Working Paper 2017-88

Addressing the Challenges in Using Qualitative Data in Qualitative Comparative Analysis

Debora de Block
Water Systems and Global Change Group
Wageningen University and Research

Barbara Vis
Department of Political Science and Public Administration
Vrije Universiteit Amsterdam

Abstract: This paper addresses an issue that so far has received relatively little attention in methodological discussions about Qualitative Comparative Analysis (QCA), namely the challenges researchers face when using qualitative data in QCA analyses. Building on a literature review of 22 empirical studies using qualitative data for QCA, we discuss the challenges and ways to address these: (1) Which qualitative data to use? (2) How to determine the thresholds for in- and exclusion of a set? (3) How to establish the degree to which a case is 'in' or 'out' of a set? (4) How to differentiate between those concepts that are truly absent and those that are not mentioned? (5) Which sensitivity tests to conduct for assessing the robustness of the findings based on qualitative data? (6) How to present the calibration process transparently (and concisely)? By discussing these challenges using especially current practices in QCA-studies that are mostly informed by qualitative data, we aim to contribute to the best practices in QCA research.

Keywords: qualitative comparative analysis; qualitative data; calibration

---

COMPASSS Working Paper 2017-88
Available at http://www.compasss.org/wpseries/deBlockVis2017.pdf
COMPASSS Working Paper Series Managing Editor, Claude Rubinson

Addressing the challenges in using qualitative data in Qualitative Comparative Analysis

Debora de Block[a] & Barbara Vis[b]

[a] Water Systems and Global Change group, Wageningen University and Research,
email: debora.deblock@wur.nl.

[b] *Corresponding author:* Department of Political Science and Public Administration, Vrije Universiteit Amsterdam, email: b.vis@vu.nl, web: www.barbaravis.nl / www.highriskpolitics.org.

Abstract

This paper addresses an issue that so far has received relatively little attention in methodological discussions about Qualitative Comparative Analysis (QCA), namely the challenges researchers face when using qualitative data in QCA analyses. Building on a literature review of 22 empirical studies using qualitative data for QCA, we discuss the challenges and ways to address these: (1) Which qualitative data to use? (2) How to determine the thresholds for in- and exclusion of a set? (3) How to establish the degree to which a case is 'in' or 'out' of a set? (4) How to differentiate between those concepts that are truly absent and those that are not mentioned? (5) Which sensitivity tests to conduct for assessing the robustness of the findings based on qualitative data? (6) How to present the calibration process transparently (and concisely)? By discussing these challenges using especially current practices in QCA-

studies that are mostly informed by qualitative data, we aim to contribute to the best practices in QCA research.

Keywords

Qualitative Comparative Analysis; Qualitative data; Calibration

1. Introduction[1]

Almost 30 years after Charles Ragin (1987) introduced the approach in the social sciences, *Qualitative Comparative Analysis (QCA)* is starting to turn into a "mainstream" approach in several disciplines, such as sociology and political science (Rihoux et al. 2013). Despite this trend towards becoming a mainstream approach, QCA is also still in development. Current methodological discussions focus, for example, on the (in)compatibility of regression analysis and QCA (Fiss et al. 2013; Thiem et al. 2016; Vis 2012), the preferred type of solution: intermediate or parsimonious (e.g., Baumgartner, 2015; Schneider & Wagemann, 2012), and sensitivity diagnostics, robustness analyses or model specification (e.g., Thiem et al. 2016; Skaaning 2011; Marx 2010; Baumgartner & Thiem 2015; Thiem 2014).[2]

In this paper, we focus on an issue that so far has received relatively little attention in methodological discussions about QCA: the challenges researchers face when using qualitative data in QCA and how to address these. Given that QCA was originally conceived as an approach for case-oriented comparative research (Ragin, 2013: 171), which often

1 The ideas in this paper have been presented at the 4th International QCA Expert Workshop in Zurich, Switzerland in 2016. We thank all participants for their useful comments and suggestions. Additionally, we thank the two anonymous reviewers of the COMPASSS Working Paper Series, the series' editor – Claude Rubinson –, and Federico Iannacci, Eva Thomann and Zsofia Tóth for their constructive feedback on earlier version of this paper. Barbara Vis' research is funded by a VIDI grant from the Netherlands Organisation for Scientific Research (grant nr.: 452-11-005).

2 Additionally, there are studies criticizing QCA as an approach (e.g., Lucas & Szatrowski, 2014; Paine, 2016). Such studies, however, have received their fair share of criticism themselves (e.g., Fiss, Marx, & Rihoux, 2014; Ragin, 2014; Thiem & Baumgartner, 2016).

uses qualitative data, this may seem like a redundant question to ask. In fact and interestingly, however, it is not. For instance, while teaching QCA to graduate students, we received many questions about using qualitative data. Questions included for instance: Can you use qualitative data in QCA? (y*es*) Is it meaningful to do so? (*that depends*, among other factors on the aims of the study, an issue we do not address here, but see e.g. Schneider & Wagemann 2012). How to go about collecting, calibrating – i.e., interpreting 'measures relative to external standards' (Ragin, 2013: 172) – and analysing qualitative data for QCA? (*more on that to follow*). Remarkable little guidance exists on how to answer these and related questions (cf. Scheck McAlearney et al. 2016, but see Basurto & Speer 2012 and Tóth et al. 2016). Existing studies using qualitative data in QCA are typically unclear about how they have calibrated their data (cf. Basurto & Speer 2012: 169; Tóth et al. 2016). At the same time, techniques common in other methodological approaches – such as coding interviewees' answers according to an "arbitrary", predetermined scale that is subsequently quantified – cannot be used for set calibration since fuzzy-set values should be based on theoretical and substantive (case) knowledge (cf. Basurto & Speer 2012).

