10.4337/9781839100871.00042

23. Outcome Harvesting

GENOWEFA BLUNDO CANTO

Abstract

Outcome harvesting is a qualitative approach to *ex post* evaluation of social change results. Rather than testing for a specific impact of the intervention at hand, it consists in, in collaboration with stakeholders: broadly making sense and collecting evidence on outcomes, investigating how these were produced and whether and how the intervention may have played a role in this process, before substantiating these outcomes with external sources. Outcome harvesting is a utilisation-focused approach: it aims at producing knowledge for action. It is particularly useful in the case of complex interventions. In the case of complex interventions, when the effects of an intervention are previously not known or identified, or when the intervention has been significantly modified since its inception.

Keywords: Qualitative methods, case study, observable changes, outcome statements, harvest, impact

I. What does this approach consist of?

Outcome harvesting is a qualitative approach to monitoring and evaluation in which harvesters, meaning evaluators, identify, formulate, verify, analyse and interpret outcomes. Outcomes are defined as observable changes in the behaviour of individuals or collective actors such as groups, communities, organisations, institutions. Outcomes are changes in the practices, relationships, policies, actions, and activities of these individuals or actors. An outcome could be the increased enrolment of women in vocational training or the use of an innovative management practice by a farmers organisation. Outcomes can be positive or negative, intended or unintended, expected or unexpected. They are not to be confused with impacts, or the ultimate consequences of the intervention in wellbeing dimensions of the targeted beneficiaries, such as health or education. In outcome harvesting terms, impacts (e.g. in health, education) cannot be achieved without behaviour change (e.g. application of feeding practices, training attendance). Outcome harvesting therefore focuses on these behaviour changes.

Outcome harvesting provides evidence on what outcomes were achieved and how an intervention contributed, and the meaning of these outcomes in light of the evaluation questions. Outcome harvesting does not focus on the progress or achievement of intended, expected or planned outcomes of the intervention, but collects evidence on what has been achieved and works backwards to identify whether and how the intervention contributed to these changes. In the harvest, all the changes that actually happened are collected, which allows to capture also unintended outcomes, positive and negative. The evaluator facilitates their identification and the search for evidence of how they were achieved, and of the contribution of the intervention. Outcome harvesting is utilisation-focused: its purpose is to serve uses of the evaluation findings by the intended users, meaning those who will make decisions based on these findings. The focus is on learning from the evaluation in order to take action.

The key element of outcome harvesting is outcome statements that describe the change, who made it, when and where, what was the plausible contribution of the intervention's activities, strategies and outputs to each change, and the significance of the change. As an example: since 2015, the agricultural department of Senegal has been issuing a newsletter presenting meteorological forecasts and targeted agricultural recommendations using a language and format adapted to farmers and the wider public, democratising access to scientific knowledge. The intervention contributed to this change through a study on farmer’s preferred communication channels and formats and a series of scientific communication trainings.

Outcome harvesting is carried out in six steps, through which outcome statements are refined and evidenced. The first step is to design the harvest in order to respond to the intended uses of the findings, defined by their primary users. The second step is to review documentation to identify and formulate draft outcome statements. The third step is to engage with the people who know how the intervention contributed to these changes, who review and refine the outcome statements with the evaluator until a set of precise statements is identified. Clearly defining what has changed, the contribution of the intervention and the significance of the change allows us to bound which outcomes are taken forward for the assessment and which are not. The idea is to strive for fewer verifiable outcomes for which to collect evidence in the next step. The fourth step is the substantiation of the outcomes with external sources that are independent from the intervention but knowledgeable about the outcome and can validate the contribution of the intervention. Substantiation allows us to verify the accuracy of outcomes, but also to enrich understanding of the change and the contribution of other actors or interventions. Other changes linked to the intervention that the primary users had not identified can emerge during the substantiation step. The fifth step is to analyse and interpret the outcome statements, systematising the evidence to answer the evaluation questions defined in the design step. The sixth step is about supporting users to use findings for their intended uses.

The 6 steps of outcome harvesting

1. Design the harvest
2. Review documentation, draft outcome statements

3. Engage with stakeholders
4. Substantiate the outcomes with external sources
5. Analyse and interpret outcome statements
6. Support use of findings by stakeholders

II. How is this approach useful for policy evaluation?

Outcome harvesting is a retrospective or ex post evaluation method, but it can also be used to monitor progress during an intervention. It is particularly useful for complex interventions where the programming context is unpredictable, uncertain, dynamic, and planned courses of actions are likely to change. Outcome harvesting is especially appropriate to: 1) monitor and evaluate interventions on emerging challenges for which little information and evidence exists; 2) generate evidence of what changes were achieved by an intervention that did not pre-define changes or had very general pre-defined changes; 3) provide evidence on changes generated directly and indirectly, intended or unintended; 4) evaluate an intervention that has changed significantly from what was originally planned.

This approach can be used to monitor an ongoing intervention by periodically and systematically collecting information and learning on the social changes achieved and the intervention's contribution to them. It can be combined with qualitative or quantitative methods that address other evaluation questions. For instance, it can complement findings of methods that quantify the well-being related impacts to which have led the changes harvested.

Outcome harvesting focuses on the contribution of the intervention's actions to social change, or the plausible and logical relationship between intervention and change, rather than the exact part attributable to this action. Unlike other contribution-based approaches (see separate chapter on contribution analysis), outcome harvesting does not depart from investigating the cause-effect link to intended outcomes. On the contrary, it collects "all" the observed results and then works backwards to reconstruct the contribution of the intervention to these changes. Indeed, the main interest of outcome harvesting lies in its ability to account for the dynamics of social change by adopting an open and broad approach in identifying these changes, even when they are unexpected or unintended. For this reason, it is particularly suited to the evaluation of complex interventions where uncertainty, dynamic contexts and adaptation are common.

III. An example of the use of this approach: the scaling of a weather and climate information policy in Senegal

The following presents a case study using outcome harvesting within a broader evaluation approach to assess outcomes of the scaling of Weather and Climate Information Services (WCS) in Senegal and how research actions contributed to these outcomes (Blundo-Canto et al. 2021). WCS are the production, translation, transformation, transmission, access and use of scientific information on weather and climate to support decision-making. In Senegal, the dissemination of weather and climate forecasts along with recommendations for economic sectors and actors expanded from pilot research projects to a national-level strategy over two decades. The evaluation of the outcomes of this scaling process had an accountability and a learning objective, and was based on three components. The reconstruction of the history of the innovation (Douthwaite et Ashby 2005): the detailed timeline and interconnection of events, factors, actions and actors that marked the scaling of WCS. The Outcome harvesting (Wilson-Grau 2018): the assessment of changes in observable practices, relationships, policies, actions, and activities of the actors involved in and affected by the scaling of WCS, and the contribution of research partnerships to these outcomes. The analysis of the impact pathway (Douthwaite et al. 2003): the causal chain leading from research actions of the National Agency of Civil Aviation and Meteorology (ANACIM by its French acronym) and its partners to the outcomes identified and their perceived medium and long-term effects.

