Chapter 10

# A NEW METHOD FOR POLICY EVALUATION?

*Longstanding Challenges and the Possibilities of Qualitative Comparative Analysis (QCA)*

Frédéric Varone
*Université catholique de Louvain*

Benoît Rihoux
*Université catholique de Louvain*

Axel Marx
*Hogeschool Antwerpen*

## 1. INTRODUCTION

According to E. Vedung (1997: 3), evaluation is "*the careful retrospective assessment of the merit, worth and value of administration, outputs, and outcome of government interventions, which is intended to play a role in future, practical action situations*". Thus, the purpose of evaluation research is to measure the effects of a policy against the goals it sets out to accomplish. Hence, it implies the application of systematic research methods for the assessment of program design, implementation and effectiveness.

In fact, several handbooks are dedicated to evaluation designs, which constitute the technical part of an evaluation process and relate to the collection and interpretation of empirical data on policy outputs and outcomes (see analytical distinction below). Various methodological designs are possible, such as longitudinal and/or cross-sectional quantitative research, different types of experimental designs (including quasi-experiments and natural experiments), or different types of (comparative) case research. With each design also come specific types of data (from the most quantitative to the most qualitative) and data analysis techniques. Although there are many ways in which outputs and outcomes assessment can be conducted, these methodological options are not all equivalent: some produce more credible estimates of policy effects than others. Therefore, it is not surprising that there

is still a deep divide and fierce academic struggle among the advocates of quantitative versus qualitative methods of policy evaluation.

In this exploratory chapter, we investigate the potential added-value of QCA in such a methodological debate. Indeed, up till now, QCA and policy evaluation have very seldom been explicitly linked[1], and never in a systematic way. This is quite surprising, as clear parallels can be drawn between some key features of QCA (and, hence, the preoccupations of its initiators) and key preoccupations of policy evaluators.

The following two sections set the stage of "policy evaluation" and "QCA": we first provide a brief definition of policy evaluation, and then lay out the fundamentals of QCA. Next, we identify four methodological challenges that policy evaluators typically face. From there on, we examine to what extend QCA may offer some innovative answers to these longstanding methodological issues. Finally, we identify some of the remaining challenges for QCA, if it is to become a very useful tool for policy evaluation.

## 2. DEFINING POLICY EVALUATION: MEASUREMENT AND VALUE JUDGMENT

We define a public policy as a body of decisions and activities adopted and carried out by interdependent public and private actors – with varying values, beliefs, interests, institutional allegiances and resources – in order to resolve, in a coordinated and targeted manner, a collective problem that has been socially constructed and politically defined as public in nature (Knoepfel, Larrue and Varone 2001). Each public policy is thus based on some "causal theory" that consists of assumptions about the causes of the problem to be solved and about the intended impacts (on actors' behavior) of the implemented policy tools.

Numerous authors have tried to create a diagram conveying the unfolding of the decision and implementation processes involved in a public policy. The overall impression that emerges from the literature is one of a "policy life cycle" starting with the emergence and perception of a public problem, followed by the agenda-setting stage, the policy formulation, the implementation phase and, finally, the evaluation of the policy effects (Jones 1970). At this last stage of a policy cycle[2], the policy analyst aims to determine the results and effects of a public policy in terms of the production of administrative acts (policy *outputs*), the changes in behavior of target groups, and problem resolution (policy *outcomes*). Thus, policy evaluation represents an empirical test of the validity of the "causal theory" underlying the public policy.

Consequently the emphasis is placed on links between the administrative

services responsible for implementing the public policy, the target groups whose behavior is politically defined as one of the (in)direct causes of the societal problem to be solved, and the final beneficiaries who endure the negative consequences of this public problem. For example, public environmental protection agencies, as well as industry, impose decontamination measures on polluting industrial companies in order to improve the quality of the air breathed by people living in the vicinity of factories, whilst the economic promotion and finance agencies grant tax exemptions to small and medium-sized companies which employ job-seekers by creating new jobs.

The main tasks of a policy evaluation are, on the one hand, to measure the outputs and outcomes of the public policy and, on the other hand, to formulate a judgment on the value, merit or worth of these policy effects with reference to criteria and explicit standards. As it were, there are numerous different such criteria. For example, the criterion of *relevance* (or appropriateness) examines the link that exists – or should exist – between the goals as defined in the policy design, on the one hand, and the nature and pressure of the public problem to be solved, on the other hand. Thus, a policy is described as relevant if the goals implicitly formulated in the laws and regulations, and sometimes concretized in administrative action plans (i.e. policy *outputs*), are adapted to the nature and temporal and socio-spatial distribution of the problem that the policy is intended to solve. The criterion of *effectiveness* is directly connected with the category of policy *outcomes*. It refers to the relationship between the anticipated effects of a policy and those that actually emerge in social reality. The evaluation of the effectiveness of a policy is generally carried out on the basis of a comparison between the target values (i.e. goals) defined in the policy design and the effects actually triggered among the policy's end beneficiaries. The criterion of *efficiency* focuses on the relationship between the resources invested during policy implementation and the effects achieved. It describes the ratio between the costs and benefits of a policy. The criterion of *economy*, which is rooted in a more managerial rationale, relates the administrative outputs produced to the resources invested. Thus, it evaluates the efficiency (in a narrow sense) of the administrative implementation processes. Further evaluation criteria are also discussed in the literature and applied in concrete evaluations. For example, E. Ostrom (1999: 48-49) refers to policy evaluation in terms of six criteria namely: economic efficiency, fiscal equivalence, redistributional equity, accountability, conformance to general morality and adaptability.

In this chapter, we mainly focus on the first ambition of a policy evaluation (the production of a valid and reliable measure of policy effects) and do not consider explicitly the aspect of value judgment (which always requires a previous measurement of the policy effects). The four methodological

challenges of policy evaluation that are discussed in the third section are all related to the measurement of policy effects. Before addressing these challenges, we briefly introduce some fundamental notions of Qualitative Comparative Analysis (QCA).

## 3. QCA IN A NUTSHELL

Qualitative Comparative Analysis (QCA) is a method that was launched some 15 years ago by Charles Ragin in a prize-winning volume (Ragin 1987; Ragin and Rihoux 2004; Rihoux 2003). It is both an approach (and research design) and a specific technique for the analysis of data.