With this paper, we aim to contribute to the best practices in QCA research by presenting a set of directions on how to handle qualitative data in QCA. These directions are based on a review of 22 QCA-studies that use various types of qualitative data. Hereby, we also contribute to the discussion on how to conduct multi-case qualitative research. We discuss how the 22 studies have, or have not, addressed the challenges

that arise for the following six questions: (1) Which qualitative data to use? (2) How to determine the thresholds for in- and exclusion of a set? (3) How to establish the degree to which a case is “in” or “out” of a set? (4) How to differentiate between those concepts that are truly absent and those that are not mentioned in interviews? (5) Which sensitivity analyses to conduct for assessing the robustness of the findings based on qualitative data? And (6) how to present the calibration procedure transparently (and concisely)?

Note that while the examples we use come mainly from studies using medium-sized data sets with say between 10 and 50 cases, our directions also hold for studies using qualitative data that have a higher number of cases. The directions are intended both for crisp set QCA, in which cases’ membership in sets is either “fully in” (1) or “fully out” (0), and for fuzzy-set QCA, in which cases’ membership in the sets varies between “fully in” (1) or “fully out” (0) (e.g., a four-value fuzzy-set: 1, .75, .25 and 0; or a continuous fuzzy set with values from 1 through 0).

This paper is structured as follows. First, we discuss the two existing studies focusing on how to use qualitative data in QCA – Basurto and Speer (2012) and Tóth et al. (2016) –, whereby we highlight their contributions and identify the issues they leave unaddressed. Next, we explain how we selected the literature for our review of QCA-applications using qualitative data. Subsequently, we discuss these studies following the six questions introduced above. We finalize the paper with a reflection on how our results contribute to best practices in QCA.

2. How to use qualitative data in QCA? Two existing approaches

The two existing studies on how to use qualitative data in QCA – Basurto and Speer (2012) and Tóth et al. (2016) – focus especially on the process of calibration. This focus is understandable, since calibration is a key step in any QCA. How to calibrate *quantitative* data is an issue various QCA-scholars addressed. For example, Ragin's (2008) chapter on the calibration of fuzzy sets is devoted exclusively to this – admittedly important – issue. Conversely, Ragin provides no practical advice for researchers on how to perform the task of calibrating *qualitative* data. In their text book on set-theoretic methods, also Schneider and Wagemann (2012: 32-41) offer practical advice on how to calibrate only for quantitative data. Moreover, Schneider and Wagemann's discussion on the calibration process in general – so including that of qualitative data – remains rather abstract. They indicate, for instance, that this process needs to be *transparent* and that the resulting 'set [needs to have] *high content validity* for the concept of interest' (Schneider & Wagemann 2012: 32, emphases added). The latter is, indeed, advice that should be followed. But how to do this? Schneider and Wagemann (2012: 24) state that researchers here need to *make use of knowledge that is external to the data*, for instance coming from "obvious facts", things that are generally accepted in the social sciences, or 'the knowledge of the researcher accumulated in a specific field of study or specific cases'. This requires much work from a researcher, who could use for instance data from 'interviews, questionnaires, data obtained with participant

observation or focus groups, and organizational analysis, quantitative and qualitative content analysis (…)' (Schneider & Wagemann, 2012: 32). All this is very much true. But it is also difficult to do and the advice on what to do is hardly concrete. Consequently, researchers who are new to the process of calibration probably do not know how to go about doing this.

To the best of our knowledge, Basurto and Speer (2012) and Tóth et al. (2016) are the only two studies focusing explicitly on how to use *qualitative* data in QCA. Basurto and Speer propose a stepwise procedure to calibrate qualitative (interview) data to qualitative classifications with associated fuzzy-set values, which we summarized in Appendix I. In a nutshell, the six steps of their procedure are the following: (1) operationalize the conditions and the outcome; (2) develop the qualitative thresholds or anchor points and elaborate the qualitative interview guideline; (3) conduct a content analysis of the raw interview data; (4) summarize the coded qualitative data; (5) determine the fuzzy-set scale and define the fuzzy-set values; and (6) assign and revise the fuzzy-set values.

The second study that provides helpful suggestions on how to use qualitative data in QCA is that of Tóth et al. (2016). Tóth et al. aim to increase the validity and replicability of the calibration of qualitative data and to this end introduce the so-called Generic Membership Evaluation Template (GMET). We summarized the steps of the GMET in Appendix II. This GMET is to be used for each case for each condition or outcome (e.g., with 10 cases, 3 conditions and 1 outcome, there are 40 GMETs to be filled in). Again in a nutshell, the six steps of Tóth et al.'s approach

are: (1) provide an overall case description from the perspective of the specific condition; (2) list the condition's dimensions or sub-measures (what we label attributes);[3] (3) for each of these attributes, provide information on: the specific context, the direction of the effect on membership (positive or negative), the attribute's relative intensity or importance (high, moderate or low), and an illustrative quote; (4) provide supporting quantitative data, if applicable; (5) provide set membership score; and (6) summarize the argumentation for giving this set membership score. Hereby, this approach enables theory-informed, transparent decision making on assigning membership scores to conditions based on qualitative data.[4]

Calibration is an important analytical step in a QCA analysis, and also the majority of challenges we encountered relate to this process (see

3 We follow Goertz and Mahoney's (2012) terminology for qualitative research, which means that we use the terms concepts, attributes and data instead of variables and indicators (and sub-measures).