The six steps of the outcome harvesting approach guided the implementation of the three components. In the design step, evaluators discussed with the leading change agent (ANACIM) the design of the harvesting, identifying documentary sources and key actors that would need to participate in the formulation step. In step two, existing documentation was reviewed to reconstruct the history of the innovation and pre-identify outcomes that could be linked to the scaling process, which were discussed with the change agent. In the formulation step, a workshop with 16 representatives of national and local actors involved in the scaling process was organised to reconstruct its key events, actors, actions, and contextual factors. The workshop allowed us to identify other actors involved in the process and some additional outcomes. The outcomes identified in the previous three steps were systematised and formulated, and subsequently substantiated through individual interviews with 44 knowledgeable informants. These informants were independent from the leading change actor, the ANACIM. Nonetheless, the particularity of the policy process studied, a scaling up from a pilot project to a nationally-wide action involving many actors and sectors made so that some of the participants of the workshop were interviewed in the substantiation process to validate and provide evidence of the outcomes for which they were knowledgeable informants. With the evidence from the substantiation step, outcomes were refined as well as the contribution of the research actions of the ANACIM and its partners along with other actors and factors. In order to support use, three restitution workshops at the national and local level were carried out, in which results of the outcome harvesting were presented and discussed by participants, as well as possible ways forward based on the knowledge generated.

The key findings of the approach combining outcome harvesting, innovation histories and impact pathway analysis can be summarised as follows. In the past two decades, Weather and Climate Services have been serving as key policy instruments to tackle increased rainfall variability and extreme climate events that affect vulnerable rural communities in the West African Sahel. The innovation history starts in the 1980s when, following a devastating drought, the Agriculture-Hydrology-Meteorology regional centre created the first multidisciplinary working group to facilitate the development of WCS, their interpretation, and their dissemination and uptake. In the 2000s, the Senegalese ANACIM partnered with national and international research actors to set up pilot decentralised multidisciplinary working groups to facilitate uptake of forecasts and recommendations at local level. Climate information was combined with agricultural advice in a language that used local concepts, habits and practices to make this information actionable by farmers. The multidisciplinary working groups met regularly during the rainy season to discuss transmission of forecasts and recommendations. Individuals that were influential in their farming communities were trained on forecasting concepts and interpretation in order to increase uptake by other farmers. In the years leading to 2018, WCS and multidisciplinary working groups were sequentially scaled up to most departments in Senegal.

The outcome harvesting allowed us to identify how climate information was incorporated in sectoral and national adaptation plans, strategies and programmes, as well as in the coordination of actions of multiple actors at the local level. It also allowed us to identify other sectors beyond agriculture, including fisheries, energy and water resources protection that were using WCS, showing that the outcomes generated crossed institutional, sectoral and governance boundaries. Beyond the actions of the ANACIM and its partners, this process was supported by a favourable global funding environment. By combining outcome harvesting with impact pathway analysis, it emerged that the scaling of WCS to new users, sectors and uses happened through five axes: 1) continuous improvement of WCS, 2) emergence and consolidation of WCS facilitators, 3) inclusion of WCS in action planning, 4) active mobilisation to sustain the scaling of WCS, and 5) empowerment of actors to take up new roles. The factors underlying the scaling process can be summarised as intended actions by the research partners, including capacity strengthening, knowledge sharing and action platforms, and creation of interaction opportunities; national and international financial support; and an enabling political environment. The continuous improvement of WCS through feedback from its users reinforced the scaling process, resulting in increased access to WCS for the population. The outcome harvesting also allowed to capture challenges raised by the expansion of WCS as new users and uses emerged. There is a growing demand for improved quality and finer-grain WCS that are delivered at the right time to make decisions, which requires significant investment. Issues of trust in who delivers the information, how it is produced, and how to interpret it can be an obstacle as WCS reaches more users but capacity building campaigns do not. Public-private partnerships could play an important role, but at present, the involvement of the private sector in delivering WCS is limited. The results of the evaluation were used for accountability of the research partners involved, for the production of scientific knowledge on the scaling of these policy instruments, and to discuss key challenges that the actors involved in WCS production, dissemination and use need to overcome.

IV. What are the criteria for judging the quality of the mobilisation of this approach?

Outcome harvesting focuses on evaluating outcomes, not impacts. It does not focus on counting beneficiaries or measuring the well-being effects they experience. It is important to make this explicit when judging the quality of this method. Its purpose is to identify intended, unintended, expected or unexpected changes in practices, relationships, policies, actions, or activities and assess the contribution of an intervention to these changes. Such an evaluation focuses on the use of findings by intended users for their intended uses, therefore its quality needs to be judged in terms of these intended uses (e.g. accountability or learning for adaptive management). For instance, the substantiation step is important when the purpose is accountability and final assessment but can be overlooked or carried out in a lighter way when the purpose is internal learning or monitoring.

To judge outcome statements, outcome harvesting uses the SMART criteria: each statement needs to be Specific, sufficiently detailed to be appreciated by any reader; Measurable, providing verifiable quantitative and qualitative information; Achieved, a plausible link between the contribution of the intervention and the outcome can be established; Relevant, the outcome is significant in light of the intervention's purpose; Timely, the outcome occurred close to the time the evaluation is carried out, even if the contribution of the intervention happened a significant time before.

V. What are the strengths and limitations of this approach compared to others?

The key feature that makes outcome harvesting stand out compared to other evaluation methods is its focus on achieved outcomes independently of whether they had been planned or not, allowing to capture unintended or unexpected outcomes, both positive and negative. The method provides a systematic and structured way to identify these changes and to work backwards to determine whether and how the intervention contributed to them. Outcome harvesting produces quantitative and qualitative data to describe outcomes. However, it does not provide a quantitative assessment of these outcomes. Rather, it informs on the processes and strategies that have led to a quantitative outcome measured with other methods. It cannot be used for impact measurement. Applications of outcome harvesting that do not carry out the substantiation step, are better used for internal learning than for accountability as they do not include validation from independent and knowledgeable sources.

As other utilisation-focused approaches, outcome harvesting focuses on making the evaluation useful for its users and for the intended uses of the evaluation findings. The intended uses can be learning, decision-making, planning, accountability, informing partners, and so on, depending on what is agreed with the primary users at the design stage. This choice guides how the method is applied and the weight given to the substantiation of social change outcomes and the intervention's contribution. The extent to which the contribution of the intervention is assessed in the harvest will be higher when the use is accountability at the end of an intervention than when the intended use is learning for adaptive management during the intervention.

Some bibliographical references to go further

This methodological note draws significantly from:

Wilson-Grau, Ricardo. 2018. *Outcome Harvesting: Principles, Steps, and Evaluation Applications*. IAP.

Blundo-Canto, Genowefa. and Andrieu, Nadine. and Soule Adam, Nawalyath. and Ndiaye, Ousmane. and Chiputwa, Brian. 2021. "Scaling Weather and Climate Services for Agriculture in Senegal: Evaluating Systemic but Overlooked Effects". *Climate Services*, 22 (April): 100216. https://doi.org/10.1016/j.cliser.2021.100216.

Additional references mentioned:

Douthwaite, Boru. and Ashby, Jacqueline. 2005. "Innovation Histories: A Method for Learning from Experience". *The Institutional Learning and Change* (ILAC) *Initiative*, 4. https://hdl.handle.net/10568/70176.

Douthwaite, Boru. and Kuby, Thomas. and van de Fliert, Elske. and Schulz, Steffen. 2003. "Impact pathway evaluation: an approach for achieving and attributing impact in complex systems". *Agricultural Systems, Learning for the future: Innovative approaches to evaluating agricultural research*, 78(2): 243-65. https://doi.org/10.1016/S0308-521X(03)00128-8.