### 3.1 QCA as an Approach

As an approach, QCA develops a "synthetic strategy", which ambitions to «*integrate the best features of the case-oriented approach with the best features of the variable-oriented approach*» (Ragin 1987: 84).

Indeed, on the one hand, QCA meets some key strengths of the qualitative approach (Ragin 1987: 12ff; De Meur and Rihoux 2002: 20ff). The first one is its holistic character: each individual case is considered as a complex entity (a «whole») which needs to be comprehended and which should not be forgotten in the course of the analysis. Thus it is a case-sensitive approach. Furthermore, it develops a conception of causality which leaves some room for complexity. This is a truly central feature of QCA: multiple conjunctural causation. This implies that : A/ most often, it is a combination of conditions[3] that eventually produce a phenomenon (the « outcome »[4]); B/ several different combinations of conditions may very well produce the same outcome; and C/ depending on the context, on the « conjuncture », a given condition may very well have a different impact on the outcome. This implies that different « causal paths » – each path being relevant, in a distinct way – may lead to the same outcome (De Meur and Rihoux 2002: 28-30). Causality is viewed as context- and conjuncture-sensitive (as in policy evaluation indeed; see below). Hence, by using QCA, the researcher is urged not to « *specify a single causal model that fits the data best* » (which is often done with most conventional statistical techniques), but instead to « *determine the number and character of the different causal models that exist among comparable cases* » (Ragin 1987: 167).

On the other hand, QCA also ambitions to meet some key strengths of the quantitative approach (Ragin 1987: 12ff; De Meur and Rihoux 2002: 20ff). Firstly, it allows one to analyze more than a few cases, and from there on to produce – to a certain extent – some generalizations. Secondly, it relies on formal tools (Boolean algebra) and is analytic in nature, in the sense that each

case is reduced to a series of variables (a certain number of *conditions* and an *outcome*). At the same time QCA is not radically analytic, as it leaves some room for the holistic dimension of phenomena. Thirdly, it is a replicable analysis, in the sense that « *a researcher B who uses the same variables and makes the same choices as a researcher A will reach the same conclusions as the latter*» (De Meur and Rihoux 2002: 27ff). This replicability also opens up the way for other researchers to verify or falsify the results obtained in the analysis. Finally, the Boolean technique allows one to identify «*causal regularities*» that are *parsimonious*, i.e. that combine only a few conditions, and not all the conditions that have been considered in the model.

Besides constituting a middle way between the holistic and analytic strategies, QCA is particularly well-suited for « small-N » or « intermediate N » situations and research design (De Meur and Rihoux 2002: 24). Moreover, QCA allows one to consider both phenomena that vary qualitatively and phenomena that vary quantitatively. Both of these phenomena can be operationalized in the conditions and outcome variables used for software treatment (De Meur and Rihoux 2002: 32). So, while « cases do matter, and each case matters » in QCA (in this sense, it has a strong « qualitative » preoccupation), QCA really lies at the crossroads of qualitative and quantitative analysis.

## 3.2 QCA as a Technique

QCA has been developed in the form of a software. The latest version (the «crisp» part of the fs/QCA software) is still under development, but already available as a freeware; so is TOSMANA, a software developed at the Marburg University, that performs similar analyses with some additional features (see also Cronqvist and Berg-Schlosser, in this volume).[5]

The key philosophy of QCA as a technique is to « *(start) by assuming causal complexity and then (mount) an assault on that complexity* » (Ragin 1987: x). The tool which is used for this purpose of reducing complexity is Boolean algebra, the « algebra of logic ». It would be impossible to give a clear idea of all the technical details and steps in this article.[6] In a nutshell, the researcher must first produce a raw data table, in which each case displays a specific combination of conditions (with «0» or «1» values[7]) and an outcome (with the « 0 » or « 1 » value). The software then produces a *truth table* which displays the data as a list of *configurations* – in a more synthetic way, as several different *cases* may very well display the same configuration[8]. Then the key step of the analysis is Boolean minimization: by using Boolean algorithms, the software reduces the long Boolean expression (which consists in the long description of the truth table) into a much shorter expression (the *minimal formula*) that shows the causal regularities – the different causal

paths, called *prime implicants* – that were, in a way, « hidden » in the data. It is then up to the researcher to interpret this minimal formula.

Two more strengths of QCA as a technique deserve to be mentioned. On the one hand, it can be used for at least five different purposes (De Meur and Rihoux 2002: 78-80). The most basic use is simply to summarize data, i.e. to describe cases in a more synthetic way (by producing a table of configurations). Hence it can be a useful tool for data exploration, for instance to construct typologies in a more inductive way (for a more detailed discussion of typology-building with set-theoretic methods, see Kvist, in this volume). It can also be used to check the coherence within the data: when the researcher discovers contradictions, this allows him/her to learn more about the individual cases. The third use is to test existing theories or assumptions, i.e. to eventually corroborate (validate) or refute (falsify) these theories or assumptions. QCA is hence a particularly powerful tool for theory-testing (for example for testing the "causal theory" underlying a public policy). Fourthly, it can be used to test some new ideas or assumptions formulated by the researcher (i.e. not embodied in an existing theory); this can also be useful for data exploration. Last but not least, it allows one to elaborate new assumptions or theories : the minimal formula obtained at the end of the analysis can be exploited and interpreted – i.e. confronted with the cases examined – and eventually lead the researcher to put forward some new segments of theory, in a more inductive way.

On the other hand, in the course of the procedure, the researcher is confronted with choices. For instance, he/she must decide whether or not he/she wants to obtain the shortest solution possible (i.e. achieve a maximal level of parsimony). If this choice is made, this means that some *logical cases* (also called *remainders*, i.e. cases that exist logically, but that have not been observed in the data) will be included in the «black box» for the Boolean minimization[9]. The point is that the researcher may very well reject this option, and hence prefer more complexity over more parsimony.[10] One also has to make clear choices on which variables to include and how to dichotomize them. The bottom line is that QCA is a particularly transparent technique, insofar as it forces the researcher not only to make choices on his/her own (he/she decides, not the computer), but also to justify these choices, from a theoretical and/or substantive perspective.