4 Note that Tóth et al. present the GMET as an alternative to what they see as two core calibration strategies currently used in QCA-studies: (1) dichotomisation of data and (2) using quantitative data to inform the process of calibration (what they call: setting quantitative anchor points). Tóth et al. consider the second strategy an appropriate one in some cases, but indicate that especially in the context of qualitative research, the meaning of the same quantitative score may differ across cases. In the latter case, using quantitative anchor points will fail to result in valid membership scores. Their GMET thus allows for quantitative data to inform the calibration process, but it does neither require this nor is it the crucial element in the procedure. Regarding dichotomizing data as a calibration strategy, Tóth et al. (2016: 4) state that the majority of QCA-studies using qualitative data opt for crisp sets over fuzzy ones because calibrating qualitative data in fuzzy sets is challenging. We agree that it is, but we beg to differ that dichotomizing is an easy way out. Also when dichotomizing data, researchers need to establish when a case is "in" or "out" of a set, because of which also they face the intricate question how to do that (see Li et al. 2016 for an example). This is why we focus both on studies using fuzzy sets and crisp sets.

section 5). Still, when using qualitative data in QCA, researchers face other challenges, such as which data to use, how to analyse them, how to present the calibration process, and which sensitivity analyses to conduct. On such issues, Basurto and Speer (2012) and Tóth et al. (2016) hardly provide guidance.[5]

Additionally, when following one of their approaches, researchers are likely to face concrete challenges that are not addressed in these studies. For example, how to follow Basurto and Speer's (2012) recommendation to take some qualitative aspects into consideration (cf. Tóth 2016: 5)? We assume that researchers using Basurto and Speer's approach have indeed grappled with this, since most of them fail to make different calibration steps and motivations behind the various choices explicit (Chai & Schoon 2016; Chatterley et al. 2014; Chatterley et al. 2013). Chai and Schoon (2016: 32), for example, only state that they have used the procedure of Basurto and Speer (2012) and do not discuss what this entailed. Moreover, in both the approach of Basurto and Speer and that of Tóth et al., it remains unclear how the qualitative data inform and justify the determination of the qualitative breakpoints, especially the cross-over point with value 0.5. And whereas filling in the GMET is rather straightforward, the final decision on how to attribute the final fuzzy set score remains to some extent subjective.[6]

5 Note that Basurto and Speer (2012: 165) do argue that 'decisions on contradictions in the data and the information based on which they were made need to be transparent in presenting the analysis results'.

6 Tóth et al. (2016: 6) acknowledge that there is a subjective element to using the GMET, but indicate that nonetheless it 'provides a means to ensure transparency of the calibration by making the researcher's judgment (…) visible'.

Summing up, Basurto and Speer's (2012) six-step approach and Tóth et al.'s (2016) GMET provide valuable guidelines for researchers on how to calibrate qualitative data. The aim of our paper is therefore not to propose a whole new approach, but rather to provide directions on how to handle qualitative data in QCA, focusing especially on the issues they leave unaddressed. Hereby, we thus take up the invitation of Basurto and Speer (2012: 170) for 'an open discussion on how to improve the use of qualitative interview data in QCA'.

3. Selection of QCA-studies using qualitative data

To demonstrate how QCA-researchers use qualitative data in their analyses, we reviewed existing empirical studies using such data. To this end, we first selected peer-reviewed articles from the bibliography on the COMPASSS website (www.compasss.org, last accessed on November 10, 2016). We examined the papers' titles and abstracts. When we considered these to be relevant, we read the methods section to see whether it contained information about the use of qualitative data. This search process led to the selection of 3 papers. Additionally, we used Scopus to find papers that referenced Basurto and Speer (2012) (n=10, accessed on October 20, 2016) and selected 4 relevant papers, again based on assessing the methods sections. A similar search on ISI Web of Science yielded no additional papers. We further determined the relevance of the 7 papers discussed by Tóth et al. (2016), leading to the selection of 3 additional papers for our analysis. Finally, we derived 12 papers through snowballing, i.e. we identified relevant papers based on references in

already selected papers. We excluded papers that did not include an empirical analysis (such as Scheck McAlearney et al.'s 2016 study on the potential of QCA in health research), or of which the type of data used was not clearly qualitative (e.g., Ide's 2015 analysis based on existing data and advise from [but not "true" interviews with] experts). This search process resulted in the 22 papers included here. The studies cover a broad range of topics and geographical locations, from biodiversity conservation in Costa Rica (Basurto 2013) to school sanitation in Bangladesh (Chatterley et al. 2014) and from policy change in Switzerland (Fischer 2014) to environmental justice policies in the US (Kim & Verweij 2016). Table A1 in Appendix III gives an overview of all selected papers and provides information on how each study addressed the six questions. We discuss these questions in more detail in the sections below.

4. Data Sources

The first question we looked into, is which type or types of data were used in the 22 studies we reviewed. Table 1 summarizes the information on this from table A1 in Appendix III.

