Qualitative Comparative Analysis as an Evaluation Tool. Lessons From an Application in Development Cooperation

Pattyn, V.; Molenveld, A.; Befani, B.

Qualitative Comparative Analysis as an Evaluation Tool: Lessons From an Application in Development Cooperation

Valérie Pattyn[1], Astrid Molenveld[2,3], and Barbara Befani[4]

Abstract

Qualitative comparative analysis (QCA) is gaining ground in evaluation circles, but the number of applications is still limited. In this article, we consider the challenges that can emerge during a QCA evaluation by drawing on our experience of conducting one in the field of development cooperation. For each stage of the evaluation process, we systematically discuss the challenges we encountered and suggest solutions on how these can be addressed. We believe that sharing this kind of lessons learned can help evaluators become more familiar with QCA, shedding light on what it is to be expected when considering the application of QCA for an evaluation, at the same time reducing unfounded fears and promoting awareness of traps and requirements. The article can be insightful and potentially inspirational for both commissioners and evaluators.

Introduction

More than a quarter of a century after its introduction in the social sciences (Ragin, 1987), qualitative comparative analysis (QCA) found its way to the evaluation field. The method is increasingly considered a valuable alternative or complement to existing evaluation methods. That a widely disseminated study entitled "Broadening the Range of Designs and Methods for Impact Evaluations" (Stern et al., 2012) gives central attention to QCA is but one proof for this claim. QCA combines strong points of both qualitative and quantitative methods, aiming at "meeting the needs to gather in-depth insight into different cases and to capture their complexity, whilst at the same time attempting

---

[1] Institute of Public Administration, Leiden University, Leiden, the Netherlands
[2] Department of Public Administration and Sociology, Erasmus University Rotterdam, Rotterdam, the Netherlands
[3] Public Administration and Management, Antwerp University, Antwerp, Belgium
[4] Research Fellow, University of Surrey, Guildford, United Kingdom

Corresponding Author:
Valérie Pattyn, Institute of Public Administration, Leiden University, Leiden, the Netherlands.
Email: v.e.pattyn@fgga.leidenuniv.nl

to produce some form of generalization" (Befani, 2013; Rihoux & Lobe, 2009, p. 472). By systematically comparing cases as configurations of conditions and outcomes, evaluators can search for prevalent patterns and identify redundant conditions or conditions that do not seem to make any difference to explain a certain phenomenon. Applied to the evaluation field, the method in the first place serves a learning purpose: Via QCA, the evaluator can unravel explanatory patterns for "success" and "failure" of existing cases, with the possibility to inform potential future cases. The number of QCA applications in the field of evaluation is still rather scarce, especially when compared to the number of applications in academic circles, including the policy analytical literature (see, for instance, Rihoux, Rezsohazy, & Bol, 2011).

The ambition of this article is to address the potential and challenges of applying QCA in an evaluation context by sharing lessons learned from an evaluation in the field of development cooperation. Although every effort has been made to clarify the main concepts used in the QCA, the article is not intended to be a primer on the method and the reader might benefit from preliminary consultation of other QCA texts written for evaluators to fully grasp the method's characteristics and potential (Befani, 2016).

The evaluation, commissioned by the Dutch nongovernmental organization (NGO) Hivos, concerned two media programs conducted in Kenya and Tanzania and focused on the explanatory conditions that trigger a response from powerful actors, following publication of critical media products (e.g., article, documentaries, TV shows, etc.). Common to many evaluation settings, the aim was to systematically investigate the actual role of certain conditions that have a central position in the media programs' theory of change. Although our case example concerns the area of development cooperation, most of the lessons learned are inspirational for other contexts as well, since the opportunities and challenges we faced are typical of evaluation, rather than a specific policy field.

The authors of the article have QCA experience in both academic and evaluation settings, which enables them to identify challenges that arise when QCA is used and how these can be solved. The challenges discussed are of two types: a first type concerns issues commonly experienced by all QCA researchers, but which may be more at stake in an evaluation context. The second type of challenges relate to the application of a method developed in an academic context to an applied evaluation setting. The lessons learned can contribute to give QCA a firmer and a well-established position in the evaluation toolbox, which can be of use for both evaluation commissioners and evaluators.

The article is structured in three parts. In the first part, we present the abovementioned evaluation project, which we will use as an example throughout the text. The presentation of the project should help the reader understand why some challenges can arise when applying QCA. In the second section of the article, we introduce the basic characteristics of the method, and we summarize its potential for the evaluation field. These two sections set the scene for the actual core of the article: in third section, we systematically go through the various stages of the evaluation process and discuss which challenges may emerge when applying QCA. For each stage of the evaluation cycle, we discuss how we addressed the challenges in the case evaluation.

A QCA Evaluation of Media Support Programs in Kenya and Tanzania

In Kenya and Tanzania, the Dutch NGO Hivos administered two funds (Tanzania Media Fund [TMF; which developed into Tanzania Media Foundation in October 2015] and Kenya Media Programme [KMP]). Both funds were established to financially support the media (journalists, radio makers, and media houses) for the realization of investigative and critical media products (articles, documentaries, etc.). This intervention was driven by the assumption that the supported media products generate an effect in terms of an (accountability) response from politicians, businesses, or NGOs. In the most "successful" cases (in terms of generating a response), tangible and concrete

actions are taken by one or more of the abovementioned actors following the publication of a specific media product, with the intention to address the structural problems. With the aim to study the success of these funds running since 2008 in Tanzania, and since 2011 in Kenya, Hivos launched an evaluation. To clarify, the type of responses following the publication of media products can be many. Examples can be as diverse as, for example, the Tanzanian inspector general of police reshuffling regional police commanders following a newspaper report by a grantee and the Tanzanian Food and Drugs Authority banning a certain type of milk powder following another publication.

Three goals of the evaluation led the commissioner to choose QCA as the main method for this specific evaluation.

- The media programs could only be fully grasped through a lens of *causal complexity*: Most media products were to be conceived as a combination or "package" of conditions that "produces" a certain effect. For instance, and hypothetically: An article can be written about a very salient topic but might only trigger a response when it is combined with a strong media echo (i.e., other media picking up the story). In QCA jargon, this phenomenon is coined *as conjunctural causation*: only *in combination with other conditions* will a condition produce a certain outcome. Furthermore, the method does not assume the idea of a uniformity of causal effects (A leads to B).
- The commissioner was puzzled by the fact that a certain outcome could be triggered by numerous, nonexclusive combinations of conditions. In QCA terms, this is called *equifinality*. To elaborate on our hypothetical example, a response from a powerful actor can be triggered when an article is about a very salient topic, combined with a strong media echo or when an article is about a very political urgent topic, where politicians are explicitly named and shamed. The method is *not necessarily oriented toward the formulation of one single causal model that fits the data best* (Rihoux & Lobe, 2009) but allows for equifinality. No matter their frequency of occurrence, all rival explanations for a certain outcome are (in principle) theoretically equivalent (Schneider & Wagemann, 2012).
- Added to this, the commissioner and local staff members shared the assumption of *causal asymmetry*, meaning that if the presence of a particular (combination of) conditions is relevant for the outcome, its absence is not necessarily relevant for the absence of the outcome. Again applying this to our hypothetical example: If we would observe that actor response is triggered by producing a media product on a highly salient topic, that is also picked up by other media, it would be wrong to conclude that a media product that features a less salient topic, and that is neither picked up by other media, is doomed to "fail" in terms of actor response. Explaining the presence or the absence of an outcome thus requires separate analyses in their own right (Schneider & Wagemann, 2012, p. 6), as there can be different dynamics at play.