Hence QCA really forces the user to always keep an eye on theory... and the other eye on the real-life, complex cases behind the coded data, not only on the tables and formulae produced by the software. Thus QCA is both theory-driven and inductive: although induction does play an important role, there is quite a significant input of theory in QCA (Ragin 2004; see also Befani and Sager in this volume). For instance, quite clearly, the selection of variables that will be used for the analysis – and the way each variable is

operationalized – must be theoretically informed (De Meur and Rihoux 2002: 40).

## 4. METHODOLOGICAL CHALLENGES OF POLICY EVALUATION

Every public policy or action program evaluation faces (at least) four methodological challenges that are intertwined: How to identify the causal mechanisms underlying the policy "outcomes line"? How to measure the net policy effects? How to produce counterfactual evidence? How to triangulate methods and data? In the following sections, we summarize these traditional evaluation issues.

### 4.1. Explaining Policy Effects: "Testing the Program Theory"

Each program or public policy is based on (most of the time implicit) "outcome line", "causal chain", "theory of action", "policy rationale", etc. This outcome line consists in beliefs, assumptions and expectations about the nature of the change brought about by program action and how it results in the intended policy outcomes. Thus, every policy can be interpreted as a theoretical construction whose consistency and rationality must be questioned analytically by the evaluators: *"a policy can be interpreted as a theoretical construction, in the sense that it implies an a priori representation of the measures implemented, of the actors' behaviour, of the sequence of measures undertaken and of the effects produced on society "* (Perret 1997: 292, our translation). The first task of a policy evaluator is thus to re-construct this program theory.

Such a *program theory* is generally understood as a causal theory: it describes a cause-and-effect sequence in which certain program activities (administrative outputs) are the instigating causes and the social benefits (policy outcomes) are the effects that they eventually produce. Within a program theory one can further distinguish an *impact theory,* relating to the nature of the change on outcomes brought about by program action (links between outputs and outcomes), and a *process theory,* depicting the program's organizational and resources plan (links between the implementation arrangement and outputs).

The model of causality of a public policy is always a normative representation of the "operation" of society and the State. Proof of its validity comes through implementing and evaluating the effects of public policies. For an empirical analysis it is therefore necessary to distinguish the elements that

constitute this outcome line. Program evaluation thus involves empirical testing of the validity of the causality model on which the program is based. The analysis concerns both the relevance of this program theory and the scope of its practical implementation.

Evaluation studies might (often) identify failures within the program designs and, thereby, explain missing policy outcomes. The ineffectiveness and adverse effects of certain policies often derive from false or incomplete hypotheses of the impact and process theories. Several ineffective policies can be found in the field of urban traffic planning. For example, nowadays, the management of *public* parking spaces is one of the solutions adopted in order to direct, level off and reduce private motorized transport and, in particular, the volume of traffic arising from commuting. As a new transport policy measure, residents' parking disks are intended to restrict the periods during which non-residents can park in certain city neighborhoods. The aim of this policy measure is to remove commuter traffic from residential neighborhoods and to improve the quality of life of local residents and traders. Evaluation studies on the contribution made by the residents' parking-disk model to the reduction in the volumes of commuter traffic in the cities of Zurich, Basel and Bern (Switzerland) conclude that this measure remains largely ineffective (Schneider *et.al.* 1990, 1992, 1995). Between 70% and 85% of commuters using private means of transportation already had *private* parking spaces prior to the introduction of the disk. A clear majority of the target groups (i.e. commuters) have their own private parking spaces or the use of one owned by their employers, thus they do not have to adapt their behavior (by ensuring their mobility using public transport). This is an example of the incorrect choice of policy instruments, of the bad formulation of the action hypothesis of the impact theory.

Hopefully, evaluations can also conclude that a public policy is effective and produce the intended effects. Anyway, such an assessment should also be based on a careful investigation of the outcome line. An example of an effective policy is that of public support for home ownership in Switzerland. In 1970, an average of 28.1% of households were home owners (three quarters of the Swiss people are tenants; they have to rent their place of residence). This percentage was very low compared with other European countries. Thus, the Swiss Confederation passed a bill (1974) supporting residential construction and access to home ownership. This bill contained the following measures to reduce the initial costs incurred by future home owners: a federal guarantee, a reduction in the price of land, and non-reimbursable supplementary reductions. The main objective of this policy was to increase the rate of individual residential property owners in Switzerland. According to an evaluation of this bill (Schulz *et.al.* 1993), the federal support of access to home ownership had the desired effect. Up to 1991, some 15.747

construction projects were financially supported by the Confederation (outputs). Access to home ownership with the help of public support was primarily of assistance to young households which, in view of their limited finances, would not otherwise have had a opportunity to become home owners. Thanks to this measure the proportion of home owners increased during the study period (around 15 years) to reach 31.3% (i.e. outcomes in accordance with the objective). Furthermore, the bill had other indirect positive effects: in a period of recession, the support of access to home ownership constituted an important asset for the economy. This was the case for example in 1991 (a weak period in the construction sector) because 20% of family housing built was supported by federal aid.

These two examples illustrate how useful it is to follow - both conceptually and empirically - the whole outcome line of public policy, in order to find out and explain where the program theory could be incomplete or even absolutely false.

Challenge 1: *"The evaluator should test the program theory of the public policy to be evaluated. He/she should reconstruct the outcome line of the public policy even if it remains implicit in the policy design".*

## 4.2. Isolating Net Policy Effects: "Purging the Confounding Factors"

The starting point for an evaluation of a public policy is the identification of one or more measurable outcomes that should represent the goals of the program (see examples above). A critical distinction must be made here between gross outcomes and net outcomes. *Gross outcomes* consist of all the changes (in an outcome measure) that are observed when assessing a public policy. Gross outcomes are normally easily measured as the differences between pre- and post-program values on outcome measures.

*Net outcomes*, also referred to as *Net effects*, are much more difficult to isolate. These are the changes on outcome measures that can be reasonably attributed to the program and not to other contextual variables. In other words, gross outcomes include net effects of course, but they also include other effects that are not produced by the program to be evaluated, i.e. that are produced by *other factors and processes* occurring during the period under consideration (such as other public policies, contextual events, etc.).