Table 1: Summary of the types of qualitative data used

Types of (qualitative) data	Examples
Interviews	Basurto 2013; Basurto & Speer 2012; Chai & Schoon 2016; Chatterley et al. 2013; 2014; Crilly 2011; Fischer 2014; *Henik 2015*; Kirchherr et al. 2016; Li et al. 2016; *Metelits 2009; Tóth et al. 2016*; Van der Heijden 2015; Vergne & Depeyre 2016; Verweij & Gerrits 2015; Verweij et al. 2013; Wang 2016

Existing documents/ archive material	Basurto 2013; Basurto & Speer 2012; Chai & Schoon 2016; Crilly 2011; Fischer 2014; Kim & Verweij 2016; Kirchherr et al. 2016; Li et al. 2016; Van der Heijden 2015; Vergne & Depeyre 2016; Verweij 2015; Verweij & Gerrits 2015; Verweij et al. 2013
Data from observations (e.g. photos, site visits)	Chatterley et al. 2013; 2014; Verweij & Gerrits 2015; Wang 2016
Focus groups	Chatterley et al. 2013; 2014
Participant observation	Verweij & Gerrits 2015; Verweij et al. 2013

Notes: Details are provided in Table A1 in Appendix III. Studies listed in italics use only one type of data.

Table 1 indicates that the large majority of these studies used more than one source of qualitative data (Henik 2015; Metelits 2009; Tóth et al. 2016 are exceptions with only one source of data). Kim and Verweij (2016), for example, used reports to determine the initial adoption level of environmental justice policies, which they then corroborated with the results of surveys conducted by other researchers. Tóth et al. (2016) included a row called “supporting quantitative data” in their GMET-template to allow for the inclusion of quantitative data that can inform the process of qualitative data calibration. In their own empirical analysis, they did not use such quantitative data though. Peer-reviewed papers, press research, interviews and survey data were collected and triangulated by Kirchherr et al. (2016) to enhance the validity of their findings. Given that using multiple data sources allows for the triangulation of data, we consider the use of multiple sources of data a best practice. Still, the studies that did use more than one type of data typically did not discuss how the data were triangulated. They also did not discuss what had been done in those situations where different types of data conflicted.

When statements or different data sources contradict one another, a decision must be made on which data to base the final value. This decision needs to be presented in a transparent way. A good justification for the choice made is especially important when one type of data imply that a case is “in” the set, while another type of data indicate the case is “out” of the set. The reason is that in those situations, the decision made will likely influence the substantive findings. This is an issue that requires more attention.

In almost all of the 22 studies, interviews were among or the only source of data (Verweij 2015 and Kim & Verweij 2016 are exceptions here in not using interview data). A challenge that is especially relevant for collecting interview data, but also applies when analysing for instance existing documents, is how to define the outcome and conditions such that they are interpreted in the same way by different researchers. We recommend researchers to try to determine, if this is possible, in an early stage of the research process whether a concept is multi-interpretable or not. If it is, it should either be defined more strictly, so as to make it unambiguous, or the concept should be sub-divided into various attributes (in our case in “facilitating policies and regulations” and “restricting policies and regulations”). A second best option is going back to the data to get a more detailed understanding of a specific condition.

5. Calibration

As we indicated earlier, assigning set membership to the outcome and conditions (i.e. calibration) is usually challenging for researchers working

with qualitative data. Existing literature typically does not provide guidance on: (1) where to place the thresholds for in- and exclusion of a set and (2) how to establish the degree to which a case is "in" ($0.5 < x \leq 1$) or "out" of the set ($0 \leq x < 0.5$) – question 2 and 3 we focus on here. These two questions are related, since oftentimes answering question 3 on the degree of membership will also require an answer to question 2 on the in- and exclusion of a set. Below and in Table A1 in Appendix III, we have separated them as much as possible for pedagogical purposes.

How to determine the thresholds for in- and exclusion of a set?

To determine where to place the thresholds for in- and exclusion of the set, researchers have various options at their disposal. Table 2 summarizes different strategies; Table A1 in Appendix III provides more details.

For example, Chatterley et al. (2014) develop a rubric to assign codes for the outcome and the conditions based on data from interviews, focus group and observations. Whereas these codes are useful to rate the conditions and outcome for each case, Chatterley et al. (2014) do not provide a justification for assigning the thresholds for membership and non-membership. Another example is Kirchherr et al. (2016), who use both a 4-value and 2-value scheme to assign fuzzy-set values, depending on the depth of information available for each case and the nature of the condition. The operationalization of some fuzzy-set values was based on existing quantitative indices, while others were based on an iterative process relying on multiple semi-structured expert interviews and an

online survey. It remains unclear how Kirchherr et al. (2016) used the qualitative data to establish the fuzzy-set values, although the online appendix with the fuzzy-set scores for some sub-dimensions per case gives some insights in the researchers' reasoning. Moreover, while the thresholds for in- and exclusion of a set are rather straightforward for the data based on indices (e.g., a ranking is used), it is unclear how the thresholds based on the (combination of) qualitative data are determined.

Table 2: How to determine the thresholds for in- and exclusion of a set?

Approach	Examples
Develop a rubric to assign codes to outcome and conditions.	Chatterley et al. 2013, 2014; Fischer 2015; Henik 2015
Construct an imaginary case for full-membership based on the case context, and a case for non-membership based on theoretical knowledge. The thresholds for in- and exclusion are then placed somewhere in-between these values.	Basurto & Speer 2012
Apply the GMET where qualitative anchor points are based on a combination of the positive or negative direction on a case's membership and the relative importance of the attribute.	Tóth et al. 2016
Conduct a cluster analysis by using, for example, Tosmana (Cronqvist 2016).	Kim & Verweij 2016; Li et al. 2016; Vergne & Depeyre 2016
Base the thresholds on a large gap in the numerical data between the various cases (and preferably complement this with other approaches).	Li et al. 2016; Vergne & Depeyre 2016

Note: Details are provided in Table A1 in Appendix III.