24. Cultural Safety

LOUBNA BELAID AND NEIL ANDERSSON

Abstract

A concept originating from nursing sciences and here applied in an innovative way to evaluation, cultural safety refers to an approach aimed at ensuring that the evaluation takes place in a "safe" manner for the stakeholders, and in particular for the minority communities targeted by the intervention under study, i.e. that the evaluation process avoids reproducing mechanisms of domination (aggression, denial of identity, etc.) linked to structural inequalities. To this end, various participatory techniques are used at all stages of the evaluation. Cultural safety is compatible with all types of methods. It contributes to making the evaluation more relevant and useful for stakeholders and will likely increase their self-determination.

Keywords: Mixed methods, participation, indigenous evaluation, culturally responsive evaluation, inequalities, racism, decoloniality, fuzzy cognitive mapping

I. What does this approach consist of?

The cultural safety approach was developed in response to the observation that the evaluation process could reproduce mechanisms of domination linked to structural inequalities, particularly concerning indigenous peoples or in post-colonial contexts. For example, one study assessed the perception of three psychometric scales used to diagnose depression in the Inuit and Mohawk populations of Quebec. The study results showed that the three scales were not culturally safe. Participants disliked the numerical assessment, the self-report (as opposed to supportive interaction) and the focus on symptoms rather than supportive factors (Gomez Cardona et al. 2021).

Cultural safety aims to produce 'an environment for people where there is no aggression, challenge or denial of who they are, what they need' (Williams 1999). A Maori nurse originally developed the concept in response to the racism and discrimination faced by Maoris in healthcare settings (Papps and Ramsden 1996). Culturally safe evaluation aims for stakeholders to feel that their cultures are respected and strengthened by the evaluation.

Cultural safety goes beyond another approach more commonly promoted in evaluation based on cultural sensitivity and competence (culturally responsive evaluation). Indeed, beyond simply paying attention to cultural differences, cultural safety considers the power imbalances, institutional discrimination, racism and colonial relations that can interfere with the design and implementation of services and programmes (Curtis et al. 2019). The concept is thus situated within the spectrum of critical postcolonial theories and aims for social justice. The concept of cultural safety has been extended beyond Maori communities to any group that differs from its care or service providers regarding age, gender, socio-economic status, ethnicity, religion or disability (Smye and Browne 2002).

The concept has attracted the attention of research and evaluation approaches that challenge unidirectional and conventional perspectives centred on the point of view of the person conducting the study, with a minimal contribution or benefit to the participants (Smith 2012; Cram 2016; Katz et al. 2016). It also calls for a revisiting of concepts, methods, values and evaluation approaches stemming from Western epistemologies and draws attention to the validity of worldviews of indigenous and minority groups (Smith 2012; Belaid et al. 2022)

Cultural safety aims to ensure that evaluation is beneficial and relevant to these communities. It aims to empower them and can contribute to increasing their self-determination (Gollan and Stacey 2021). Cultural safety changes the direction of evaluation by incorporating the participants' perspective.

Five key principles characterize cultural safety in evaluation (Wilson and Neville 2009; Cameron et al. 2010; Andersson 2018): (i) participation (ii) partnership (iii) ownership (iv) critical reflexivity and (v) protection of identities, beliefs, cultural values and worldview.

1. *Participation*: refers to the involvement of stakeholders throughout the evaluation (Cameron et al., 2010). Procedural or symbolic participation should be distinguished from genuine participation. Procedural participation allows for structured stakeholder input at specific stages of the process, for example, structured interviews with key informants. Token participation may involve a paid or unpaid stakeholder 'representative' participating in certain evaluation activities. Authentic participation includes co-ownership of the evaluation, active engagement in the analysis of evidence and a role in designing solutions based on the evaluation findings.

2. *Partnership* formalizes participation, often evolving into genuine participation as the evaluation unfolds. It is about establishing equitable relationships between the evaluation team and stakeholders, whether they are communities, patients or staff. The evaluation team should clarify the potential scope of these relationships at the outset, strive to maintain them, and allow them to evolve as stakeholders increase their capacity throughout the evaluation (Cameron et al. 2010).

3. *Ownership of the evaluation process, results and governance*: cultural safety enables stakeholders to take ownership of the process (Andersson 2018). This can start with circumscribing the purpose of

the evaluation within the limits of the funded objectives, or 'having a voice' in what is being evaluated and how, and being involved in the evaluation activities, including interpreting the results.

4. *Critical reflexivity*: The starting point for culturally safe evaluation is for evaluators to reflect on their values and beliefs, social position, power and privilege (Wilson and Neville 2009; Browne et al. 2016). This involves an awareness of the historical relationship between evaluators and indigenous and minority communities, the history of colonialism, and the systemic racism and discrimination that these communities may still face (Cameron et al. 2010).

5. *Protection* strengthens research ethics by protecting indigenous and minority groups from exploitation and reinforcement of negative representations or accounts (Wilson and Neville 2009). This implies that their knowledge, values and epistemologies are also valued alongside Western scientific epistemologies and methods (Cameron et al., 2010). Indigenous communities thus want to see evaluation rooted in their worldview (Belaid et al. 2022).

Several frameworks and guidelines address how to develop a more fair evaluation when involving indigenous and minority groups (Wilson and Neville 2009; Cameron et al. 2010; Gollan and Stacey 2021). Here we present the framework developed by Andersson and colleagues to clarify how cultural safety unfolds at different evaluation stages (Cameron et al. 2010; Andersson 2018).

Formulation of the purpose of the evaluation

Ideally, a culturally safe evaluation emanates from a community request. This increases the relevance of the evaluation, ensuring alignment with community priorities. In practice, many evaluations are commissioned from a problem defined by donors, leaving less room to formulate or rename the purpose of the evaluation.

Culturally safe evaluation chooses to build on the strengths of communities. Dominant social groups often label indigenous and minority groups as 'at risk', 'vulnerable' or 'marginalized'. These negative labels do not necessarily reflect how these communities see themselves and significantly reduce the space and conditions for improvement (Wilson and Neville 2009). When stakeholders define or rename the object of evaluation, highlighting their strengths in addressing it, this reduces the negative labelling and opens the way for improvement.

Fuzzy cognitive mapping (FCM) allows groups or individuals to frame the evaluation problem in terms of their own understanding ot it (Andersson and Silver 2019). In a fuzzy cognitive mapping session, participants build a flexible causal model of how they see the problem by providing concepts and lexicon that are familiar to them. They describe the factors that influence the problem and discuss the directions and strengths of each relationship that impacts the problem. This mapping is called "fuzzy" because it assesses the relative influence of each relationship in the map: participants are asked to rate this influence on a scale of 1 (weak) to 5 (strong). This approach enriches the evaluation but also changes its appropriation.

FCM supports cultural safety in several ways. Requiring no literacy or language skills, it promotes equity and inclusion (Andersson and Silver 2019). It reduces the stigma or shame that individuals or groups who belong to previously excluded groups may feel. In FCM sessions, groups are organized by gender, age and type of stakeholder (e.g. patients, providers, programme managers) to ensure that all voices are represented. Facilitators meet with each group separately, reducing power imbalances between groups. Ideally, facilitators should be of the same gender and age and share the language and socio-cultural realities of the participants to reduce hierarchy with them. Facilitators are trained to reduce bias.