The evaluation of a public program that aims at reducing energy consumption by industry and households provides one example of this difficulty. The instruments of that policy are typically information campaigns to enhance energy efficiency of industrial production processes and of the use of heating systems in individual houses. The evaluator may analyze the

evolution of the energy consumption statistics before and after the information campaign. He might observe that there is a clear decrease in the overall energy consumption of both the industry and household sectors. However, he/she cannot conclude (without further in-depth analysis) that this decrease is directly linked to the energy policy put in action. It may well be the case that the industry decreases its energy consumption because there is an economic recession and thus less industrial production. In the same way, the decrease in household energy consumption can result from a less cold winter and thus lower heating needs. Alternatively, it may well be that both industry and households face an important increase of energy prices and decide – for financial reasons that have nothing to do with the information campaign – to reduce their energy consumption. It may also be the case that a more complex combination of all these other evolutions (i.e. not linked to the program) is at work.

Challenge 2: *"The evaluator should purge the confounding factors (other public policies, external factors)".*

## 4.3 Estimating the Policy Deadweight: "Producing Counterfactual Evidence"?

The crux of the evaluation of a program with respect to a particular outcome is a comparison of what did appear after implementing the program (see points 3.1 and 3.2 above) with what would have appeared at the outcomes level had the program *not* been implemented. This pivotal element of any evaluation, which can never be observed and can never be known for certain, is known as the *counterfactual.*

The counterfactual (that should be estimated by the evaluator) is the quantitative score or level at which the outcome of interest would have be found had the program (to be evaluated) not taken place. In other words, the evaluator should compare a *policy-on* with a (fictional) *policy-off* situation. The real impact of the policy is then considered to be the difference between the real changes (at the outcome level) after the policy intervention with the fictional changes estimated in a situation without the policy intervention. Such an analysis prevents evaluators from over-estimating the effects of the public policy.

For example, the evaluator should determine the extent to which a public financial assistance scheme for job creation did make it possible to create new jobs and how much the absence of public financial assistance would have affected the creation of new jobs. If, say, public assistance has taken the form of job creation grants given to companies which would have created jobs anyhow (i.e. even without public assistance), the term *deadweight* is

employed for those who have benefited from the financial support. The net effect of the policy is then obtained by substracting the deadweight effect from the gross effects.

The evaluation of rural development policy in Denmark provides another example. Farmers were to receive a subsidy for the purpose of diversifying their activities. In a survey carried out as a part of an evaluation, the farmers were asked whether they agree with the following sentence: "the support received conditioned the implementation of my diversification project". About 75% of the assisted farmers gave a positive answer. An initial estimation of the deadweight would therefore be around 25%. However a complementary survey was carried out on farmers who had requested assistance but had been rejected for various reasons. The results revealed that all the farmers had implemented their diversification project even without the public assistance. Deadweight can therefore be considered to be 100%. This conclusion is nevertheless hasty, in so far as some of the farmers were not selected for assistance precisely because the project selection committee judged them capable of realizing their project without public support. In such a case, a more in-depth analysis would be necessary to gauge the real deadweight (European Commission 1999:113).

Challenge 3: *"The evaluator should estimate counterfactual evidence (in order to identify the deadweight of a public policy)".*

## 4.4. Comparing Comparable Cases: "Triangulating Methods and Data"

Policy evaluation within the European context (enlarged EU, clusters of regions, etc.) or within a federal country (several federated entities such as Belgian regions and communities, Swiss *cantons*, German *Länder*, etc.) is generally based on interregional comparisons. Various evaluations are launched by funding bodies (e.g. the EU or the central level of power in federal states) to evaluate how a similar program (e.g. a European Directive or a federal law) is implemented, and whether it has produced effects in various countries and/or regions. As a matter of fact, the Europeanization of domestic public policies, namely the response of the domestic policies to the EU policies (minimalist definition in Featherstone 2003), is a growing political reality. The definition of Europeanization is still under discussion in the literature (Cowles *et.al.* 2001; *Héritier et.al.* 2001; Featherstone and Radaelli 2003; Börzel and Risse 2003; Radaelli 2004) and it still covers a very broad range of theoretical and empirical issues. However, in order to explain the impact of a European policy on the domestic policy outputs and outcomes, one must above all scrutinize the behavior of domestic policy actors. Indeed,

Europeanization processes are always mediated by the domestic implementation networks, following different paths.

The methodological challenge is then to compare one (similar) European program, its implementation by domestic actors and its effects in a « small-N » design, with a limited number of various countries, regions and/or administrations implementing the same (causal mechanism of the) European program in various political, administrative, economic and social contexts. The methodological issue consists in identifying all the conditions (at the European level as well as at the domestic level) leading or not leading to the expected policy outputs and outcomes.

As a matter of fact, the evaluator may identify more than one unique (causal) path to the policy outputs and outcomes: more than one combination of (domestic) conditions can account for favorable policy effects. This is quite obvious within the European Union or within federal countries, as practical experience shows that policy effectiveness is often strongly dependent upon national and/or regional settings as well a upon sector-specific features, and that different cultural, political and administrative traditions often call for differentiated implementation schemes.

Furthermore, policy evaluation ideally requires additional comparisons in time (before and after the program implementation; see challenge 2) and space (regions with the program and regions without the program; see challenge 3). Thus, the triangulation of several comparisons is a crucial factor for the methodological quality of an evaluation design. The question is how to develop an evaluation design that combines diachronic and synchronic comparisons and, simultaneously, is still feasible from an economic point of view (e.g. costs of data collection and analysis). Furthermore, the evaluation should ideally compare comparable cases. In real-life policy evaluations, however, the empirical cases to be compared by the evaluator (for example the same policy implemented in all Swiss cantons) are defined by the political reality and not on the basis of methodological considerations. This hinders the development of a comparative evaluation design within a well-defined set of comparable cases.