Another way to going about determining the thresholds for in- and exclusion of a set is the approach suggested by Basurto and Speer (2012). They first construct two imaginary ideal cases, one representing full membership in a set (1) and one representing full non-membership (0). Whereas the former imaginary case is largely based on knowledge about the case context, the latter case is largely informed by theoretical insights. The thresholds for in- and exclusion of the set, then, are put in

between the two “extreme” values. This could be a useful approach, but also one that is probably not applicable for all studies because it requires both a high level of contextual information and much theoretical knowledge.

An inductive approach that several studies adopt is to use the possibilities for setting thresholds that are available in QCA-software, particularly the threshold setter in Tosmana (Cronqvist 2016). Also this can be a useful approach, but it should never be used mechanically. Researchers using the threshold setter should always check the face validity of the thresholds set and possibly complement this approach by another strategy.

How to establish the degree to which a case is “in” or “out” of a set?

A second issue relating to calibration is how to establish the degree to which a case is “in” or “out” of a set. While the overview in Table A1 in Appendix III shows that all studies pay attention to this important issue, many studies do not make fully clear how the qualitative data has been used to this end.

For example, in his QCA-analysis of satisfactory outcomes in the implementation phase of infrastructure projects, Verweij (2015) used both qualitative and quantitative data to calibrate the outcome and the conditions.[7] For one of Verweij’s conditions, management and cooperation

7 Like in other studies that combine quantitative and qualitative data to calibrate conditions and/or the outcome (e.g., Vis, 2010), the quantitative material “dominated” the calibration here in that it is the benchmark that may be adjusted based on the qualitative material.

orientations in the projects, he used only qualitative data (general management summaries, stakeholder environment reports, and market cooperation reports). Verweij coded these documents using qualitative data-analysis software, constructed qualitative case descriptions, and cross-compared these in two iterations. In the latter step, 'a few small contradictions led to re-examination of some project data, recoding, and a recalibration' (Verweij, 2015: 1883). While doing so is indeed common practice in QCA, as Verweij notes (idem), this statement fails to make clear what he has done exactly, and why. Consequently, it provides other researchers with little guidance on how to proceed in their own work (which, admittedly, was also not Verweij's objective).

The same holds for Verweij and colleagues' (2013) study of what makes governance of spatial planning projects work. Verweij et al. used a variety of qualitative sources to calibrate their outcome and conditions, including semi-structured interviews, documents and participant observation. Using this information, '(…) the cases were given qualitative and/or quantitative descriptions for each separate condition. (…) Next, each case was given a membership score of 1.00, 0.67, 0.33, or 0.00 on each of the conditions after an iterative dialogue between theoretical and substantive knowledge took place within the research team. (…) After several rounds of analyses during which [the researchers] added to each other's case knowledge, the result was a raw data matrix [with the calibrated scores]' (Verweij et al., 2013: 1040). Again, this statement does not specify what exactly has been done in the process of calibration. Note that in line with good QCA-practice, Verweij et al. (2013) published the

reasoning behind most of their coding in their Appendix, allowing other researchers to replicate their work or assess the validity of the scores. Yet, these tables do not include – understandably – what for our purposes is especially interesting: the decisions behind the coding decisions.

More generally speaking, several studies describe the process of how the degree to which a case is “in” or “out” of the set has been established, for example by using multiple coders, but do not detail the reasoning behind this process (e.g., Chai & Schoon, 2016). To some extent this holds also for Van der Heijden (2015). Using a systematic coding scheme and qualitative data analysis software (Atlas.ti), he explored the data systematically and gained insights ‘into the “repetitiveness” and “rarity” of experiences shared by the interviewees, and those reported in the existing information studied’ (p. 581). However, *how* this was done exactly is not made clear.

Some studies are brief in discussing their coding decisions, such as Chatterley et al. (2013: 416) who only state that ‘when possible, we used multiple data sources as recommended in QCA literature’. In other studies, such as Basurto (2013), there is no information on how the interview data were translated into the fuzzy set values.

Although generally little insight is thus provided in how the researchers decide which values to assign to qualitative data, some suggestions are made on how to establish the degree to which a case is “in” or “out” of a set qualitative data. The first option is to directly ask interviewees to give their answer guided by a Likert scale or by pre-determined options corresponding with a certain value. Fischer (2014)

used interview data from approximately 250 face-to-face semi-structured interviews to derive values for the outcome and three conditions. To calibrate his outcome (i.e. policy change), Fischer used the following approach. First, he directly asked interviewees 'to indicate their perception of change on a scale from 1 (close to status quo) to 5 (major policy change)' (p. 350). Subsequently, Fischer (2014: 350-351) states that the 'average perception of actors about the outcome of each process was transformed into a fuzzy-value by simply rescaling the value on the 1–5 scale to a 0–1 scale'. He used the same approach for the calibration of two of the three conditions. Fischer's approach of taking the average of the actors' perceptions would be one way to calibrate interview data. Straightforwardness is an advantage of this approach. But it is also a mechanistic approach that may not result in a valid fuzzy-set.[8] This approach of take-the-average-of-interviewees'-perceptions should therefore be applied only when the resulting fuzzy-set is a valid one (to be assessed through, for instance, triangulation of data). In a later study, Fischer (2015) used the mean of two measures based on the interviewees' responses: (1) the "part-of-the-sum" measure, which sums up actors' reputational power and expresses the sum as the part of the total power of all coalitions) and (2) the average measure. Fischer (2015: 253, note 9) states that the former measure typically overestimates 'the power of coalitions that contain a lot of actors with little power', whereas the latter

8 A drawback of this approach is that it provides a valid fuzzy-set only if the average adequately reflects actors' perceptions. If, however, there is one or more interviewee with a strongly diverting perception – formulated in more quantitative terminology: when the standard deviation is high –, taking the average fails to result in a valid fuzzy-set.