The results of the fuzzy cognitive map can inform the evaluation beyond problem formulation. It allows for stakeholder input into the questionnaire design, combining existing literature with stakeholders' views and understandings of the mechanisms of change (Dion et al. 2019; Dion et al. 2022). Giles and colleagues used the tool to capture how a Mohawk community understands the factors influencing diabetes (Giles et al. 2007). Sarmiento and colleagues used the tool to explore how indigenous communities in Guerrero State perceive factors that influence maternal health to better design interventions (Ivan Sarmiento, Paredes-Solís, et al. 2020; Iván Sarmiento, Zuluaga, et al. 2020).

Ethics

Institutional ethics boards and evaluation ethics committees almost invariably rely on Western epistemologies, expecting all aspects of the evaluation to be clarified before the evaluation begins (Cameron et al. 2010). Yet cultural safety involves participant input into evaluation protocols and tools, which is not usually possible before the evaluation begins. Culturally safe evaluations should also seek the approval of Aboriginal committees whenever possible and respect the ethical guidelines for Aboriginal research. In addition, they should apply the principles of ownership, control, access and possession regarding how "data and information from Indigenous peoples will be collected, protected, used and shared. These principles aim to enhance the information governance of Indigenous peoples" (Nations 1998).

Research design

Cultural safety can be applied to qualitative, quantitative and mixed methods evaluations. For example, Andersson and colleagues used a randomized controlled trial to evaluate local interventions to reduce domestic violence in partnership with 12 Aboriginal women's shelters in Canada (Andersson et al. 2010). The shelter directors requested the randomized controlled trial. They felt the need to show the impact of their programme to apply for more funding (Andersson et al. 2010).

Development of data collection instruments and methods

Evaluation teams are often encouraged to use standardized questionnaires to benefit from their validity and reliability. Many conventional questionnaires focus on risk factors, and deficits rather than the strengths and resilience that characterize many indigenous and minority group worldviews. Cultural safety requires more flexibility, particularly in the variables included in the questionnaires. This means including factors that stakeholders perceive as important, even if they are not part of a standardized questionnaire. Very often, it is possible to define the themes of a questionnaire through a participatory process, for example, the fuzzy cognitive mapping mentioned above.

Data collection in culturally safe evaluations can be quantitative or qualitative. Each method can introduce substantial bias when controlled too closely by the evaluation team. Careful recruitment and training of a local data collection team who share a similar social and cultural context with the community may be a better strategy. Not only does this increase acceptability and response rates, but the evaluation promotes skill development in the community.

Data analysis and interpretation

Whether using qualitative or quantitative methods, the analysis and interpretation of data should reflect the worldviews of indigenous and minority groups. For example, in inductive content analysis, there are participatory categorisation and coding options (Liebenberg, Jamal, and Ikeda 2020). Even when the analysis is computerised, which makes wider participation difficult, frequent checking by community members and the separation of analysis (data analysis) from interpretation (what the results mean) help to support the voice of participants and increase the accuracy and relevance of the analysis.

Participants' voices also play a potential role in formal statistical analysis. Andersson and colleagues use the weights generated by fuzzy cognitive mapping as Bayesian *a priori*, incorporating pre-existing beliefs and knowledge as a prior probability distribution. This allows the integration of the indigenous perspective into statistical analysis.

Communication and knowledge transfer activities

Dissemination and transfer of knowledge are essential phases in evaluation. At these stages, there is a high risk of exploiting participants and using the results to project a situation that the communities could not handle. Instead of separating knowledge transfer activities and treating them as the final product of the evaluation, a culturally safe evaluation integrates them into the evaluation process (Kothari, McCutcheon, and Graham 2017). This approach involves all stakeholders – programme staff, funders, participants and, where possible, policymakers – in the evaluation from the outset. All are thus involved in the design, implementation and interpretation. Andersson and colleagues have developed a protocol called SEPA (socialising evidence for participatory action) that allows these steps to be integrated into the research. The

SEPA protocol involves stakeholders in the production of evidence; it also implies presenting the research data so that they can participate not only in its interpretation but also in developing solutions in dialogues. Solutions are thus contextualised, implemented and evaluated (Ledogar et al. 2017).

II. How is this approach useful for policy evaluation?

Cultural safety is the first question that should be asked when evaluating a given programme. In addition to answering *ex-ante* and *ex-post* evaluation questions, the cultural safety approach helps to amplify the voices of participants and beneficiaries. It should be a requirement for all evaluations of public services. For example, in programmes aimed at addressing the needs of Aboriginal or minority people, involving them in the evaluation, hearing how they define the problem (what their needs are and how they perceive the relevance of the evaluation) and understanding how they perceive the cultural appropriateness of potential solutions can only strengthen the policy and make it more appropriate (Cram 2016; Cameron et al. 2010).

Cultural safety helps to avoid implementation barriers, increasing the programme's effectiveness by including indigenous worldviews, needs and priorities (effectiveness evaluation). It can help to reduce inequalities (short-term impact). It can empower these communities and eventually lead to their self-determination (long-term impact).

III. An example of the use of this approach in reproductive health

A reproductive health project in a post-conflict region of northern Uganda used the cultural safety framework to improve reproductive health outcomes. The project received funding from a Canadian organisation (Belaid et al. 2020; Belaid et al. 2021).

The civil war (1986-2006) between the Lord's Resistance Army (North) and the Museveni government (South) displaced more than 90% of the population in this region. This has increased long-standing tensions between northern and southern Uganda (Laruni 2015).

Northern communities are still recovering from the conflict. The region has low poverty, social opportunity and human development indicators (Esuruku 2019). The conflict has had a negative impact on health services, deteriorating maternal health (Chi et al. 2015b, 2015a). Women and girls in this region are less educated and poorer. They are much less likely to give birth in a health facility (Uganda Bureau of Statistics (UBOS) 2012).

Before launching the project, we invested time in building relationships with local stakeholders and developing networks to involve community members. Stakeholders included women and men of different ages, traditional midwives, service providers and government officials. We involved these groups in all activities of the programme design. Each group defined perinatal care outcomes according to their worldview. We used fuzzy cognitive mapping to collect and compare perspectives. Group discussions clarified the lexicon and cultural concepts associated with perinatal care. We used these concepts to design the questionnaire, as far as possible, also using standardised questions corresponding to the concepts identified by the stakeholders.

The groups met in a series of deliberative dialogues to discuss local evidence, generate lists of potential strategies for improving access to perinatal care and design a programme. We invited the groups to discuss who should deliver the programme, how and with what content. Participants identified several barriers to accessing perinatal care and proposed strategies to address the problems in a culturally safe way.

Cultural safety helped identify problems in perinatal service provision. Reflection on local evidence generated feasible community-led solutions. This, in turn, increased trust between community members and service providers.

IV. What are the criteria for judging the quality of the mobilisation of this approach?

Cultural safety can only be judged by programme beneficiaries (Wilson and Neville 2009; Cameron et al. 2010; CIET/PRAM 2022). However, evaluation teams can reflect on the following questions:

- Do participants/beneficiaries report feeling culturally safe during the evaluation?
- How will the intended beneficiaries be involved in each evaluation phase?
- Do the terms of the evaluation lend themselves to partnership?
- Does the evaluation build on the strengths of communities?
- Does the evaluation increase ownership of the project or service and the evaluation products?
- How are the methods adapted to the specific culture?