With these assumptions in mind, the evaluation was geared toward the following evaluation questions:

1. Under which conditions do the media products trigger response from powerful actors?
2. Under which conditions do the media products not generate any response from powerful actors?

Two of the authors of this article were appointed as external evaluators and were given the explicit mandate to answer the two questions by means of QCA. The evaluation started in fall 2013 and was concluded early 2015. Table 1 summarizes some key elements of the evaluation, such as the conditions, the outcomes, and the number of cases. Throughout the article, we will elaborate on several of these elements.

Table 1. Key Information About Cases,[a] Outcomes, and Conditions.

Key Information About Cases, Outcomes and Conditions		
Number of cases	Kenya	Tanzania
− Finished media products	87	527
− Media products covered in the evaluation	43 (49%)	217 (41%)
Number of qualitative comparative analysis analyses	4 Outcomes per country = 8 analyses	
Number of outcomes	4	
Presence response of		
1. powerful actors	• Fully in the set: to a strong degree—actor explains and justifies, and structural actions are taken	
2. citizens		
Absence of response of	• More in than out of the set: to a certain degree—actor explains and justifies, and actions are taken but do not address structural problems (e.g., symbolic actions)	
1. powerful actors		
2. citizens	• More out than in the set: to a small degree—actor explains and justifies but no subsequent actions	
	• Fully out of the set: no response at all	
Number of conditions	10	
Related to the individual journalists	High journalistic experience; high education	
Related to the individual media products	Strong media echo, high reach of media outlet, high currency of issue, high coverage of wrongdoings, high coverage of solutions, strong depth of story, and high salience of issue	

[a]Given the sensitivity of the evaluand (the media articles) and to protect the anonymity of the journalists, we will not elaborate on specific cases in this article.

QCA: What's in a Name?

In the scope of this article, we restrict ourselves to a concise explanation of the method. More extensive ontological and technical details can be found in specialized methodological textbooks such as Ragin (1987, 2000, 2008), Rihoux and Ragin (2009), and Schneider and Wagemann (2012) or in more recently published evaluation-specific outlets such as Befani, (2016).

As hinted at in the introduction of this article, QCA is often portrayed as a bridge builder between contextualization in terms of accounting for the idiosyncratic nature of specific cases and generalization in terms of unravelling trends across these specific cases (Verweij & Gerrits, 2013). QCA became initially popular in situations where researchers are confronted with a number of cases that is too small to apply variable-oriented methods, such as regression analysis, and too large to apply in depth within-case qualitative methods, such as process tracing (Beach & Pedersen, 2013). Via statistical methods, one can conduct different types of systematic comparisons, but these do not allow for rich contextual explanations and causal complexity. The focus on averages, typical for such methods, can come at the cost of understanding the complexity of individual cases. QCA is particularly suited to overcome this difficulty. Because of its potential to account for causal complexity and allowing for generalization, the method is also increasingly applied in settings that concern large-*n* (of which our case illustration is an example). While we, in our evaluation, do use the method to discover patterns across cases, the method can as well be deployed for other purposes including systematically summarizing and ordering qualitative data and inductively generating new theoretical propositions (Berg-Schlosser, De Meur, Rihoux, & Ragin, 2009; Verweij & Gerrits, 2013).

QCA belongs to the family of set theoretic methods. *Sets* are well-determined groups, in which the cases are *member* of to a certain extent, for example, and again hypothetically, it can be rather

easily argued that a journalist with a PhD degree would be a full member of the set: high education. In QCA, every case is conceived as a combination of conditions and one particular outcome. Conditions can be conceived as causal variables, determinants, or factors (Rihoux & Ragin, 2009, p. xix). In an evaluation context, an outcome will usually refer to constructs such as "effect," success, or "impact." By relying on set theory, QCA provides the possibilities to identify (configurations of) condition(s) that are sufficient and/or necessary for a certain outcome to occur.

- Will a condition (or a combination of conditions, i.e., configuration) 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 ($X \rightarrow Y$). For instance, an article that covers many solutions to the problem describes the root causes and explicitly addresses the wrongdoings of officials can constitute a sufficient path to actor response.
- 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 ($X \leftarrow Y$). For example, if we would find that all articles that lead to a response of a powerful actor, describe root causes, the latter can be qualified as a necessary condition.

Necessity and sufficiency are often of core interest in evaluation studies. To make their distinction and relevance clear, we exemplify this with the relationship between receiving development aid and the levels of famine in a country. A large amount of development aid is neither a necessary nor a sufficient cause for the absence of famine in a country. Having low famine rates does not necessarily imply that a country got a large amount of development aid. However, a large amount of development aid can be sufficient to reach low famine next to (e.g., a productive harvest season) or in combination with other factors (effective food supply and absence of government corruption). Using QCA, the evaluator can thus disentangle such relations between (contextual) conditions and the outcome.

In crisp set QCA (csQCA), the original version of QCA, conditions and outcome need to be translated in binary terms, 1 or 0. Conditions or outcomes assigned a score of 1 should be read as present (or high, or large, . . .), while variables with a score of 0 are regarded as absent (or low, or small, . . .). The data in this stage are transformed into categories that express qualitative differences in kind (Vink & Van Vliet, 2009). The latter procedure is coined *calibration* in QCA (Ragin, 2000; Rihoux & Lobe, 2009). In response to the binary calibration of data (0 and 1) common to csQCA, two other QCA techniques have been developed: fuzzy set QCA (fsQCA) and multivalue QCA (mvQCA). The "fuzziness" in fsQCA—the technique applied in the present study—basically refers to the fact that cases can have partial membership in a particular set. Set membership scores can take all values of the interval between 0 (=*fully out*) and 1 (=*fully in*). In fsQCA, cases are given a set membership score for each configuration. This implies that a single case can have partial membership in several configurations at once (Vink & Van Vliet, 2009). Unlike csQCA, the outcome is also fuzzy calibrated. fsQCA is thus most suited for the analysis of phenomena that not only vary qualitatively, in kind, but also quantitatively, in degree (Ragin, 2009). Similar to csQCA, fuzzy sets start from the observation that many social phenomena are dichotomous "in principle" but take account of the fact that empirical manifestations of these phenomena in practice often differ in degree (Schneider & Wagemann, 2012, p. 14). mvQCA, the third technique, resembles the crisp set variant but allows for the placement of more than one threshold in one or more conditions (not on the outcome). So, for instance, for a condition with three categories, the mvQCA values can be 0, 1, or 2. Similar to csQCA, the different values on a particular condition are discrete, while in fsQCA, a case can have all degrees of membership in a particular set (Vink & Van Vliet, 2009, p. 265). In our study, we applied fsQCA to account for both differences in kind and in degree and as such keep as much

case complexity as possible. No matter the QCA variant used, this translation in QCA should be strongly anchored in substantive or theoretical case-based knowledge (see subsequently).