‘tends to underestimate the power of these coalitions’. The average of these two measures reduced the potential biases. Still, also this second approach should not be applied mechanistically and the resulting membership scores should be assessed for their face validity. Kirchherr et al. (2016) also averaged the calibrated values for the different attributes of the conditions in order to obtain values for their analysis. They also acknowledge the potential weakness of this approach, which ‘could introduce misfits between the verbal meaning of a concept and its operationalization’ (Kirchherr et al. 2016: 39). Kirchherr et al. addressed this risk by reviewing all averaged calibrations of the conditions and change or recalibrate the attributes when they found that the conditions’ values did not correspond to their averaged operationalization. Alternatives for taking the average value is taking the weakest link (i.e. the minimum value of the attributes of the concept) and substitutability (i.e. taking their maximum value) (Ragin 2000). Chatterley et al. (2014) take the minimum value of two attributes to derive at the final value for the condition, since the theory that they use suggests that either one of the two attributes should be present for the outcome to be present. Basurto and Speer (2012), conversely, take the maximum of the three measures since it does not matter for the outcome which of the three is present.

In addition to the challenge of how to determine the thresholds for in- and exclusion of a set, and the challenge of establishing the degree to which a case is “in” or “out” of a set, a third challenge relating to calibration is how to deal with concepts that are truly absent and those

that are not addressed. We turn to this challenge next.

What does a zero actually mean?

QCA scholars encounter an additional challenge when coding interviews: how to differentiate between concepts that are truly absent and those that are not mentioned in the interview? We think this issue relates to a more general question of what a zero ("fully out of the set") actually means. This question is also important for other types of qualitative data, such as existing document or archive material (and also for quantitative data for that matter).[9] If a concept is not mentioned in a document, does this mean that it is absent, or is there just no information on it in the document? Data triangulation is one way to find out which of the two it is, which is another reason why we consider using more than one source of data important. In any case, it will oftentimes be useful to explore the zeros in more detail, so as to try and find out why the condition was absent or why the information was missing. There are, for instance, different reasons why an interviewee does not address an issue (e.g., too sensitive versus irrelevant).

In the large majority of the 22 studies we reviewed, this question seems not to have been an issue though. In these cases, a zero meant truly absent (or at least it seemed). We identified one study that did come across this issue: Kim and Verweij 2016. For calibrating their outcome "American states levels of environmental justice (EJ) policy", Kim and

9 The informal *communis opinio* seems to be that the zero is "a big mess". While we cannot solve that issue here, we can highlight some of the things that researchers working with qualitative data can think about.

Verweij qualitatively assessed a range of documents and additionally used survey results from other authors. They first assigned each state to one of nine groups. States with either “No action” or “no information” were assigned a value 0 (see table 1 in their paper).[10] Subsequently, those states with levels 6-8 were coded as in the set (>0.5); and states with levels 0-5 as out of the set (<0.5). The crossover point (0.5) was set at 5.5. Their assignment of value “0” to states with both “no action” and “no information” raises questions about the comparability of the coding values.

Ideally, researchers construct their interview scheme in such a way that all concepts are addressed during the interview. Creating a separate section for each condition in the interview guideline as proposed by Basurto and Speer (2012) is a useful way to do so. It is similar to Tóth et al's. (2016) suggestion to have an initial template based on previous literature to identify the same conditions across the cases. When all concepts are addressed, a value “0” would then indeed only be assigned to concepts that are truly absent. However, due to the iterative nature of QCA, which allows for the inclusion and exclusion of conditions during the process, a lack of data about one or more concepts cannot always be avoided. A similar data deficiency can also be experienced by researchers

10 The decision to assign value “0” both to cases where no information is available and where the condition is absent may or may not influence the substantive results. In Kim and Verweij (2016), assigning value “0” both to states with “no action” and “no information” only influences the results when information would indeed be available *and* would be assigned a value 6-8, since only then would they change from being “out” of the set to being “in” the set. When instead of “0”, states would have received value 1-5, no numerical difference would exist and the results would not change since the cases would stay at the same side of the qualitative anchor.

analysing already existing data for QCA. We suggest two options to deal with such gaps in the data. First, interviewees can be approached again and specific questions can be asked about the missing concepts. This allows researchers to establish whether the concept was indeed absent in the case, or whether it was just not mentioned the first time. The values of the corresponding condition can then be changed if necessary. When it is not possible to gather additional data, for example because of practical constraints, a second-best option is to conduct sensitivity analyses. We turn to these in the next section, but mention how to deal with the specific issue of the truly absent versus not mentioned problem here. There are several sensitivity analyses possible: (1) remove the conditions where this problem occurs and assess the effect, (2) assign value “0.51” (i.e. just “in” the set) to cases that have a value that is not “truly” absent (i.e. where the concept is indeed absent) to differentiate between the two findings, and (3) exclude the cases where the concept is “not mentioned” from the analysis. With option 3, it is important to make sure the cases-to-conditions ratio is still acceptable, typically 1 condition to 3 cases (Marx 2010).