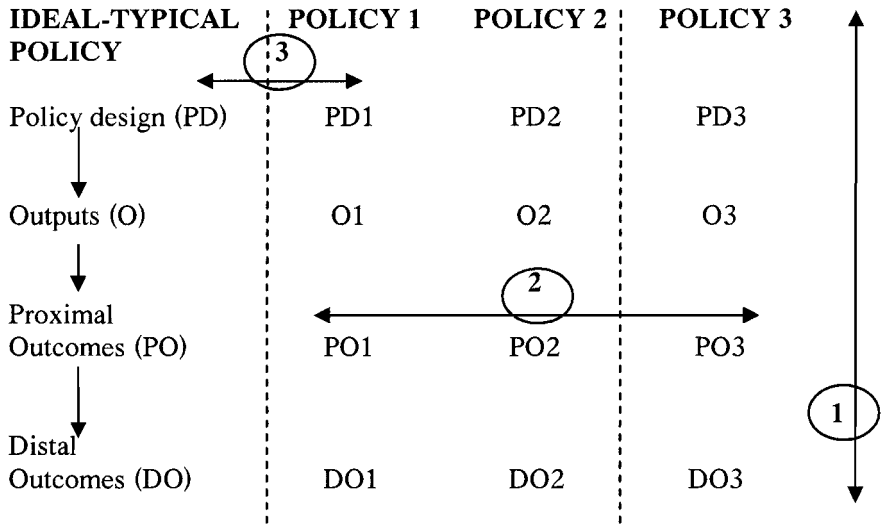
Challenge 4: *"The evaluator should triangulate many comparisons (before/after program, with/without program, cross-countries/regions/ public administrations/sectors, etc.) within a set of cases that are not (always) comparable."*

# 5. QCA ANSWERS

## 5.1 Introduction – Trying to Order Challenges and Answers

A proper QCA analysis consists out of three strategies to address the four methodological challenges of policy evaluation. First of all, since QCA is, in essence, a case-oriented strategy, it allows researchers to conduct a within-case analysis, in order to grasp dynamic within-case processes and identify causal mechanisms which link policy-design (in configuration with contextual conditions) to outcomes (cf. arrow "1" in figure below).

*Figure 10.1.* Policies and Comparative Strategies




Note: Practical example (of concept of the outcome line) for a public policy: widening of a road to boost the development of an isolated valley

- policy design: budgets allocated, bills used for compulsorily purchase of land (expropriation), information campaign, etc.;
- outputs: kilometers of new roads, enlarged intersections, improved road surfaces;
- proximal outcomes: number of cars using the road, reduced traveling time between two destinations, reduced number of accidents (but also more pollution for people living near the road etc.);
- distal outcomes: more rapid economic development of the towns served, improve access to leisure activities offered the regional capital (but also decreasing value of the neighboring land, etc.).

Secondly, QCA is comparative in nature, which allows researchers to identify differences and similarities across cases (cf. arrow "2"). Finally, both in relation to within-case analysis and cross-case analysis, QCA allows researchers to assess policy implementation vis-à-vis an ideal-typical policy-design. This Weberian ideal-type comparison is especially interesting for policy-evaluation of outcomes in theoretically well-developed fields (see also Kvist, in this volume). For example, in the case of the impact of institutions for the governance of common-pool resources, each institution can be compared to an ideal-type institution such as one constructed by Ostrom (1990). This strategy enables researchers to identify where the actual policy differs from the policy theoretically hypothesized to be effective (cf. arrow "3").

## 5.2. Challenge 1: Explaining Effects, Testing Program Theory

We will mainly examine the "process theory" side of this challenge, i.e. the challenge to analyze processes, and more specifically to identify *causal* processes linking explanatory conditions to outcomes.

In fact, such issues are very much discussed in the recent literature, also beyond the policy evaluation literature *stricto sensu*. Much attention has been paid recently to the importance of specifying causal mechanisms in order to open the black box of how interventions relate to outcomes (Hedström and Swedborg, 1998). In addition, many authors have argued that 'time matters' due to path-dependent and sequential processes (Abbott 2001; Pierson 2004) and that this should be taken into account in research designs. In contrast to general assumptions that the effect of a policy intervention can be observed after its implementation Paul Pierson (2004) argues that the time horizons of causes and outcomes are far more complex and interact in different ways. The general assumption is that the time horizon of the cause and the time horizon of the outcome are short. However, there are many instances in which this is not the case. A policy outcome can occur shortly after an intervention, but can also be spread out over a long time-period. In addition, policy interventions can have a non-linear impact in the sense that they first generate a shock, which then fades away.

The inclusion of these 'process' challenges in a research-design is not straightforward, especially as the N increases. One possibility is to focus analytic attention to the causal processes as such. This has led some authors to argue that researchers should focus more on "causal process observations" (Brady and Collier 2004; Bennett and George 2005). In order to do this, researchers need to get to grips with the cases under investigation and gain case-specific knowledge.

In this context QCA is a suitable research strategy since, while being a *comparative* strategy, it also in essence a *case-oriented* strategy which pays much attention to "case-based knowledge" (Ragin 2004) and to gathering different types of data on each case. In addition, the configurational minimal formulae produced by QCA identify the key components (conditions) of an explanation and in this way identify the main variables which are important for a mechanism based explanation. It is important to stress, at this point, that QCA as a technique (the software and its Boolean algorithms) does not identify, by itself, a "causeA → causeB → causeC" sequence which eventually produces a certain outcome. QCA does not *include* process as such. What QCA does provide, through the minimization procedure, are key "configurations" of factors (conditions). It is then up to the evaluator to interpret the minimal formula(e), e.g. in terms of sequence ("outcome line", as defined above) between the key conditions identified

Marx and Dombrecht (2004), for example, evaluated why some forms of work organization generated more repetitive strain injuries of the wrist than other forms of work organization for a limited though comparable set of cases. In a first step they compared different forms of work organization and identified different configurations which led to repetitive strain injuries of the wrist. The QCA results showed for example that, in organizations where people had to work at a steady pace, do not rotate between jobs but are able to individually decide when they can take a break, repetitive strain injuries occur. These results were at odds with existing literature which stressed that the possibility to individually insert a break is a good design principle to prevent repetitive strain injuries from occurring. The results/configurations from the QCA analysis were used to return to the cases to analyze which mechanism(s) played between the different variables to produce the outcome. By specifically focusing on causal process observations the researchers found that the key mechanism producing strain injuries was free-time maximization. Most workers saved up all their free time to go home earlier. Hence, the possibility to insert a break at own decision combined with a demanding job in which there is no variation in terms of rotation results in the occurrence of repetitive strain injuries.

Another interesting feature of QCA in this respect follows from its "multiple conjunctural causation" conception of causality (see above). Eventually, in most QCA analyses, the minimal formula will provide not only one configuration; most often, it will provide 2, 3 or 4 configurations ("terms") leading to the *same* outcome. Hence the researcher will be able to interpret (reconstruct) not one, but 2, 3 or 4 (partly or totally) different sequences. This is also coherent with real-life evaluation practice: it gives an added value vis-à-vis most quantitative methods which lead to the identification of "one best way".