Calibrating conditions and the outcome has several advantages. It is a transparent and replicable way to describe a case, increasing the study's internal validity (Berg-Schlosser et al., 2009, p. 14). Moreover, it enables the researcher to compare the different cases in a systematic and formal way. With the data being calibrated, a data matrix can be constructed which basically presents the empirically observed data as a list of configurations (i.e., a combination of conditions and an outcome). 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 summarizes 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. 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 & Wagemann, 2012).[1] The composition of the truth table enables the evaluator to verify the existence of *contradictory configurations*. Applying this to our case, contradictory configurations are media programs that share the same combination of conditions but which do not consistently correspond with the same outcome (i.e., in some cases, the presence of actor response; in other instances, the absence of response). The existence of contradictory configurations can alert the evaluator to possible measurement errors or to wrong operationalization of the conditions and/or outcome. Contradictory configurations can also urge the inclusion of extra conditions, so that cases are better discriminated from each other. The process of resolving contradictions necessitates that the evaluator goes back and forth between the cases and the analysis in an iterative way. The need for an intensive dialogue with the cases can generally be considered a strength of the technique, as it makes the evaluator and stakeholders acquire a better knowledge of the cases (Rihoux & De Meur, 2009). Once contradictions are solved, one can proceed to the most famous QCA procedure: *Boolean minimization*. The minimization procedure follows a "one-difference rule" (Baumgartner, 2012): It 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.

Challenges Faced in the QCA Evaluation Process

Although each evaluation process is unique, it typically runs through the same consecutive stages: (1) decision to evaluate, (2) establishing the evaluation design (i.e., structuring/planning the evaluation), (3) data collection, (4) data analysis, and (5) interpreting the findings.[2] Admittedly, splitting up the evaluation cycle in various stages is analytically possible, albeit in practice there will often be interaction and iteration between multiple stages (e.g., between stage of data collection and data analysis). As will become clear from the rest of the article, the iterative aspect is a key characteristic of QCA.

Nonetheless, irrespective the actual importance of a specific stage in a particular evaluation, we will discuss a series of challenges that we assume to be present in most QCA evaluations. The aim of this article is not to repeat the answers on frequently heard questions or critiques about QCA (De Meur, Rihoux, & Yamasaki, 2009), for instance, about the ontological premises behind the method or about the use of Boolean algebra. Our aim is instead to focus on the challenges that an evaluator may face when she or he decides to use QCA and to provide suggestions and possible solutions based on the experience acquired in practice. We will do this in the following paragraphs.

Stage 1: Decision to Evaluate

Determining the purpose of the evaluation. Various classification systems of evaluation purposes circulate in evaluation discourse (see, for instance, Balthasar, 2009; Vedung, 2009). Two major purposes dominate the literature: either evaluations are set up for *accountability* reasons, to provide information that can inspire decisions about program continuation, expansion, reduction, or termination, or evaluations have a *learning* purpose. Evaluation in the former perspective can contribute to the answer to the fundamental question "what works?" while the latter perspective enquires "why does it—or did it not—work?" When one would visualize the two purposes on a continuum, QCA evaluations will be closest to the learning pole rather than the accountability pole. Rather than "an effects-of-causes-stance" (what works question) which is typical for quantitative approaches, QCA follows "a causes-of-effects-stance" ("why does it, or did it not work?"—questions; Mahoney & Goertz, 2006; Vis, 2012). A QCA evaluation will be primarily oriented toward *understanding* what caused a certain effect and how. Applying this to our evaluation, the emphasis was not on the extent of actor response triggered by the media products but rather on the patterns toward response. The latter focus can sometimes be at odds with what is often the implicit assumption in an evaluation context. Donor institutions often emphasize the accountability motive, at the sacrifice of the learning purpose. Also in our case example, there was some initial skepticism about the purpose of the evaluation among the primary stakeholders in Kenya and Tanzania, precisely because it significantly deviated from the intentions of accountability oriented evaluations to which they were used. We learned that it is important to invest in explaining the purposes of a QCA evaluation and its benefits to all stakeholders. Therefore, the evaluation started with a concise training session on QCA, attended by the local (i.e., Kenyan and Tanzanian) Monitoring & Evaluation (M&E) and program officers and a representative from the Hivos main office. For instance, we explained the added value of the iterative search for conditions to add to the "explanatory model" (by solving contradictions), which was an insightful exercise for the teams. It challenged them to overthink the assumptions underpinning their program theory:

> if citizens and civil society have access to more (and) reliable information, provided by independent and critical media, they will demand (more) accountability of the state, businesses and NGOs. These will respond to the external pressure and become more accountable to its citizens. (Theory of Change— Tanzanian Media Fund)

Deciding on the locus of the evaluation. For each evaluation, it should be decided whether to conduct it in house or outsource it to consultants or universities (Pattyn & Brans, 2013). Every choice involves advantages and disadvantages. QCA evaluations are also subject to this dilemma. On one hand, a QCA evaluation requires evaluators to master the basic principles of the method and of the technical software associated with it. On the other hand, QCA as a case-based approach requires "intimate" knowledge of the cases (Rihoux & Lobe, 2009, p. 223). As we will further explain subsequently, this is necessary to be able to formulate the theory, to select the conditions and outcome, to decide on thresholds to code the conditions in binary or fuzzy values, and to interpret the findings following the analytical moment. This high emphasis on case knowledge complicates the outsourcing of a QCA evaluation but does not rule it out. In an optimal scenario, one chooses for an internal evaluator with the required QCA technical expertise, but we realize that this possibility will be rare. When choosing for an outsourced QCA evaluation, as in our case example, evaluators should invest in acquiring familiarity with the cases to compensate for the weaker case knowledge compared to that which in-house evaluators can usually rely on. This recommendation is common to most external evaluations, for instance, in case of theory-based evaluations and strongly applies to QCA evaluations. That is why we, as external evaluators, engaged in strong interaction during the entire evaluation process

with the commissioning organization and a representation of the M&E local teams in Kenya and Tanzania.

Stage 2: Establishing the Evaluation Design

Formulating the evaluation questions. A QCA evaluation is geared toward unravelling the combinations of conditions that produce a particular outcome. Conducting a proper QCA analysis is a time intensive undertaking, which is even more the case when there are multiple effects of interest to stakeholders. It is important to realize that for each outcome to explain, a separate set of conditions is to be selected, and a separate analysis has to be conducted for both the absence and the presence of the outcome. QCA requires not only that effects are known prior to the commissioning of the evaluation but also that the effects to be scrutinized are limited in number to keep the analysis manageable for all stakeholders (and to keep the budget manageable). One key challenge in a QCA evaluation is hence to reach consensus among stakeholders about the effects to be investigated and, preferably, to keep them limited in number. In evaluation, multiple strategies can help to reach such consensus. We can think of strategies as card sorting models (Davies, 1996) or color voting. In our evaluation, we used informal brainstorming methods to come to a consensus on outcomes. Within the scope of the case evaluation, it was decided to focus on two particular outcomes only: response from powerful actors (politics, businesses, or NGOs) and response from citizens. For each outcome, separate analyses have been conducted.

Selecting cases. While it is imperative in experiments to strive for as much similarity across cases as possible, except for the intervention variable (in particular in the two cases used for comparison between treatment and control), a QCA evaluation can handle many different degrees of variation. The higher the variation in cases, the stronger the validity of the findings. QCA is particularly suited to search for recurring patterns in a variety of cases. However, there is a limit to the variety that QCA can handle. When the cases are diverse to the point of being incomparable, QCA can no longer be applied. A QCA evaluation is particularly appropriate when the outcome strongly varies among cases. Our evaluation question: "under which conditions do the media trigger (non-) response from powerful actors?" was precisely based on this assumption. To answer the question, QCA systematically compares positive and negative effect cases. The more variety in the analysis, the more opportunity to see whether a certain condition contributes or is redundant to understand what works, which is a typical "learning" evaluation question.

A learning attitude toward less successful cases may be problematic or challenging in an evaluation context: Receiving parties may not always be very willing to share information about failure cases, and donors often want to know mostly about success stories. However, information about less successful cases is a prerequisite for understanding and analyzing sufficiency in QCA. In our evaluation, we emphasized the value of the less successful cases (i.e., media products that were not followed by any actor response) during the introductory workshop on the method.