Summing up, the challenge of how to differentiate between those concepts that are truly absent and those that are not mentioned can be addressed in several ways. For interviews: Construct the interview scheme such that all concepts are addressed during the interview, for example by creating a separate section for each concept (Basurto & Speer 2012). If this is not possible, approaching interviewees again to fill in the missing data is a second-best option. A third-best option would be

to conduct sensitivity analyses.

6. Sensitivity Testing

It is increasingly acknowledged that testing the robustness of the findings by means of sensitivity analyses should be part of a good QCA (Schneider & Wagemann 2012). The methodological literature on QCA pays increasing attention to sensitivity analyses (Baumgartner & Thiem 2015; Marx 2010; Skaaning 2011; Thiem 2014; Thiem et al. 2016), including how to deal with different types of errors (Maggetti & Levi-Faur 2013). And in the literature criticizing QCA (e.g., Hug 2013; Lucas & Szatrowski 2014; Paine 2016), the alleged lack of the robustness of the findings is regularly indicated as a key problem of the approach. (but see Baumgartner & Thiem 2017).

The QCA-literature provides several suggestions on how to assess the robustness of findings using sensitivity analyses. A non-exhaustive list includes: cases and conditions can be dropped or added, calibration functions changed, consistency thresholds altered (Schneider & Wagemann 2012; Thiem 2014; Thiem et al. 2016), definitions of the set values can be changed, alternative measures for a concept can be used (Basurto & Speer, 2012), and calibration thresholds of raw data into set-membership can be changed or the frequency of cases linked to configurations can be altered (Skaaning 2011). These suggestions are by no means specific for testing the robustness of findings based on qualitative data. Changing the consistency thresholds, for example, can be done irrespective of whether the data used are qualitative,

quantitative, or both (see for examples with qualitative data Tóth et al. 2016; Kim & Verweij 2016). Similarly, changing the frequency of cases linked to the configuration (in our research by including only configurations with at least two instead of one case) can be done in QCA irrespective of the kind of data used. Still, the higher is the number of cases, the more appropriate this sensitivity test becomes. Since studies using qualitative data oftentimes – but not necessarily – have a relatively low number of cases, this will in many cases not be the most important sensitivity analysis to conduct.

There are also sensitivity analyses in which qualitative data play a specific role. We list these in Table 3; Table A1 in Appendix III provides a more detailed discussion.

Table 3: Which sensitivity tests to conduct for assessing the robustness of the findings of QCA based on qualitative data?

Approach	Examples
Drop or add cases motivated by extensive case knowledge.	Kirchherr et al. 2016
Alter the attributes of a condition based on knowledge about the case context.	Kirchherr et al. 2016
Replace conditions by one of their attributes based on the importance that the data from the interviews, documents, or literature assigned to a specific attribute.	Kirchherr et al. 2016
Re-run the analysis with a new, more extreme, outcome that has – consequently – a different qualitative breakpoint (anchor point) for being "in" the set. Go back to the qualitative data to calibrate this new outcome (which can be done starting from the original outcome's calibration).	Tóth et al. 2016

First, the available qualitative data can be a strong motivator to decide which cases to drop or add in the sensitivity analysis. Kirchherr et al. (2016) include an extensive section on robustness in which they amongst other factors motivate the choices to exclude certain cases (i.e. dams) based on their case descriptions as presented in an appendix. Examples

of such motivations are the comparability of dam types and ambiguity in data collection for a specific dam compared to the others. By presenting their argumentation both in the main text and in an appendix, readers can follow and judge whether it indeed makes sense to exclude these cases. Dropping cases can be a useful way to assess findings' robustness. However, let us state once more that when dropping cases, the ratio between the number of cases and conditions may become too low, and the results unreliable (Marx 2010).

A second type of sensitivity test is conducted by altering the different attributes of the condition (Kirchherr et al. 2016). An example would be to base the membership score on only one attribute instead of multiple ones. Also here, the motivation for this choice needs to be based on knowledge about the case context. An argument could for example be that the now omitted attributes introduced noise to the condition's operationalization. This decision can for example be based on the importance assigned to a specific attribute(s) in the interviews, relevant documents or literature.

Another type of test, which we subsume under the heading of sensitivity tests but which technically is a test to understand better which factors or mechanisms "drive" the outcome, is conducted by Tóth et al. (2016: 11), who follow the procedure of Fiss (2011). To arrive at this improved understanding, a new outcome is introduced that is more extreme than the original one (in Tóth et al. 2016: *very high* relational attractiveness of the customer [RAC]). The qualitative threshold (what Tóth et al. call anchor point) for being "in the set" is higher for "very high

RAC" than it was for "RAC", meaning that some cases will no longer be "in" the set of this new outcome. The calibration of the outcome requires going back to the qualitative data and assigning appropriate (fuzzy) set values. The calibration of the original outcome can be used as a starting point for doing so.

7. Presentation of process and findings

In order to facilitate the replication of studies, the data sources and calibration process need to be presented transparently and comprehensively (Gerring 2012). Ideally, this should also be done concisely, so as to make the material easily accessible. These two goals – transparency and comprehensiveness versus conciseness – may often conflict. The way various QCA-scholars present the data calibration process, and hence the actual possibility for replication, varies strongly among the studies that we assessed. Table 4 summarizes the material from Table A1 in Appendix III on this.