- What is the anticipated impact of the evaluation on community self-determination?

V. What are the strengths and limitations of this approach compared to others?

A cultural safety framework has several advantages for evaluation teams and programme participants. It increases the local acceptability and relevance of the evaluation. It can guide the design of programmes and services, increasing their contextual appropriateness. In evaluating ongoing programmes, cultural safety is an interpretive lens for understanding how indigenous and minority communities experience these programmes and services. As an analytical lens, it can highlight how inequalities and social injustices are shaped, what changes are needed, and what barriers or facilitators to these changes are possible (Gerlach 2012).

The main challenge is to develop and agree on protocols for assessing the impact of cultural safety in the context of complex outcomes (Gerlach 2012; Tremblay et al. 2020). Cultural safety depends largely on each local context, as each cultural group is different and has its own way of seeing things and its own degree of adaptation to dominant representations (Cameron et al., 2010). However, as more and more evaluations apply a cultural safety framework, our experiences of best practices will accumulate and contribute to developing guidelines with a wide range of transferability.

Some bibliographical references to go further

Andersson, Neil. and Shea, Beverley. and Amaratunga, Carol. and McGuire, Patricia. and Sioui, Georges. 2010. “Rebuilding from Resilience: Research Framework for a Randomized Controlled Trial of Community-led Interventions to Prevent Domestic Violence in Aboriginal Communities.” *Pimatisiwin*, 8(2): 61-88.

Andersson, Neil. and Silver, Hilah. 2019. “Fuzzy cognitive mapping: An old tool with new uses in nursing research.” *Journal of Advanced Nursing*, 75, no.12: 3823-30.

Belaid, Loubna. and Atim, Pamela. and Atim, Eunice. and Ochola, Emmanuel. and Bayo, Pontius. and Oola, Janet. and Sarmiento, Ivan. and Zarowsky, Christina. and Andersson, Neil. 2020. “Marginalized women and services providers improve access to perinatal care in post-conflict Northern Uganda: socializing evidence for participatory action”.

Cameron, Mary. and Andersson, Neil. and McDowell, Ian. and Ledogar, Robert. 2010. “Culturally Safe Epidemiology: Oxymoron or Scientific Imperative.” *Pimatisiwin*, 8(2): 89-116.

Curtis, Elana. and Jones, Rhys. and Tipene-Leach, David. and Walker, Curtis. and Loring, Belinda. and Paine, Sarah-Jane. and Reid, Papaarangi. 2019. “Why cultural safety rather than cultural competency is required to achieve health equity: a literature review and recommended definition.” *International Journal for Equity in Health*, 18(1): 174. https://doi.org/10.1186/s12939-019-1082-3. https://doi.org/10.1186/s12939-019-1082-3.

Dion, Anna. and Carini-Gutierrez, Alessandro. and Jimenez, Vania. and Ben Ameur, Amal. and Robert, Emilie. and Joseph, Lawrence. and Andersson, Neil. 2022. “Weight of Evidence: Participatory Methods and

Bayesian Updating to Contextualize Evidence Synthesis in Stakeholders' Knowledge." *J Mix Methods Res*, 16(3): 281-306. https://doi.org/10.1177/15586898211037412.

Gerlach, Alison J. 2012. "A Critical Reflection on the Concept of Cultural Safety." *Canadian Journal of Occupational Therapy*, 79(3): 151-58. https://doi.org/10.2182/cjot.2012.79.3.4. https://journals.sagepub.com/doi/abs/10.2182/cjot.2012.79.3.4.

Giles, Brian. and Findlay, C. Scott. and Haas, George. and LaFrance, Brenda. and Laughing, Wesley. and Pembleton, Sakakohe. 2007. "Integrating conventional science and aboriginal perspectives on diabetes using fuzzy cognitive maps." *Social Science & Medicine*, 64(3): 562-76. https://doi.org/https://doi.org/10.1016/j.socscimed.2006.09.007. http://www.sciencedirect.com/science/article/pii/S0277953606004758.

Gollan, Sharon. and Stacey, Kathleen. 2021. *Australian Evaluation Society First Nations Cultural Safety Framework*. Australian Evaluation Society (Melbourne, Australia).

Gomez Cardona, Liliana. and Brown, Kristyn. and McComber, Mary. and Outerbridge, Joy. and Parent-Racine, Echo. and Phillips, Allyson. and Boyer, Cyndy. and Martin, Codey. and Splicer, Brooke. and Thompson, Darrell. and Yang, Michelle. and Velupillai, Gajanan. and Laliberté, Arlène. and Haswell, Melissa. and Linnaranta, Outi. 2021. "Depression or resilience? A participatory study to identify an appropriate assessment tool with Kanien'kéha (Mohawk) and Inuit in Quebec." *Social Psychiatry and Psychiatric Epidemiology*, 56(10): 1891-902. https://doi.org/10.1007/s00127-021-02057-1. https://doi.org/10.1007/s00127-021-02057-1.

Ledogar, Robert. and Arosteguí, Jorge. and Hernández-Alvarez, Carlos. and Morales-Perez, Arcadio. and Nava-Aguilera, Elizabeth. and Legorreta-Soberanis, José. and Suazo-Laguna, Harold. and Belli, Alejandro. and Laucirica, Jorge. and Coloma, Josefina. and Harris, Eva.

and Andersson, Neil. 2017. “Mobilising communities for Aedes aegypti control: the SEPA approach.” *BMC Public Health*, 17(1): 403. https://doi.org/10.1186/s12889-017-4298-4. https://doi.org/10.1186/s12889-017-4298-4.

Nations, Le Centre de Gouvernance de l’Information des Premières. 1998. “Les principes de PCAP® des Premières Nations.” Le Centre de Gouvernance de l’Information des Premières Nations.

Papps, Elaine. and Ramsden, Irihapeti. 1996. “Cultural Safety in Nursing: the New Zealand Experience.” *International journal for quality in health care: journal of the International Society for Quality in Health Care / ISQua*, 8: 491-7. https://doi.org/10.1093/intqhc/8.5.491.

Sarmiento, Ivan. and Paredes-Solís, Sergio. and Loutfi, David. and Dion, Anna. and Cockcroft, Anne. and Andersson, Neil. 2020. “Fuzzy cognitive mapping and soft models of indigenous knowledge on maternal health in Guerrero, Mexico.” *BMC Medical Research Methodology*, 20(1): 125. https://doi.org/10.1186/s12874-020-00998-w. https://doi.org/10.1186/s12874-020-00998-w.

Smith, Linda Tuhiwai. 2012. *Decolonising methodologies. Research and Indigenous Peoples*. New York, USA: Bloomsbury.

Smye, Vicky. and Browne, Annette. 2002. “‘Cultural safety’ and the analysis of health policy affecting aboriginal people.” *Nurse Res*, 9(3): 42-56. https://doi.org/10.7748/nr2002.04.9.3.42.c6188.

Williams, Robyn. 1999. “Cultural safety — what does it mean for our work practice?” *Australian and New Zealand Journal of Public Health*, 23(2): 213-14. https://doi.org/https://doi.org/10.1111/j.1467-842X.1999.tb01240.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-842X.1999.tb01240.x.

Wilson, Denise. and Neville, Stephen. 2009. “Culturally safe research with vulnerable populations.” *Contemporary Nurse*, 33(1): 69-79. https://doi.org/10.5172/conu.33.1.69. https://doi.org/10.5172/conu.33.1.69.