In some evaluations, it is not possible to include cases that are too different from one another. QCA can only account for a limited number of causal conditions and therefore, it is recommended to choose most similar (not most different) cases to hold contextual conditions constant. As for our case, we decided already in an early stage to keep the analyses of Kenya and Tanzania separate. While the theory of change in the media program is similar in both countries, different choices were made in its implementation. To give a few examples: Tanzania's media program focuses on the rural areas and intends to increase the number of rural stories. The Kenyan funds emphases high-impact stories dealing with highly salient topics as impunity and leadership, accountability, corruption, politics and elections, and so on. So, to sum up, there should be variety in the causal conditions and outcomes, but contextual factors should be comparable.

Selecting conditions. Just as there can be multiple outcomes of interest for stakeholders, there can be multiple conditions to investigate. Yet, working with a too large number of conditions involves the risk of coming to individualized explanations per case and thus inhibiting parsimony. Being based on Boolean logic, the number of conditions has a strong influence on the number of logical combinations: $2^{\text{number of conditions}}$. With every condition (each scored as 1 or 0) added to the truth table, the number of logical combinations multiplies. Given this logic, QCA evaluators are advised to keep the number of conditions included as low as possible or at least to come to an adequate ratio between the number of conditions and the number of cases (Berg-Schlosser & De Meur, 2009; Marx & Dusa, 2011). In evaluation practices, this can be a challenging undertaking, with every individual stakeholder potentially having different preferences about conditions to examine. The choice of conditions should therefore be strongly theoretically informed or be based on prior (evaluation) evidence.

Logic modeling, a technique to model the decision logic within programs (Chen, 1992; Donaldson, 2012), may be a valuable tool to make conditions explicit but still this does not rule out the fact that there can be multiple relevant conditions. The QCA literature provides several tools to deal with this issue (see, for instance, Berg-Schlosser & De Meur, 2009, pp. 27–32): conditions can be aggregated in a higher order construct, remote conditions can be analyzed separately from more proximate conditions (so-called two step QCA; Mannewitz, 2011; Schneider & Wagemann, 2006), or other techniques as the most similar different outcome/most different similar outcome method can be used (De Meur, 1996; De Meur & Berg-Schlosser, 1994). The set of conditions can also change throughout the evaluation process, when they turn out not to be the right factors to account for a certain effect or when the analysis runs into many contradictions, that is, cases that share the same characteristics but that correspond with a different outcome. Going back and forth between the conditions and the analysis is a typical characteristic of QCA. This is also where the learning objective from the method is fulfilled. The iterative process is a useful vehicle to get to know the cases in more depth.

In our evaluation, a media product can be conceived as a combination of certain conditions that is assumed to contribute to an actor response. Actor response is the outcome in this regard. In particular, the selection of conditions was the result of three phases:[3]

    i.   Before the field visit, we requested the primary stakeholders to suggest conditions that were thought to be influential in explaining why some products lead to actors' response and others not. The theory of change underpinning the media programs and outlined in strategic plan of one of the funds (TMF, 2015) constituted the basis for selection. The initial set of conditions predominantly focused on characteristics of the journalist and a selection of characteristics of the media product (media echo, quality of the media product, and reach of media outlet).

    ii.   During the field visit, the initial list of conditions was systematically discussed with all stakeholders. The abovementioned QCA training helped the teams to identify new conditions or to remove other conditions.

    iii. A preliminary QCA analysis on available secondary data identified inconsistencies in the configurations, that is, the so-called contradictory configurations. Several media products shared the same characteristics but were associated with a different level of actor response. By analyzing these media products in depth, we learned that other conditions were key to understanding what really mattered. We learned that a sufficient explanation would never be found by only including conditions referring to the journalist and the abovementioned characteristics of the media product (media echo, quality of the media product, and reach of media outlet). The contradictions highlighted the need to include more conditions about the actual content of the media product (salience of an issue and regional focus of the story).

Table 2. List of Conditions.

Category	Conditions	Explanation
Characteristics of journalist	High journalistic experience	How many years is a journalist active in the profession?
	High education	What kind of training did a journalist attend?
Characteristics of media product	Strong media echo	Has other media picked up the same message?
	High reach of media outlet	Is the media outlet regional or national?
	High currency of issue	Does the story concern problems of ordinary people and/or of officials?
	High coverage of wrongdoings of people	Does the story deal with wrongdoings?
	High coverage of solutions	Does the story provide concrete solutions to the problem described that are easy to solve?
	Strong depth of story	Does the story describe the root causes to the problem described and significant past events?
	High salience of issue	Does the story concern a topic that is perceived as "hot" in the country?
	Regional focus of story	Does the journalist write about local or national issues?

This three-phase approach finally yielded a pool of 10 conditions that constituted the basis for data collection (see Table 2).

Stage 3: Data Collection

Once the evaluation design is settled, data should be collected to answer the evaluation questions. Evaluators can rely on new, primary data and/or on secondary, already available data. No matter whether data are of primary or secondary nature, QCA, as the term itself indicates, requires the data to be *comparative*. Especially, when relying on secondary sources, the comparability of the data can be a challenge, particularly when working in different intervention settings. Stakeholders may have worked with different concepts, may have collected data at different moments in time, or may have data of varying quality. When data are insufficiently comparable, QCA evaluators can consider clustering the cases in groups with comparable data, and analyze these separately. The minimal paths leading to the outcomes (i.e., the output of the QCA analyses) can then be compared at a later stage for the different groups of cases.

Our case evaluation was characterized by big differences in data comparability and data availability. The Kenyan and Tanzanian media programs developed separate and different monitoring and evaluation frameworks throughout their existence. Prior to our evaluation, the large majority of the Tanzanian media products were already subjected to rigorous and independent content analysis. A reliable database with information on different content attributes was available. In Kenya, such content analysis was not conducted, at least not in such a systematic way. Available data were exclusively based on the perception of journalists themselves and the staff members of the program. As for the outcomes, that is, the actor's response on the media products, the picture was more homogeneous, in the negative sense. There was no fine-grained information about the outcomes in either of the countries. Primary data collection was hence essential. For both countries, we launched a survey that was distributed to all journalists. This survey helped to acquire comparable information across the countries and complemented the secondary sources. In addition, we organized semistructured (skype and phone) interviews with a selection of journalists in both countries.

A common rule of thumb, albeit often forgotten, for each evaluation (QCA or not QCA) is to design the data collection strategy as early as possible. In a QCA evaluation, this is essential. Data need not only to be comparable but should also be calibrated at the stage of data analysis (see below). The need to calibrate the data, in a crisp or fuzzy set way, needs to be kept in mind already when deciding and constructing the data collection tools.