In addition, with QCA, the evaluator is 'forced' to operationalize a dichotomous outcome (e.g. success v/s "non-success" or "failure"). Hence the evaluator can also, in a separate analysis, systematically identify the configurations – also to be *interpreted* as a sequence, as above – which lead to a "negative" outcome. Most often, with QCA there is no symmetry between the explanation of a "1" outcome and that of a "0" outcome; this also is in line with evaluation theory and practice.

## 5.3. Challenge 2: 'Net' Effects; Purging Confounding Factors

Assessing an outcome implies taking many potential different explanatory conditions into account. In order to assess the impact of the factors under investigation, the evaluator should carefully construct a research population. Indeed the careful construction of a research population is an important step in a QCA analysis, too. This can be done by applying the Most Similar Different Outcomes (MSDO) design for constructing a research population (Przeworski and Teune, 1970). This design applies two key principles:

- Principle 1: Maximize the variation on the outcome and conditions (explanatory variables) under investigation.
- Principle 2: Homogenize as much as possible on other possible explanatory conditions.

These two principles imply that the evaluator draws a clear distinction between the conditions that will be included in the model (to be tested through QCA) and those that will be left out. This implies, in turn, that the evaluator carefully constructs the research population in such a way that some possible relevant explanatory factors (apart from the program-linked conditions) are held constant in order to 'neutralize' their effect (see also Collier 1993). Hence the evaluator should come as close as possible to an "experimental" design. Those principles (and practical consequences thereof) are of course common to all well-thought comparative research designs, but they are especially relevant in the context of evaluating policy relevant outcomes which are not linked to a specific policy program.

In practical terms, these requirements are probably difficult to meet in real-life policy evaluation work, as some variation in contextual variables is still most likely to remain. One of the reasons thereof is that the research population may be a "given", e.g. defined *a priori* by the public authority (see also above). For instance, the European Commission may request the evaluation of a specific policy in all 25 EU countries. In such a situation, the evaluator will most probably consider that (national) contexts display a lot a variation – hence several contextual variables will have to be added to the

main program conditions (e.g. those identified by theory, by previous studies etc.). The problem, then, is that the evaluator will find him/herself in a problematic "few cases, many variables" situation.

One methodological strategy to meet this difficulty is to run separate QCA analyses on different clusters of cases that are sufficiently similar in terms of context, e.g. 3 clusters: Northern European EU members, Southern European EU members, and the 10 new (Central and East European) EU members. The problem with such a strategy is that it will make it more difficult to reach sufficiently general/comparable conclusions. So while being methodologically sound, this strategy may prove less interesting from the policymaker's perspective.

Another alternative strategy would be to include only *one* additional overall contextual macro-condition to the model for the QCA analysis. If, say, the evaluator is able to distinguish 2 main clusters of cases (say : the 10 new EU member states versus the 15 other states) that differ on quite *several* contextual variables, then he/she can add a condition [NEW], with a '1' score for the new EU members states and a '0' score for the other countries. If the minimal formulae reached at the end of the minimization procedure do not contain this [NEW] or [new][11] condition, then one may conclude that the contextual variables may be left out of the model. If it does contain this [NEW] or [new] condition, then the evaluator has a problem, and some of the (more detailed) contextual variables will need to be added to the model initially tested.

A third way to address this difficulty is to use QCA in a more inductive, iterative way. The first step would consist in producing a truth table adding up both all program and all contextual variables (not an exaggerated number of conditions altogether, of course). Provided there are no contradictory configurations[12], the second step would be to take out of the model, one at a time, contextual conditions, until this eventually causes contradictions. Then, finally, the evaluator can run the QCA analysis (minimization) with the shortest operational model possible, i.e. the model containing as few contextual conditions as possible, and yet still displaying no contradictory configurations.

Eventually, at the end of the second or third procedure explained above, if we take it for granted that *some* contextual variables will have been included in the model (for the reasons explained above), if the minimal formulae do not contain any of the contextual conditions, then the evaluator will be able to evacuate contextual conditions altogether (i.e. concluding that the outcome is really linked to the program – or some specific conditions within the program) and thus identify a high net effect. If not, then the researcher will most likely have to conclude that, in *some* cases (those cases covered by the prime implicants in which contextual conditions show up), the net effect is much

lower than in some other cases. In any event, QCA will not allow the evaluator to actually quantify the net effects of the program.
Hence QCA can indeed be useful to make some progress towards an assessment of 'net' effects, especially when they are configurational in nature.

## 5.4. Challenge 3: Counterfactual Evidence

On a more general level, the use of counterfactuals (broadly defined, i.e. "non-observed" cases, called "remainders" in QCA jargon) lies at the heart of the QCA minimization procedure. It is actually the resort to counterfactuals which allows one to reach more parsimonious minimal formulae. One of the strengths of QCA is also that it *explicitly* addresses the issue of counterfactuals in the course of the analysis (De Meur and Rihoux 2002, 2004; Rihoux and Ragin 2004).

This original feature of QCA can be exploited for evaluation research, as follows:

1. The evaluator should choose, in the minimization procedure of the "1" outcome, to include remainders. This allows the software to select some non-observed combinations of conditions, to which it attributes a "1" outcome score (this is a "simplifying assumption"), hence allowing to express most parsimoniously the regularities (combinations of conditions) shared by the *observed* cases with a "1" outcome. The evaluator then asks the software to produce a list of these simplifying assumptions.
2. The same goes for the minimization procedure for the "0" outcome.
3. When this is done, the evaluator can interpret these lists of non-observed cases selected by the software. If there are no contradictory simplifying assumptions (NB: this is a key requirement; for technical details, see De Meur and Rihoux 2002; Vanderborght and Yamasaki 2004), this means that the evaluator will have at his/her disposal a list of non-observed cases with a "0" outcome. Within this list, he/she will be able to check whether or not one of these non-observed cases comes close to the "absence of program" situation.