About the contributors

Mathias André is a graduate of the École nationale de la statistique et de l'administration économique (Ensae), has a doctorate in economics from the École Polytechnique and is an administrator at INSEE. He is an expert in charge of the elaboration of the national distributed accounts within INSEE, he was the rapporteur of the expert group on the measurement of redistribution and inequalities published in February 2021 and author of a report on redistributive issues for the Conseil des prélèvements obligatoires. He has published work on the microsimulation of socio-fiscal reforms and the role of various transfers in reducing inequalities (VAT, income tax, property tax). He also conducts studies on property wealth or inequalities and climate. He teaches microsimulation at Ensae.

mathias.andre@insee.fr

Neil Andersson MD PhD is Professor of Family Medicine and Director of Participatory Research at McGill University in Montreal, Canada. He specialises in large-scale community trials with an emphasis on patient and community engagement. Interventions in these trials have included prevention of vector-borne diseases, immunisation, cultural safety in childbirth, perinatal home visits, prevention of gender-based violence and access to social support for young women at risk of HIV. His contributions to trial methodology have emphasised early stakeholder engagement in the conceptualisation and design of interventions, and cluster analysis of outcomes. He advocates for the ethical and culturally safe inclusion of socially marginalised and indigenous groups in randomised controlled trials.

neil.andersson@mcgill.ca

Habibata Baldé is a social scientist with a specialisation in Applied Social Sciences in Health and expertise in qualitative research. Since 2014, she has been actively involved in health research activities. She is currently a

Technical Assistant at the African Centre of Excellence for the Prevention and Control of Communicable Diseases at the University of Conakry and a consultant for Guinea in a multi-country research project on the introduction of the Pulse Oximeter in primary health centres with the NGO ALIMA.

habicompanya@gmail.com

Carlo Barone is full professor of Sociology at the CRIS, Sciences Po, Paris. He is the director of the research group on educational policies of LIEPP, and he is affiliated to J-PAL. He conducts several policy impact evaluations in the field of education using randomized controlled trials as well as systematic reviews and meta-analyses. He has published several articles on educational inequalities, labor market inequalities and social mobility in leading sociology journals.

carlo.barone@sciencespo.fr

Loubna Belaid is an assistant professor of program evaluation at the École Nationale d'Administration Publique in Montreal, Canada. Her research interests focus on cultural safety and social justice in health programs and policies. She uses participatory approaches to co-design, implement and evaluate health programs with and for Aboriginal, racialized and marginalized communities in Canada and East Africa.

Loubna.Belaid@enap.ca

Genowefa Blundo-Canto is a social scientist at the French Agricultural Research Center for International Development (Cirad). Previously she worked as consultant and research fellow at Bioversity International and as a post-doc in impact assessment at the International Centre for Tropical Agriculture (CIAT). A development economist, she carries out research in impact evaluation of agricultural research for development (AR4D) interventions. She is particularly interested in methodological issues in impact assessment, focusing on integrated mixed methods, participatory and systemic approaches, and navigating complexity at multiple scales. She also combines participatory foresight with qualitative and quantitative impact assessment methods to broaden the scope of evaluations. Thematically, she focuses on the multidimensional and multi-scale impacts of interventions that enhance the use of agricultural biodiversity, with a special interest in food and nutrition security, social equity and power imbalances at multiple scales.

genowefa.blundo_canto@cirad.fr

Abdourahmane Coulibaly is an anthropologist, teacher-researcher at the Faculty of Medicine and Odontostomatology (Mali) and member of the IRL "Environment – Health – Societies" UCAD, USTTB, CNRST – Ouagadougou, CNRS – France. He has participated in several research programmes on the implementation of health policies, including results-based financing, technological innovations in the field of health and analysis of the resilience of health institutions. His research is largely based on the use of qualitative methods and the ethnographic approach.

coulibalyabdourahmane@gmail.com

Thomas Delahais is an evaluator and co-founder of the cooperative company Quadrant Conseil. His work focuses on the evaluation of complex interventions and in particular on contribution analysis, the evaluation of transition initiatives and the sociology of evaluation. He is a member of the editorial board of the *Evaluation Journal*.

tdelahais@quadrant-conseil.fr

Agathe Devaux-Spatarakis is a consultant and researcher for the Scop Quadrant Conseil. She conducts policy evaluation and methodological support missions for public organisations and NGOs in France and abroad. She has a PhD in political science and specialises in the development of evaluation methods adapted to social innovations and experiments, as well as the study of the use of evaluation results by policy makers.

adevaux@quadrant-conseil.fr

Emanuele Ferragina is Professor of Sociology at Sciences Po, Paris; prior to his appointment in Paris he worked as a researcher and Departmental Lecturer at the University of Oxford. Emanuele studied in Turin, Bordeaux, London and Paris and received his DPhil in Social Policy from the University of Oxford. His work has been published in several journals in political economy, sociology, political science and social policy, such as the *Review of International Political Economy*, *New Political Economy*, *Research in Social Stratification & Mobility*, *Socius*, *International Journal of Comparative Sociology*, *Political Studies Review*, *Journal of European Social Policy*, *Social Policy & Administration*, *Social Politics*, *Social Policy & Society*, *Stato & Mercato*, *L'Année Sociologique*.

emanuele.ferragina@sciencespo.fr

Nicolas Fischer is a CNRS research fellow in political science at the Centre de recherches sociologiques sur le droit et les institutions pénales (CESDIP). His recent research has focused on the administrative detention of foreigners in France and immigration policies, on the independent control of detention facilities, and more broadly on the tension between violent repression and legal protection of stigmatised populations in democracies. He is currently investigating the medico-judicial controversies surrounding capital punishment by lethal injection in the United States. His publications include *Le territoire de l'expulsion. La rétention administrative des étrangers et l'Etat de droit en France* (Lyon: ENS Editions, 2017) and, with Camille Hamidi, *Les politiques migratoires* (Paris: la Découverte coll. Repères, 2016).

fischer@cesdip.fr

Denis Fougère is Director of Research at the CNRS. He is a member of the Centre for Research on Social Inequalities (CRIS) and the Laboratory for interdisciplinary evaluation of public policies (LIEPP) at Sciences Po Paris. He teaches economics of education and statistical methods of public policy evaluation at Sciences Po Paris. He is also a Research Fellow at the

Centre for Economic Policy Research (CEPR, London) and at the Institute of Labor Economics (IZA, Bonn). His current research focuses on the evaluation of education policies and pension reforms in France. He has published articles in several international journals, such as *Econometrica*, *Review of Economic Studies*, *Review of Economics and Statistics*, *Economic Journal*, *European Economic Review*, *European Sociological Review*, *Journal of Public Economics*, *Journal of the European Economic Association*, *Journal of Business and Economic Statistics*, *Journal of Applied Econometrics*, *The Econometrics Journal*, *Journal of Population Economics*, *Labour Economics*, etc.

denis.fougere@sciencespo.fr

Lara Gautier is an assistant professor at the School of Public Health of the University of Montreal and a regular researcher at the Centre de recherche en santé publique, and at the SHERPA Research Institute, in Montreal, Canada. Trained in public health and political science, she has scientific expertise in participatory evaluation of health services using qualitative and mixed research methods.