Stage 4: Data Analysis

Calibrating the data. fsQCA, the variant of QCA that we used (see above), requires the operationalization of conditions and outcomes in fuzzy set values. Depending on the condition or outcome concerned, a fuzzy set scale can, for instance, take the following format:

- Fully in the set (fuzzy membership = 1.0)
- More in, than out of the set (membership = .67)
- More out than in of the set (membership = .33)
- Fully out of the set (membership = 0)

All other thresholds between 0 and 1 are possible, though. Key in fuzzy set analysis is the anchor point: 0.5. Cases with a score below this anchor point are seen as cases that are more out than in of the set. Inversely, cases with a score above this anchor point are considered as cases that are more in than out of the set. It is without saying that the evaluator plays a very important role in the calibration procedure by justifying the thresholds. This justification should be based on substantive and/or theoretical knowledge. The ease within which the evaluator will be capable to calibrate will depend on (a) the extent in which a condition or outcome can be objectively measured, (b) the extent of heterogeneity of the cases, and (c) the extent of variability of a condition or outcome. We systematically discuss each of these challenges.

a. Objectivity versus Subjectivity. Conditions logically vary in the degree in which they can be objectively measured. The more objective, the less discussion among stakeholders. The more subjective, the more potential for conflict. Whether the story of a media product deals with a rural or urban area, for instance, does not require thorough interpretation. More difficulties arise when calibrating a condition as *"high journalistic experience."* When can one classify journalistic experience as high or low in a particular country setting? An external evaluator will in this respect again be highly dependent on the intervention stakeholders for input. They have the case knowledge to decide what it means to be *"highly or lowly educated,"* for instance. Often, stakeholders apply these thresholds implicitly in their daily operations, without making these explicit. For instance, applicants for media program funding have always been implicitly assessed on their journalistic experience, although the Kenyan and Tanzanian program officers did not have fixed criteria to determine what *"high experience"* exactly means. Labeling thresholds makes implicit assumptions explicit and hence also subject for discussion. In addition, calibration requires the temporary ignorance of nuances for the sake of systematic comparison. Stakeholders who are daily involved in an intervention may experience difficulties to ignore certain nuances. Evaluators will need to play a brokerage role in this respect. Ideally, the evaluator can himself calibrate available data for the sake of systematic comparability. In some circumstances, though, there will be no other way than to involve program beneficiaries or other stakeholders in the calibration of data. Within the scope of our evaluation, a survey was used to measure certain conditions. The four-scale answers of respondents were straightforwardly transformed in calibrated data. For instance, for the condition *"high coverage of wrongdoings of people"* journalists were asked to what extent wrongdoings of citizens or officials was the core focus of their product. It

would have been more objective to code the media products ourselves. Within the time, and budget frame of evaluations, pragmatism is yet often the compromise.

b. Heterogeneity Versus Homogeneity. The more heterogeneous a pool of cases, the more difficult it will be to find *"common"* calibration criteria that are relevant for all cases. In a diverse setting of cases, QCA evaluators will often need to resort to relatively abstract calibration categories. To give an example, our evaluation dealt with all kind of media products ranging from newspaper articles to radio programs and documentary series. The diversity of media outlets brought about tense discussions about *"at face value"* easy to calibrate conditions as *"reach of media outlet."* Given the diversity of products, we decided to proceed with generic calibration categories as national or local outlet. Would the evaluation only cover radio programs, more fine-grained and specific thresholds would have been possible.

c. Variability versus Stability. QCA is oriented toward diversity. Conditions and outcomes should vary sufficiently across cases, to make them interesting to include, and to avoid flawed inferences in the analysis of sufficiency and necessity (see Schneider & Wagemann, 2012, chapter 9 for a discussion on the implications of skewed set membership). This adagium should be taken into account when specifying the calibration thresholds. For a condition as *"education of the journalist,"* for instance, we initially considered to set the 0.5 threshold in between *"having a diploma"* (above the threshold) and *"having a certificate or having attended a workshop"* (below the threshold). In a setting as Kenya, where journalists are relatively highly educated, this turned out to be a suboptimal diversification, since almost all journalists scored above the 0.5 threshold, which would result in set membership that is highly skewed. As a result, we decided to move the 0.5 threshold up: Journalists with a master or bachelor degree were, respectively, given a score of 1 and 0.88, whereas journalists with a diploma a score of 0.44.

The minimization process. Once calibration thresholds are set, one can proceed to the very analytical moment of QCA, also known as the minimization process. We earlier hinted at the important role of "contradictions" in this respect. As Marx and Dusa (2011, p. 109) clearly state, "QCA is built on the assumption that contradictions will always occur if the explanatory model is not correctly specified (omitted variables, measurement error, heterogeneity of the research population, etc.) or when it does not make theoretical sense." Contradictions are hence common practice in a QCA evaluation and especially in situations where the number of cases is high in comparison with the number of conditions (Marx & Dusa, 2011). Why two or more media products that share the same characteristics nonetheless correspond with a different actor response is a puzzle that can only be solved with deep case knowledge. Especially, in an outsourced QCA evaluation, with an imbalance of case knowledge between evaluator and local stakeholders, the evaluator will not always be in a position to solve the contradictions himself or herself. Just as the selection of conditions needs to be conducted in consultation with the evaluation stakeholders, a change in this selection is preferably also verified by the stakeholders. Hence, also in the stage of a QCA data analysis, input from the stakeholders is usually needed. The process itself of contradiction solving can be considered as an output of a QCA exercise. Recall that QCA evaluations have in the first place a learning driven orientation. By solving contradictions, stakeholders get to know their cases more in depth. Although the preferred aim in QCA is to resolve all contradictions, especially when analyzing a large number of cases, and despite discussions with stakeholders, there is a possibility that not all contradictions can be resolved. In that situation, the evaluator can proceed using thresholds of quasi-sufficiency and as such move toward more probabilistic statements, such as 90% of the cases with the same combination of conditions share the same outcome (Hammersley & Cooper, 2012).

In any case, evaluators better anticipate on the possible occurrence of contradictions. A strategy is to plan the data collection stage so as to include information on an extensive set of conditions,

including many that might or might not be used. This is because when adding or replacing a condition turns out to be necessary, it is often difficult to organize a new data collection round. The more conditions are already measured in an initial round of data collection, the better. The survey that we launched in our case evaluation was deliberately conceived in a broad way. It comprised questions about all conditions that were remotely thought to be of potential influence on actor response on media products, even those we, as external evaluators, were skeptical about. The survey provided us with a large and wealthy pool of information that could be relied on if supplementary conditions would need to be added at a later stage in the analysis.

Stage 5: Interpreting the Findings

A next essential step, after the application of the software,[4] is the actual interpretation of the minimal formula(e). One should understand that QCA does not open the black box of causality itself (on this issue, see Goldthorpe, 1997). It does not directly nor fully explain the how or the process behind the combinations of conditions leading toward a particular outcome. How these conditions interact and how they link with the outcome is up to the evaluator's thick case interpretation to establish. Stakeholders should be aware of the specificities of the QCA output formulae, right from the beginning of the evaluation. This applies to (a) the complexity of the findings, (b) the causal status of the findings, (c) the atemporal character of the findings, and (d) their limited generalizability.

As Rihoux and Lobe (2009, p. 486) have formulated it: "(. . .) The QCA minimal formula act like a flashlight, which indicates some precise spots to be looked at to better understand the outcome." Consider, for instance, the following path associated with high response of powerful actors in the case of the Kenyan media products. The output is presented in the typical QCA format.