Example: consider a simple model such as: [A + B + C = OUTCOME], where A, B and C identify 3 program conditions. If, say, the following non-observed case were to be selected by the software to minimize the "0" outcome:

*a b c = outcome*

(to be read as follows : a "0" score for condition A, combined with a "0" score for condition B and a "0" score for condition C, leads to a "0" outcome score): this would mean that the absence of program is very likely to lead to a "negative" outcome.

Another QCA procedure which could be usefully exploited would be to "cross" (= intersect) actually observed cases with "hypotheses". For instance, it is possible to cross the minimal formula for the "1" outcome with the following hypothesis: "a b c = outcome" (for practical examples of this intersection technique, see Peillon 1996; Watanabe 2003; Yamasaki 2003).

All this being said, the informed use of counterfactuals with QCA will not allow the evaluator to actually quantify the level at which the outcome of interest would have be found had the program (to be evaluated) not taken place – for the obvious reason that QCA is dichotomous. This might be attempted by using MVQCA or Fuzzy Sets (however, in these two latest options, the number of potential counterfactuals will become huge, so it will be much more difficult to examine these counterfactual cases in a systematic way).

## 5.5. Challenge 4: Triangulating Comparisons

As far as the number of cases is concerned, it is quite obvious that QCA is "tailor-suited" for a small-N (meaning: "intermediate-N": from ca. 10 to ca. 40-60[13]) research design. Hence it is perfectly suited for policy evaluation at the cross-national level (e.g. within the E.U.) and cross-regional level (within a country or across countries in the EU). This statement should however be qualified. One the one hand, there should be enough information on what is shared (background, contextual characteristics) between all cases and thus can be left out of the model (see above). On the other hand, there should be enough information (qualitative and/or quantitative) on each case. In a EU or within-European national context, it is reasonable to expect that such conditions are quite often met.

In such a design, it is also perfectly possible, with QCA, to take into account program-linked conditions as well as some case- and context-specific conditions. Indeed, the evaluator should inject 3 types of conditions in the model to be tested:

- program conditions;
- sector-specific features;
- context-specific features (cultural, political or administrative traditions, constraints etc.), i.e. features that are specific to the case or the case's "environment".

As far as the (causal) conclusions reached at the end of the analysis are concerned, QCA will indeed identify, most often, *more than one* combination of conditions leading to the desired output or outcome (or indeed to a negative output or outcome). This central feature of QCA (see "multiple conjunctural causation", above) is thus fully in line with the expectation that policy effects

are often strongly dependent upon national and/or regional settings as well as upon sector-specific features, and that different cultural, political and administrative traditions often call for differentiated implementation schemes. Hence if the evaluator includes the right ingredients as conditions (see above), he/she will most probably identify at least 2 or 3 different (causal) paths.

Finally, regarding triangulation, both in time and in space, there are some practical ways in which QCA can be used to this end. One quite straightforward way to proceed is to include a condition such as "program implemented Y/N" [PROGR] in the model. This allows one to compare a set of cases (times 2) before and after implementation. Insofar as there are no contradictory configurations, this will also clearly show whether or not "the program matters" for a favorable outcome, and "how the program matters" – by looking at the other conditions that are associated with [PROGR] in some prime implicants (i.e. in the terms of the minimal formula).

The same sort of *modus operandi* can be applied for regions with and without the program. This can be operationalized by including one additional condition. Alternatively, one could distinguish 2 specific subsets of cases (regions), perform 2 separate QCA analyses, and then see whether the same (causal) combinations can be identified in the 2 minimal formulae.

Such triangulations are possible with QCA at a relatively low cost, but provided that enough data has already been collected. The best situation would be one where some good quality – and comparable – case studies are already available. In such a situation, QCA can then be used to systematically compare those cases.

## 6. REMAINING CHALLENGES AND NEXT RESEARCH STEPS

In the previous sections, we have demonstrated that QCA clearly has the potential to yield added value for policy evaluation, in particular within a comparative context. Of course, one limitation of our demonstration is that we have not (yet) proved our point on real-life policy evaluation data (see however Befani and Sager, in this volume).[14] This would be an obvious next step. In the meantime, let us point at some further potential benefits of QCA for policy evaluation, as well as some remaining challenges and difficulties.

One further interest of QCA for policy evaluation is that it could support the evaluation process itself, thereby enhancing the evaluation. Indeed, QCA is a very transparent tool. Thus, for instance, once the data has been dichotomized, the evaluator (and all the stakeholders participating to the evaluation) can very easily control the coding of the conditions, modify the dichotomization thresholds for further tests, include other conditions, discuss

the robustness of evaluation results, etc.). Hence QCA is also potentially useful for pluralist, participative and empowering policy evaluation. This is also the case of the newly developing MVQCA (Multi-Value QCA; see Cronqvist 2004, and Cronqvist and Berg-Schlosser, in this volume) and Fuzzy Sets (see Kvist, in this volume), which display the further advantage that they can handle more fine-grained data (of course, the other side of the coin is that the data tables are somewhat more complex; but they are usually still not-too-complex for a more participative policy evaluation).

In addition, QCA is also useful for the synthesis of evaluations conducted on the same policy but by various evaluation teams and/or in various countries/regions. Relying on existing qualitative evaluation case studies conducted by different evaluation teams, QCA enables one to perform "systematic comparative case analysis" (see also challenge 4, above; and Befani and Sager, in this volume).

However, some challenges also remain to improve the use of QCA for evaluation purposes. The most important challenge concerns issues related to model specification and case selection. True, this is not only a challenge for a QCA-type of analysis. However, it is more important in QCA than for other types of analysis since QCA is more deterministic in nature.

In principle, it is recommended that QCA users go back and forth between theory and evidence to produce an explanatory model which might contain several causal paths to an outcome. It is important to note that multiple paths to a given outcome are all contained within an initial model which is a result of empirical inductive and theoretical deductive work. That is the reason why much attention is paid to the relationship between theory and data.

Such a research process does not really allow for contradictions to occur. All contradictions must be solved (see above) by developing an initial model on which a QCA analysis proper, i.e. Boolean minimization, can be applied (surely, contradictions can also be solved by other means; see below). Hence, in principle there is always a/one best fit model (BFM). However, it is not always possible to go back to the policy cases and collect additional data to develop this one BFM. Consequently, it is not always possible to fit the model to the data and evaluators need to work with the variables they have at hand. The problem is worse in a synthesis of evaluations where researchers are neither able to generate new data, nor to develop new explanatory conditions which require additional data-input.