lara.gautier@umontreal.ca

Pauline Givord has been Head of the Economic Studies Department of the French National Institute for Statistics and Economic Studies (INSEE) since December 2022. Previously, she held several positions within INSEE, focusing on economic studies on a wide range of subjects and statistical methodology. An expert in econometric methods for evaluating public policies, she also participated in the creation of the SSP-Lab at INSEE, which aims to promote innovation in data sources and data science methods relating to the statistical output of the official statistical system (SSP). She has also worked at the Centre for Economic and Statistical Research (Crest), the OECD, and the Directorate for Research Coordination (Dares) at the Ministry of Labour, where she was in charge of monitoring the evaluation of the Investment Plan in Skills. She is a graduate of the Ecole Polytechnique and ENSAE, holds a PhD in economics and a habilitation to supervise research.

pauline.givord@insee.fr

Charlotte Halpern holds a PhD in political science and is Tenured Researcher in Political Science at the Centre for European Studies (CEE) of Sciences Po and codirector of the Environmental policy research group at the Laboratory for interdisciplinary evaluation of public policies (LIEPP). Her published works examine processes of policy change and the relationship between social mobilisations and dynamics of state restructuring. Her current research focuses on the governing of sustainable transition policies in European cities. Recent publications include special issues and articles in journals (e.g., *Comparative European Politics*; *West European Politics*; *Politique européenne* …) and two edited volumes, *Policy analysis in France* (Policy press, 2018 co-ed. with P. Hassenteufel and P. Zittoun) and *Villes sobres* (Presses de Sciences Po, 2018). She is the scientific director of the Sciences Po Executive master programme on Territorial Governance and Urban Planning, and teaches comparative public policy analysis, urban governance and environmental policies at Sciences Po and AgroParisTech.

charlotte.halpern@sciencespo.fr

Quan Nha Hong is an assistant professor at the School of Rehabilitation of the Université de Montréal and a researcher at the Centre de recherche interdisciplinaire en réadaptation du Montréal métropolitain (CRIR) – Institut universitaire sur la réadaptation en déficience physique de Montréal (IURDPM). She is an occupational therapist with research training in clinical sciences (M.Sc., Université de Sherbrooke), in health technology assessment (M.Sc., Université de Montréal), and in primary care (Ph.D., McGill University). She also completed a postdoctoral fellowship at the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) at University College London (UCL) and at the Institut national d'excellence en santé et en services sociaux (INESSS). She is particularly interested in research methods, including systematic reviews and mixed methods, and knowledge transfer to support evidence-informed decision-making.

quan.nha.hong@umontreal.ca

Nicolas Jacquemet is Professor of Economics at the University of Paris 1 Panthéon-Sorbonne and at the Paris School of Economics, where he is in charge of the Master in Economics and Psychology. He co-authored various textbooks targeting either or both master students and/or professionals of policy evaluation on econometrics (De Boeck) and on experimental methods in economics (Cambridge University Press, Economica). His work in the field of public policy evaluation combines various experimental methods (testing, field experiments, controlled and laboratory experiments), as well as survey or administrative data, and focuses on e.g., discrimination in hiring, tax evasion, the management of health care supply or ressource sharing within households. He regularly contributes to the public debate, and has collaborated on a book for the general public presenting the contributions of behavioural economics to the design of public policies (La découverte, Repères collection).

nicolas.jacquemet@univ-paris1.fr

Sarah Louart is a doctoral student in socio-economics of health, attached to the Lille Centre for Sociological and Economic Studies and Research (Clersé, University of Lille). She studies access to health care for poor populations and the processes of introduction and diffusion of health innovations. She is currently working with the NGO ALIMA and the Institut de Recherche pour le Développement (IRD) in Dakar on the realist evaluation of a project aimed at disseminating a health innovation in primary health centres in four West African countries.

sarah.louart@gmail.com

Ana Manzano is an associate professor in public policy at the University of Leeds (UK) and a senior applied social scientist expert in evaluation and healthcare policy, systems and practice. Her areas of expertise are realist evaluation, advanced qualitative methods, and the relationship between methods, evidence and programme evaluation. She has worked in international evaluation studies in Europe, Africa and South Asia. Manzano regularly publishes in the main field journals in evaluation, methodology and health policy and systems, including *Evaluation*, *Implementation Science*, *Evaluation and Program Planning*, *the International Journal of Social Research Methodology*, *Social Sciences and Medicine*, and *Health Policy and Planning*. Manzano was part of the RAMESES II study that developed reporting standards for conducting realist evaluations; and she is the Director of Advanced Qualitative Methods at the White Rose Social Sciences Doctoral Training Partnership (UK).

A.Manzano@leeds.ac.uk

Valérie Pattyn is an Associate Professor at the Institute of Public Administration of Leiden University (The Netherlands), and is partially affiliated to KU Leuven Public Governance Institute (Belgium). Her work focuses on public policy, evidence-informed policy making, and policy evaluation. Her current research programme deals with issues such as the development of evaluation systems and the impact on policy decisions; evidence and policy advice production and use within and outside the civil service; and policymaking under conditions of uncertainty. Besides basic research, she has developed a substantial track record in policy consulting and evaluation research in various policy fields. She serves in various national and international networks including the European Group for Public Administration (co-chair Permanent Study Group Policy Design and Evaluation); the Flemish Evaluation Association (co-ordination committee); the Dutch Evaluation Association (core group); and COMPArative Methods for Systematic cross-caSe analysis (advisory board).

v.e.pattyn@fgga.leidenuniv.nl

Clément Pin is an assistant professor in sociology at INSHEA, member of GRHAPES. He is a researcher affiliated with the Laboratory for interdisciplinary evaluation of public policies (LIEPP) and the EMA laboratory (Cergy Paris University). His research focuses on: 1) educational and university policies, 2) school and career guidance systems and instruments, 3) local mediation (school-family relations and more broadly territorial governance), 4) evaluation methods (qualitative and mixed). He has recently published with Carlo Barone "L'apport des méthodes mixtes à l'évaluation" (*Revue française de science politique*, 2021/3) and with Agnès van Zanten "The Impact on French Upper Secondary Schools of Reforms Aiming to Improve Students' Transition to Higher Education" (*Oxford Research Encyclopedia of Education*, OUP, 2021).

clement.pin@wanadoo.fr

Pierre Pluye (†) was a professor in the Department of Family Medicine and an associate member of the School of Information Sciences at McGill University. He co-directed the association "Méthodes mixtes francophonie" (MMF) and the "Valorisation des données" axis of the Unité Système de Santé Apprenant Québec. He was a member of the Canadian Academy of Health Sciences, and a founding member of the Quebec Network of Research Focused on Frontline Practices, among others. In 2017, he received the "Researcher of the Year" award from the College of Family Physicians of Canada. In 2021, he received the Doctoral Teaching Award from the Association of Northeastern Universities (Canada/USA) which recognizes excellence and innovation in doctoral teaching. Through his expertise in mixed methods and mixed reviews, he tremedously contributed to the development of these methods. His latest research focused on evaluating and improving the effects of online health information. The carreer of Professor Pierre Pluye, who passed away on August, 1st, 2023, is retraced here by the directing committee of

Méthodes mixtes francophonie : http://methodesmixtesfrancophonie.pbworks.com/w/file/fetch/154045617/PP_2023-08-01.pdf