High education AND National media outlet AND Strong coverage of wrongdoings AND Strong coverage of background of problems

The formula draws our attention to the explanatory power of a combination of four conditions. An interview with one of the journalists who produced a media product that is representative for this path, helped us to better understand the QCA output. The journalist concerned broadcasted an investigative documentary about the audit of elections. The documentary received high response from politicians, both from the ruling majority and opposition. A press conference was held, and the documentary paved the way for a national dialogue about the elections. The story being nationally broadcasted on one of the biggest TV stations was a major explanatory factor for its success. Add here the fact that the journalist engaged in extensive investigative work. He considered it his role to unravel things that people did not know before. Being highly educated probably helps in this regard.

a. The type of (long) formula presented above is representative for the complexity of findings resulting out of a QCA analysis only (and usually the initial findings are even longer/more complex). Commissioners and other evaluation stakeholders may not be familiar with this. It is up to the evaluators, like we did, to translate the complex formulae in simpler terms or in recommendations, to the extent that this is desirable and possible. A suggestion to make the formulae more understandable can be to select a case per path revealed and describe how the path applies to that particular case. This makes it easier for evaluation stakeholders to understand how a certain path links to the outcome. No matter how the evaluator *"sells the story,"* we learned that it is beneficial to engage in strong expectation management prior to the start of the QCA project.

b. Related to this, it is important for stakeholders not to misinterpret the paths: Conditions are merely to be interpreted as associations with the outcome. High journalistic experience, for instance, turned out to be condition of strong empirical relevance associated with the absence of response of powerful actors in Tanzania. Why this is the case is a question that could not be simply answered on the basis of the QCA analysis. We needed stakeholder input to provide a meaning. For the local M&E team, this finding confirmed their intuitive feeling

that experience is not equal to journalistic talent. Many older journalists received their journalistic education in a repressed political system, which impacts their style of writing up until today. In addition, experienced journalists sometimes feel themselves too mature to be influenced by mentor advice, offered by TMF. Our QCA experience taught us that sufficient time for sense making should be foreseen in the evaluation process, unless the evaluators have a strong in-depth knowledge of all cases involved. In any case, to unravel the very causal mechanism behind the paths, the QCA analysis should best be complemented with methods like contribution analysis, realist evaluation or other theory-based approaches; albeit this might require an expansion of the evaluation (see Befani, 2016 on how these can be combined with QCA). In aiming to have a full understanding of the causal story, it is also important to conduct the necessity analysis and identify the necessary conditions. This also helps in simplifying QCA models and potentially obtain simpler solutions to the Boolean minimization sufficiency analysis (see Befani, 2016 for the different ways to synthesize the data set and different strategies to simplify QCA models).

c. We want to emphasize that the QCA analytical moment is in essence a static moment. Policy interventions, in contrast, always incorporate a dynamic component. The time dimension as such is not directly covered by the technique. When the time dimension is relevant, the evaluator needs to bring in his or her own case knowledge to interpret the sequence of conditions and their relationship vis-à-vis the outcome in temporal order. If necessary, the stakeholders can be consulted also at this stage. Several solutions have been laid out in the field of QCA applications to overcome the challenge of including the temporal dimension. An overview of different (sometimes very technical) scenarios is given in De Meur, Rihoux, and Yamasaki (2009, pp. 161–163), in Schneider and Wagemann (2012, pp. 263–274) and more recently in Fischer and Maggetti (2016). Possibilities include the addition of a condition that is explicitly constructed to discriminate time-related factors, or to include some quality of time in existing conditions, such as whether media echo quickly followed the production of a journal article (Fischer & Maggetti, 2016). Alternatively, one can consider the application of temporal QCA (TQCA; Caren & Panofski, 2005) where the evaluator can specify the sequence of causal conditions related to each case. Applications using the technique are still very rare, as TQCA is only applicable to csQCA and to studies with a limited number of conditions (Fischer & Maggetti, 2016). Another possibility still is to conduct separate QCA analyses for various points in time. When panel data are available, for instance, it is possible to examine each segment with a distinct analysis (Castro & Arino, 2013; Fischer & Maggetti, 2016).

d. The type of patterns should be considered as "contingent generalizations" (George & Bennett, 2004, p. 84). The level of external validity of QCA findings is usually relatively modest, in comparison with typical statistical interference. The actual extent of generalization of QCA solutions depends on the evaluators' choices about the extent of inclusion of logical remainders (i.e., the configurations not empirically observed) and on the size of the sample population. In our evaluation, we chose to proceed with the intermediate solutions, as we included logical remainders in our analysis that matched the analysis of necessity. Strictly speaking, the paths we found hold true for our area of analysis in the first place, that is, the media products investigated, and can be extrapolated toward contexts that share sufficient similar features (Berg-Schlosser et al., 2009, p. 12). Furthermore, while our evaluation targeted the entire population of media products that were produced with the help of KMP and TMF, the actual number of media products analyzed was lower and corresponded with the number of products about which we could obtain complete information. The latter can be considered a restriction in view of the external validity of our findings but is not unique to the QCA method. Again, expectation management in terms of external validity will be key before the QCA evaluation starts.

Table 3. Summary of Possible Challenges and Solutions in a QCA Evaluation.

Stage in Evaluation Process	Possible Challenges for QCA Evaluators	Lessons Learned
Decision to evaluate		
Determining purpose	Learning driven nature of QCA at odds with accountability purpose some evaluation stakeholders are more familiar with	Explaining QCA purpose to all stakeholders
Deciding on the locus	QCA requires technical knowledge and case-based knowledge	External evaluators should strongly interact with stakeholders during the entire evaluation process to compensate for their less "intimate" case knowledge
Establishing the evaluation design		
Formulating the evaluation questions	A QCA analysis requires each outcome to be investigated separately. But stakeholders may have multiple outcomes of interest	Consensus building among stakeholders to keep the outcomes of interest limited in number (via logic modeling, card sorting techniques, brainstorming, color voting, etc.)
Selecting cases	QCA requires case variability on conditions but preferably also on outcomes	Sensitizing stakeholders to share information about all cases or beneficiaries, no matter whether a certain intervention triggered success or failure
	QCA requires comparability of cases	Separate analysis for groups of cases that are not sufficiently comparable
Selecting conditions	QCA requires keeping the number of conditions limited. But stakeholders may have many conditions of interest	Consensus building among stakeholders to keep the number of conditions limited in number. The selection of conditions is ideally (program) theory informed
Data collection	QCA requires comparability of data	Investing in primary data collection and/or splitting the QCA analysis in groups of cases for which comparable data are available
	One shot data collection moment can be at odds with iterative character of QCA	Data collection should ideally be done in an extensive way, since extra data collection is not always possible within the frame of an evaluation process
Data analysis		
Calibrating the data	Ease of calibration depends on the extent to which a condition or outcome can be objectively measured, the extent of heterogeneity of the cases, and the extent of variability of a condition or outcome	• Making the implicit assumptions explicit • Deciding whether fine-grained or generic thresholds are most appropriate on the basis of case/theoretic knowledge • taking into account the (non-) variability of the data.
The minimization process	Contradictions can occur in QCA, that is, the same combination of conditions that correspond with a different outcome	• Input of stakeholders needed to solve contradictions • Conduct data collection in extensive way, so that data is available, if extra conditions need to be added to the analysis at a later stage.

(continued)

Table 3. (continued)

Stage in Evaluation Process	Possible Challenges for QCA Evaluators	Lessons Learned
Interpreting the findings	The QCA analysis does not open the black box of generative, mechanism-based causality Complex output of QCA can be challenging to understand	• Expectation management right from beginning of evaluation • Allocate time for sense making of QCA output, preferably in interaction with program stakeholders • If possible, the QCA evaluation should be complemented with methods able to unravel the causal mechanisms at play behind the paths

Note. QCA = qualitative comparative analysis.

Conclusion

The list of challenges in a QCA evaluation is long but probably not longer than for many other non-QCA-based evaluations. Table 3 summarizes all the abovementioned challenges and the corresponding solution(s). We emphasize that the list of challenges is based on our evaluation experience only. Yet, we believe that many of the challenges described are likely to occur in other QCA evaluation settings.