The key challenge here consists in developing criteria for the selection of an explanatory model, which adheres to a configurational logic and can be used for a QCA-type of analysis that produces more or less parsimonious results. In other words, which initial model should be chosen to conduct a QCA-type of analysis? Given a number of variables one can develop many different models to be processed in a QCA-analysis. The total number of

possible models is given by $2^k - 1$ (with k = # variables). Since many policy evaluators are confronted with at least ten possible relevant explanatory variables (conditions), the number of models to choose from becomes very large.

In addition, it could be argued that QCA works best – in terms of parsimony and identification of multiple causal paths – when one works with between 4-7 variables. Yet, in many types of evaluation research more than 7 explanatory variables might be identified as potentially significant components to explain an outcome. Hence it becomes quite crucial to select a valid model to conduct a QCA analysis (see also Amenta and Poulsen 1994). The selection of the model is hampered by the fact that two problems occur when one works with either too few or too many variables in QCA.

On the one hand, if one includes too many variables, a problem of uniqueness might occur, i.e. each case is then simply described as a distinct configuration of variables. This results in full complexity and no parsimony, which might be of limited relevance to policy-makers. With a limited set of cases, this problem starts to occur from 8 variables onwards. On the other hand, if one uses too few variables the probability of contradictions, i.e. the fact that an identical model/configuration both explains successes and failures of policy, increases. This problem easily occurs with models of less than 4 variables, which indicates that there is an important omitted variables bias.

How can these two problems be solved? As far as the problem of uniqueness is concerned, the only solution is to develop limited explanatory models. This implies that the number of variables of an explanatory model should be significantly lower than the number of cases (see also Marx, 2005). With regards to the problem of contradictions, there are several possible ways to deal with it (see e.g. Clément 2004; De Meur and Rihoux 2002; Vanderborght and Yamasaki 2004). First, a new more homogenous and comparable research population can be constructed, by including new cases or removing cases (however this is not always possible when cases are "given"; see above). Secondly, new variables could be included in the explanatory model. Thirdly, existing variables – including the outcome variable, possibly – could be recoded or reconceptualized. A final way to deal with contradictions is to keep only those configurations which contain at least two or more cases for the minimization procedure, since it is often the case that contradictions are generated because only one contradictory case occurs. These contradictions are disregarded when one specifies that at least two or more cases should be covered by a given configuration. The drawback of this decision is that it decreases the number of cases in an analysis and hence excludes possible relevant configurations (and real-life cases, from the decision-maker's perspective). This is especially problematic when one works with (biased) samples and when the aim is to explore data and generate

hypotheses. Concerning the latter it is best to exclude as few cases as possible and hence proceed with an analysis of all possible cases.

However, all the possible solutions we have sketched here often only produce partial solutions. Moreover, it should be noted that there seems to be a trade-off between the two main problems of contradictions and uniqueness. The shorter (i.e. the most parsimonious) the models, the more contradictions; the more extensive the models, the less possibility to summarize data and obtain parsimonious explanations. Hence, increasing the number of variables to solve the problem of contradictions is not really a solution. The risk with increasing the number of variables is not only indeterminacy but also uniqueness.

How then to proceed? How to select a suitable model or models for evaluating policies? One possible solution to the problem of model specification is to develop a "two-step approach" in QCA, i.e. to distinguish "remote" from "proximate" conditions (Schneider and Wagemann forthcoming). A problem with the two-step approach is that it is very sensitive to contradictions. A first step of the two-step approach, with a small number of "remote" conditions, will most often result in a very high proportion of contradictions.

Another possible way is to be tolerant of contradictions and try to identify the model which best fits the data in terms of balancing the number of contradictions and the number of configurations. In other words, the aim is to find a model with jointly the least configurations (reduction of complexity – parsimony) and the least contradictions. This means that one does not select the model with the least configurations or the least number of contradictions, but the model which scores best on the two criteria combined. This 'sub-optimal' model can then be used for further QCA analysis.

## NOTES

1 Some work by Sager (2002, 2004), Befani (2004) and Balthasar (2004) is, however, specifically dedicated to the field of policy evaluation. See in particular Befani and Sager, in this volume.

2 Note that a policy evaluation can also be undertaken before the public policy is formally adopted (ex ante evaluation) or during its implementation process (on-going or mid-term evaluation).

3 In QCA terminology, a *condition* stands for an « explanatory variable », or an « independent variable ». NB : it is *not* an independent variable in the statistical sense of the term.

4 In QCA terminology, as indeed in evaluation terminology, the *outcome* is the ultimate « dependent variable ».

5 For more information, see the "software" page on the COMPASSS resource site (http://www.compasss.org). See also Cronqvist (2004).

6 For an accessible presentation of some key elements of Boolean logic, see (Ragin 1987: 103-163). For more details on the concrete steps of the analysis and use of the software, see (De Meur and Rihoux 2002; Ragin and Rihoux 2004; Rihoux *et.al.* 2003).

7 The TOSMANA software also allows multivalue coding.

8 To put it short : a *configuration* is a given combination of some conditions (each one receiving a « 1 » or « 0 » value) and an outcome (receiving a « 1 » or « 0 » value). A specific configuration may correspond to several observed cases.

9 In technical terms: the software will give a « 0 » or « 1 » outcome value to these logical cases, thus making *simplifying assumptions* about these cases.

10 Actually, some middle paths also exist, between maximal complexity and maximal parsimony (Ragin and Rihoux 2004).

11 In QCA notation, [NEW] (uppercase) stands for a "1" score and [new] (lowercase) for a "0" score.

12 A contradictory configuration is a combination of conditions with the *same* condition values that leads to *different* outcome values (a "1" outcome for some cases, and a "0" outcome for some other cases), thereby producing a logical contradiction. See also Cronqvist and Berg-Schlosser, in this volume.

13 Actually it is also possible to use QCA with less than 10 cases. As for the upper limit, it is also possible to treat much more than 60 cases, as long as it is possible to gain some case knowledge for each case included (Ragin and Rihoux 2004).

14 Note that QCA has already been used in quite numerous policy *analysis* applications (see full list via: http://www.compasss.org), but indeed not specifically policy *evaluation*, with a few exceptions (see note 1 above).