Estelle Raimondo is a Methods Advisor at the Independent Evaluation Group (IEG) of the World Bank, where she leads evaluations and advises IEG teams on methodological design and innovation. With over ten years of experience in development policy evaluation, she is a Faculty Member of the International Program for Development Evaluation Training (IPDET) and sits on the Board of the European Evaluation Society. Her research has been published in several international peer-reviewed journals and books. She received her PhD in evaluation from George Washington University and a double master's degree in international economic policy from Sciences Po Paris and Columbia University.

eraimondo@worldbank.org

Thomas Rapp is an associate professor in Economics at LIRAES (Université Paris Cité), co-director of LIEPP's "Health Policies" research group, and holder of the AgingUP! Chair. He is specialised in health economics, economics of ageing and health policy analysis. He is the author of more than 70 publications (articles, reports, book chapters) on these themes. He was a Harkness fellow at Harvard (2015-2016), a health economist at the OECD (2017-2019), and a visiting professor at Harvard, Columbia and the Catholic University of Rome. He is associate editor of the scientific journal *Value in Health*. For the past 10 years, his research programme has been funded by several grants, notably from the French National Research Agency, the CommonWealth Fund in New York, and the Innovative Medicines Initiative of the European Commission.

thomas.rapp@u-paris.fr

Anne Revillard is an Associate professor of sociology at Sciences Po, member of the Centre for research on social inequalities (CRIS) and director of the Laboratory for interdisciplinary evaluation of public policies (LIEPP). Her research explores the interplay between law, public policy, and the transformations of systems of inequality linked to gender and disability. She has notably contributed to reflections on policy evaluation based on a focus on qualitative methods, and through an approach in terms of policy reception, centred on the viewpoints of people who are the targets of public policies.

anne.revillard@sciencespo.fr

Valéry Ridde is Director of Research at CEPED, a joint research unit of the Université Paris Cité and the Institut de recherche pour le développement (IRD). He is currently based at the Institute of Health and Development (ISED) of the Cheikh Anta Diop University in Dakar (Senegal). His research and evaluation work focuses on universal health coverage, health service financing, programme evaluation, public health policy and health promotion.

valery.ridde@ird.fr

Émilie Robert is an associate professor at the School of Public Health of the Université de Montréal. A specialist in the realist approach to evaluation and knowledge synthesis, she trains research teams and assists them in the design and implementation of their evaluative research projects. Émilie has carried out mandates for several international and provincial, public and non-governmental, and academic organisations. Her approach is rooted in knowledge-based evaluation and the development of evaluative thinking.

emilierobert.udem@gmail.com

Lou Safra holds a Phd in cognitive science from the Ecole Normale Supérieure (Paris). She has been an assistant professor of political psychology at Sciences Po since 2018. She is a member of CEVIPOF, and also associated to the Institut d'Études Cognitives (Laboratoire de Neurosciences Cognitives & Laboratoire de Neurosciences Cognitives et Computationnelles, École Normale Supérieure, Paris) and affiliated to the Laboratory for interdisciplinary evaluation of public policies (LIEPP). In her research, she applies cognitive science concepts and methods to the study of political and social behaviour. She is particularly interested in the causes and consequences of social, political and economic inequalities, analysing both the reactions to these inequalities, the behavioural effects of these inequalities on disadvantaged populations, and the way in which public policies can contribute to accentuating or limiting these inequalities. To this end, it combines the use of laboratory experiments with the analysis of international survey data and cultural objects.

lou.safra@sciencespo.fr

Acknowledgements

This book is the result of the scientific momentum generated since 2011 by the Laboratory for interdisciplinary evaluation of public policies (LIEPP), which created a space for dialogue between different methodological traditions and evaluative approaches. This book, which is strongly supported by the community of researchers involved in LIEPP's activities, aims to facilitate and nurture this exchange. I would therefore like to begin by thanking the 25 authors who, either already involved in the LIEPP project or called upon for the occasion, have given their time to prepare the didactic chapters that make up this book, following a common framework of questioning and strict requirements in terms of editorial format, giving rise to numerous rewrites. Launched in the summer of 2022, this publication project was completed in less than a year thanks to the enthusiastic commitment and responsiveness of contributors committed to communicating their passion for research in accessible terms, in an approach based on scientific mediation and inter-method dialogue.

The LIEPP also provided financial support for this editorial project, enabling it to be published in open access. As such, this book has benefited from the support provided by the ANR and the French government under the "Investissements d'Avenir" program LABEX LIEPP (ANR-11-LABX-0091, ANR-11-IDEX-0005-02) and the IdEx Université Paris Cité (ANR-18-IDEX-0001).

At LIEPP, the texts were translated and initially formatted by Konstantinos Papadopoulos, who was a key player in this project and whom I would like to thank for his involvement. Ariane Lacaze then did a great deal of editorial work and monitored the proofreading by the authors: many thanks to her for her professionalism and efficiency.

I would also like to thank Samira Jebli, Andreana Khristova and Latifa Lousao, from the LIEPP team, who contributed to this project on an *ad hoc* basis.

My sincere thanks go to Editions Science et Bien Commun, and especially Erika Nimis, for hosting this dual publication (in French and English), and for their responsiveness and the care taken in preparing the book.

About the editor

Les Éditions science et bien commun is a project of the Association science et bien commun (Science and the Common Good Association) (ASBC), a nonprofit organization registered in Quebec in July 2011.

The ASBC

The mission of the Association is to support and disseminate trans-university research that promotes the development of science that is pluralistic, open, fair, multilingual, non-sexist, non-racist, socially responsible, in order to promote the common good.

For more information, write to info@scienceetbiencommun.org, or subscribe to the Twitter account @ScienceBienComm or the Facebook page https://www.facebook.com/scienceetbiencommun.

Les Éditions science et bien commun

An innovative editorial project with the following values:

- open access digital publishing, in addition to other formats
- multidisciplinarity, as much as possible
- publication in multiple languages, including African national languages or Creole, in addition to French
- internationalization, to bring together authors from different countries and to write for an audience from different countries and cultures

- Cognitive justice is our chief aim:
- each collective book, even if it is the proceedings of a conference, should aspire to parity between women and men, juniors and seniors, authors from the North and from the South; in any case, all books should avoid imbalance between these points of view;
- each book, even if written by only one person, should endeavour to include references to both the countries of the North and the South, in its themes or in its bibliography;
- each book should aim for accessibility and "readability", limiting jargon, even if it is scientifically oriented and peer-reviewed.

The catalogue

The catalogue of Éditions science et bien commun (ESBC) is composed of books that respect the values and principles of ESBC set out above.

- Scientific works (collectively written or monographs) which may be original unpublished manuscripts, or from theses, dissertations, colloquia, seminars or research projects, digital republications or university textbooks. Unpublished manuscripts will be peer-reviewed in an open manner, unless the authors do not wish to do so (see the evaluation point above).
- Works of citizen or participatory science, popular science or local and traditional knowledge, whose aim is to make this knowledge available to a wider readership.
- Essays on science and science policy (e.g. in social studies of science or science ethics).
- Anthologies of texts already published but not accessible on the web, in a language other than French, or which are not open access but of demonstrated scientific, intellectual or heritage interest.
- Textbooks or educational books for children

For open and universal access, through digital technology, to scientific books published by authors from countries in the Global South and North.

For more information: write to info@editionscienceetbiencommun.org