QCA is increasingly given attention as a new method for policy evaluations (Stern et al., 2012). The number of applications, though rapidly increasing, is yet still scarce. An application of QCA in an evaluation context needs to pay attention to several issues: We emphasize what we believe are some of the most important ones.

QCA has a *learning oriented perspective* and is only to a lesser extent accountability focused. The method therefore requires a constant and iterative dialogue with all primary stakeholders. This approach deviates from standard evaluation methods, where the external evaluator has a more distant position from stakeholders.

The method requires the *translation of cases* to a relatively few number of conditions and outcomes that are relevant and measurable across all cases. In an evaluation context, unlike in a research one, stakeholders have stakes in policy interventions and are more sensitive to the actions and choices of the investigator. For example, for some stakeholders, it can be challenging to disregard some of the unique features of specific interventions, even if they are not useful for the analysis or the systematic comparison of the cases.

Unlike experimental methods, QCA is *geared toward variation*. Its basic "mandate" (for the sufficiency analysis) is to explain why some cases show the presence of the outcome and other cases show the absence outcome. When translating this in an evaluation, this implies that the assumption that some interventions have led to good results, and other to suboptimal results, is preliminary and must be made from the start, setting the tone for the rest of the evaluation. Diversity is also required for the conditions. Only conditions that vary sufficiently across cases can be considered in the analysis, otherwise the conclusions will be obvious and will not need any analysis in the form of a Boolean minimization. We highly recommend a careful, small selection of those conditions with the highest explanatory power.

The challenges addressed in this article should not discourage evaluators. We refer to many works that emphasize the potential of the approach (and associated technique) for evaluation. The contribution of this article is to show that testing QCA in real-life evaluations brings lessons learned

that are relevant or specific to evaluation processes and that should be discussed and disseminated if QCA is to be used more widely in evaluations. We deliberately decided to focus on the challenges because evaluators need a more practical understanding of the negative scenarios that can potentially arise when applying this method in their daily job. The good news is that, on the basis of our experience, these challenges can be overcome, and we illustrated how we managed to do it in the evaluation used as a case example.

We realize that certain solutions may bring new challenges, and that every evaluation is to some extent different. There is no "one size fits all solution" that would apply to all QCA evaluations. The lessons learned are, however, not specific to development cooperation cases only but have relevance for evaluations in all policy fields in principle. The challenges we describe will be most pronounced in situations where evaluation stakeholders never conducted any QCA evaluations before, and/or where stakeholders are socialized in an accountability oriented evaluation culture, rather than in a learning-focused environment. Given the relative novelty of this approach in evaluation and the need to discuss what evaluators should expect, we hope that our experience will shed some light on broadly relevant issues and ideally contribute to a wider debate (see also Befani, 2016; Baptist & Befani, 2015) on the benefits and challenges of practically applying QCA to real-life evaluations.

Appendix

Table A1. Key Information About the Analysis of Necessity and Sufficiency for the Case Example.

Analysis Findings		
Necessary conditions: threshold: 0.9 (cf. best practice in QCA research; Schneider and Wagemann, 2012)	Comparing the analyses of necessity for the absence and presence	Some conditions are necessary for either the presence or absence of the outcome but are necessary for both to a large extent. We consider these as *trivial necessary conditions*
		Some conditions are necessary for either the presence or absence of the outcome. We consider these as *relevant necessary conditions*
Sufficient conditions: threshold: 0.8 (cf. best practice in QCA research; Schneider and Wagemann, 2012)	fsQCA[a] analysis, truth table analysis with Quine–McCluskey algorithm	We revealed several paths per outcome (intermediate solution-assumptions based on analyses of necessity). We visualized paths by distinguishing between core causal and peripheral conditions (cf. Fiss, 2011). In the evaluation report, we narratively discussed the intermediate solution

<table>
<tr><td></td><td></td><td colspan="2" align="center">Kenya</td><td colspan="2" align="center">Tanzania</td></tr>
<tr><td></td><td></td><td colspan="4" align="center">Response of powerful actors</td></tr>
<tr><td></td><td></td><td>Presence: 3 paths</td><td>Absence: 5 paths</td><td>Presence: 8 paths</td><td>Absence: 11 paths</td></tr>
<tr><td></td><td></td><td colspan="4" align="center">Response of citizens[b]</td></tr>
<tr><td></td><td></td><td>Presence: 5 paths</td><td>Absence: 6 paths</td><td>Presence: 3 paths</td><td>Absence: 2 paths</td></tr>
</table>

Note. Author's own compilation. fsQCA = fuzzy set QCA; QCA = qualitative comparative analysis.
[a]Ragin and Davey (2009). [b]The outcome "citizen response" was only added at a later stage in the evaluation, at a moment when the conditions for the outcome "response of powerful actors" were already selected. It was decided to proceed with the same conditions for the two outcomes. However, the analysis confirmed that the explanatory model for citizen response was suboptimal (i.e., the solution coverage was very low).

Notes

1. The qualitative comparative analysis (QCA) software assists the researcher in making these assumptions on the plausibility of the logical remainders. Based on this input, the software automatically generates three types of solutions: complex (called conservative solution by Schneider & Wagemann, 2012, p. 162), intermediate, and parsimonious. The intermediate solution is always a subset of the most parsimonious solution, and a superset of the most conservative/complex solution that does not consider any remainders (Schneider & Wagemann, 2012, p. 164). Importantly, logical remainders can differ in theoretical and empirical plausibility (Ragin & Sonnett, 2004). In our evaluation, we applied a relatively restrictive stance by barring all counterfactuals that conflicted with the necessary conditions found for each of the outcomes. In other words, we proceeded with the so-called intermediate solutions, which correspond with a medium level of inference (Ragin & Sonnett, 2004).
2. The stage of "communicating the findings" is beyond the scope of this article. Discussing different presentation formulae of QCA findings would require a lengthy paper in its own right.
3. This three-phases approach should not be confused with the procedure and logic underpinning two-step QCA (Mannewitz, 2011).
4. In the Appendix, we give a summary table (Table A1) with technical details about the analyses of necessity and sufficiency, and the number of paths revealed for the outcomes investigated in our case example.

References

Balthasar, A. (2009). Institutional design and utilization of evaluation: A contribution to a theory of evaluation influence based on Swiss experience. *Evaluation Review*, *33*, 226–256.

Baptist, C., & Befani, B. (2015). Qualitative comparative analysis: A rigorous qualitative method for assessing impact *Coffey "How To" Note*Coffey. Retrieved June 26, 2017, from http://www.coffey.com/assets/Ingenuity/Qualitative-Comparative-Analysis-June-2015.pdf

Baumgartner, M. (2012). Detecting causal chains in small-n data. *Field Methods*, *25*, 3–24.

Beach, D., & Pedersen, R. B. (2013). *Process-tracing methods: Foundations and guidelines*. Ann Arbor: University of Michigan Press.

Befani, B. (2013). *Multiple pathways to policy impact: Testing an uptake theory with QCA*. Centre for Development Impact Practice Paper 5. Brighton, UK: Centre for Development Impact. Retrieved June 26, 2017, from http://www.ids.ac.uk/publications/ids-series-titles/cdi-practice-papers

Befani, B. (2016). *Pathways to change: Evaluating development interventions with QCA*, Rapport till Expert-gruppen för biståndsanalys [Report for the Expert Group for Aid Studies-EBA], Expert Group for Aid Studies. Report 05/16. Stockholm, Sweden: EBA. Retrieved from http://eba.se/en/pathways-to-change-evaluating-development-interventions-with-qualitative-comparative-analysis-qca/#sthash.nyGxIej9.dpbs.

Berg-Schlosser, D., & De Meur, G. (2009). Comparative research design: Case and variable selection. In B. Rihoux & C. C. Ragin (Eds.), *Configurational comparative methods. Qualitative comparative analysis (QCA) and related techniques* (pp. 19–32). Thousand Oaks, CA: Sage.

Berg-Schlosser, D., De Meur, G., Rihoux, B., & Ragin, C. C. (2009). Qualitative comparative analysis as an approach. In B. Rihoux & C. C. Ragin (Eds.), *Configurational comparative methods. Qualitative comparative analysis (QCA) and related techniques* (pp. 1–18). Thousand Oaks, CA: Sage.

Caren, N., & Panofsky, A. (2005). TQCA A technique for adding temporality to qualitative comparative analysis. *Sociological Methods & Research*, *34*, 147–172.

Castro, R. G., & Arino, M. (2013). *A general approach to panel data set-theoretic research* (COMPASSS Working Paper Series, 76, 2013). COMPASSS. Retrieved June 26, 2017, from http://www.compasss.org/wpseries/GarciaCastroArino2013.pdf

Chen, H. T. (1992). *Using theory to improve program and policy evaluations* (No. 290). New York, NY: Greenwood Press.

Davies, R. (1996). Hierarchical card sorting: A tool for qualitative research. *Monitoring and Evaluation NEWS*. Retrieved from http://www.mande.co.uk/docs/hierarch.htm

De Meur, G. (1996). La comparaison des systèmes politiques: Recherche des similarities et des differences [The comparison of political systems: Establishing similarities and differences]. *Revue Internationale de Politique Comparée*, *3*, 405–437.

De Meur, G., & Berg-Schlosser, D. (1994). Comparing political systems: Establishing similarities and dissimilarities. *European Journal of Political Research*, *26*, 193–219.

De Meur, G., Rihoux, B., & Yamasaki, S. (2009). Addressing the critiques of QCA. In B. Rihoux & C. C. Ragin (Eds.), *Configurational comparative methods. Qualitative comparative analysis (QCA) and related techniques* (pp. 147–178). Thousand Oaks, CA: Sage.

Donaldson, S. I. (2012). *Program theory-driven evaluation science: Strategies and applications*. New York, NY: Routledge.

Fischer, M., & Maggetti, M. (2016). Qualitative comparative analysis and the study of policy processes. *Journal of Comparative Policy Analysis: Research and Practice*, 1–17. doi:10.1080/13876988.2016.1149281

Fiss, P. C. (2011). Building better causal theories: A fuzzy-set approach to typologies in organization research. *The Academy of Management Journal*, *54*, 393–420.

George, A. L., & Bennett, A. (2004). *Case studies and theory development in the social sciences*. Cambridge, MA: MIT Press.

Goldthorpe, J. H. (1997). Current issues in comparative macrosociology: A debate on methodological issues. *Comparative Social Research*, *16*, 1–26.

Hammersley, M., & Cooper, B. (2012). Analytic induction versus qualitative comparative analysis. In B. Cooper, J. Glaesser, R. Gomm, & M. Hammersley (Eds.), *Challenging the qualitative-quantitative divide: Explorations in case-focused causal analysis* (pp. 129–169). London, England: Bloomsbury.

Mahoney, J., & Goertz, G. (2006). A tale of two cultures: Contrasting quantitative and qualitative research. *Political Analysis*, *14*, 227–249.

Mannewitz, T. (2011). *Two-level theories in QCA: A discussion of Schneider and Wagemann's two-step approach* (COMPASSS Working Paper, 64). COMPASSS. Retrieved June 26, 2017, from http://www.compasss.org/wpseries/Mannewitz2011.pdf

Marx, A., & Dusa, A. (2011). Crisp-set qualitative comparative analysis (csQCA): Contradictions and consistency benchmarks for model specification. *Methodological Innovations Online*, *6*, 103–148.

Pattyn, V., & Brans, M. (2013). Outsource versus In-house? An identification of organisational conditions influencing the choice for internal or external evaluators. *The Canadian Journal of Program Evaluation*, *28*, 43–63.

Ragin, C. C. (1987). *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkeley: University of California Press.

Ragin, C. C. (2000). *Fuzzy set social science*. Chicago, IL: University Chicago Press.

Ragin, C. C. (2008). *Redesigning social inquiry: Fuzzy sets and beyond*. Chicago, IL: University Chicago Press.

Ragin, C. C. (2009). Qualitative comparative analysis using fuzzy-sets (fsQCA). In B. Rihoux & C.C. Ragin (Eds.), *Configurational comparative methods. Qualitative comparative analysis (QCA) and related techniques* (pp. 87–121). Thousand Oaks, CA: Sage.

Ragin, C. C., & Davey, S. (2009). fs/QCA (Version 2.5) [Computer Programme]. Tucson: University of Arizona.

Ragin, C. C., & Sonnett, J. (2004). Between complexity and parsimony: Limited diversity, Counterfactual cases and comparative analysis. In S. Kropp & M. Minkenberg (Eds.), *Vergleichen in der Politikwissenschaft* [The comparative method in political science] (pp. 187–197). Wiesbaden, Germany: VS Verlag fur Sozialwissenschaften.

Rihoux, B., & De Meur, G. (2009). Crisp-set qualitative comparative analysis (csQCA). In B. Rihoux & C. C. Ragin (Eds.), *Configurational comparative methods. Qualitative comparative analysis (QCA) and related techniques* (pp. 33–68). Thousand Oaks, CA: Sage.

Rihoux, B., & Lobe, B. (2009). The case for qualitative comparative analysis (QCA): Adding leverage for thick cross-case comparison. In D. Byrne & C. C. Ragin (Eds.), *The Sage handbook of case-based methods* (pp. 222–243). London, England: Sage.

Rihoux, B., & Ragin, C. (2009). *Configurational comparative methods. Qualitative comparative analysis (QCA) and related techniques*. Thousand Oaks, CA: Sage.

Rihoux, B., Rezsöhazy, I., & Bol, D. (2011). Qualitative comparative analysis (QCA) in public policy analysis: An extensive review. *German Policy Studies*, *7*, 9.

Schneider, C. Q., & Wagemann, C. (2006). Reducing complexity in Qualitative Comparative Analysis (QCA): Remote and proximate factors and the consolidation of democracy. *European Journal of Political Research*, *45*, 751–786.

Schneider, C. Q., & Wagemann, C. (2012). *Set-theoretic methods for the social sciences. A guide to qualitative comparative analysis*. Cambridge, MA: Cambridge University Press.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012) Broadening the range of designs and methods for impact evaluations (Department for International Development (DFID) Working Paper 38). London, UK: DFID.

Tanzanian Media Fund. (2015). *Tanzania Media Fund Strategic Plan. June 2012–May 2015*. Dar es Salaam, Tanzania: Hivos.

Vedung, E. (2009). *Public policy and program evaluation*. New Brunswick, NJ: Transaction.

Verweij, S., & Gerrits, L. M. (2013). Understanding and researching complexity with qualitative comparative analysis: Evaluating transportation infrastructure projects. *Evaluation*, *19*, 40–55.

Vink, M. P., & Van Vliet, O. (2009). Not quite crisp, not yet fuzzy? Assessing the potentials and pitfalls of multi-value QCA. *Field Methods*, *21*, 265–289.

Vis, B. (2012). The comparative advantages of fsQCA and regression analysis for moderately large-N analysis. *Sociological Methods and Research*, *41*, 168–198.