Table 4 demonstrates that most reviewed studies provide *some* information on the calibration procedure (Aversa et al. 2015 provide too little information). Quite a few studies provide (quite) some information but not all that would be required for full transparency. These latter studies are thus not comprehensive (Basurto 2013; Chai & Schoon 2016; Chatterley et al. 2013 & 2014; Crilly 2011; Henik 2015; Kim & Verweij 2016; Metelits 2009; Vergne & Depeyre 2016; Verweij 2015; Verweij & Gerrits 2015).

Table 4: How to present the calibration procedure transparently (and concisely)?

Approach	Examples
Table in main text, full information	Kirchherr et al. 2016; Toth et al. 2016 (for 1 GMET)
Table in main text, partial information	Basurto 2013; Chai & Schoon 2016; Chatterley et al. 2013 & 2014; Crilly 2011; Kim & Verweij 2016; Li et al. 2016; Metelits 2009; Vergne & Depeyre 2016; Verweij 2015; Verweij & Gerrits 2015; Verweij et al. 2013
Text boxes	Basurto & Speer 2012
Discussed in words in main text, typically partial	Chai & Schoon 2016; Chatterley et al. 2013; Crilly 2011; Henik 2015; Kim & Verweij 2016; Kirchherr et al. 2016; Li et al. 2016; Verweij 2015
Discussed in words in appendix, typically partial	Vergne & Depeyre 2016
Table(s) in appendix, full information	Fischer 2014; Kirchher et al. 2016; Li et al. 2016; Thomann 2015; Van der Heijden 2015; Verweij et al. 2013; Wang 2016
Table(s) in appendix, partial information	Basurto 2013

There are also some studies that are comprehensive in this regard. Kim and Verweij (2016: 3), for example, include a table with the motivation of the assignment of US states to a specific category based on a combination of descriptions that fits the specific state and the results of a survey study, indicated with a superscript symbol next to the state. Fischer (2014) presents the calibration of outcome and conditions in tables in appendices. Both studies use a rather straightforward way to derive at the calibrated and related fuzzy-set values by respectively referring to survey results and by directly asking interviewees to “score” their outcome and conditions and subsequently taking the average. Hence, replicability of their findings should also be rather straightforward. Deriving at similar results becomes more complicated when the data needed as input for a specific sub-dimension cannot be directly derived

from interviewees' answers. While journal space limitations often makes the disclosure of all details of the calibration process within an article challenging, using (online) appendices, an option that a growing number of scientific journals are offering nowadays, is a way to give more insight in the argumentation of researchers (Basurto & Speer 2012). This suggestion is taken up by various of the studies that we reviewed (Basurto 2013; Fischer 2014; Kirchherr et al. 2016; Thomann 2015; Wang 2016).

8. Discussion

With this paper, we aimed to contribute to best practices in QCA research by discussing six challenges that QCA-researchers face when using such data. By examining 22 QCA-studies using qualitative data, we demonstrated that each challenge can be addressed in several ways. Which direction(s) the researcher ultimately chooses will depend, amongst other things, on the specific research question, the type of data that are already available, and available time and financial resources. However, we would like to highlight five directions that we think every researcher using qualitative data for QCA should take into account.

First, the validity of a study that uses qualitative data enhances significantly by the use of multiple data sources (that is, by data triangulation), which is why we consider this a direction to follow. Of course, the benefits of data triangulation also apply to qualitative studies that do not use QCA, and to quantitative studies using QCA. Furthermore, a study's transparency increases substantially if the researcher makes

explicit and presents clearly the choices he or she had made (to the extent that this is possible given issues of for example confidentiality). For QCA, what is particularly important in this regard is that the researcher are clear about his or her choices when different data sources contradict one another to such an extent that there is disagreement about whether the case is “in” or “out” of the set. Ideally, this clarification is done in the main text (at least in summary).

Second, generally speaking, QCA-researchers could be even more explicit about how they derive at certain thresholds for in- and exclusion of a set. Depending on the type of data (to be) collected, these thresholds can for example be determined by constructing an imaginary, ideal-typical case, be based on a classification of the responses given in interviews, or by using cluster analysis.

Third, QCA-researchers could also pay more attention to the zeroes in their calibrated data, since it is likely that these zeroes are a “big mess” – not just in studies using qualitative data for that matter. In this paper, we discussed the issue of cases whose condition(s) or outcome are coded zero because they are “not mentioned” (or not identified in, for example, documents) versus cases whose condition(s) of outcome are coded zero because they are “truly absent”. This specific issue of not-mentioned versus truly absent seemed not to be a problem in most of the studies we reviewed; the majority of authors did not make note of it. Still, based on among other things informal discussions with QCA-researchers using qualitative data, we believe that the zeroes are likely to be a substantial problem in QCA. It goes beyond this paper’s scope to discuss

this issue in more detail, but it is an important area for further research.

Fourth, our literature review showed that conducting sensitivity tests in (qualitative) QCA is not yet common practice. Various tests are especially suitable when dealing with qualitative data, such as changing the number of cases, altering the conditions, or re-running the analysis with a more extreme outcome.

Finally, when presenting the data in a transparent yet concise way, a balance should be sought in giving brief explanations and/or illustrations in the main text and using tables in the main text and/or in (online) appendices.

The main limitation of our research is probably that our review may not have been exhaustive. The reason for this is mainly that we did not conduct a fully systematic review but instead used the various search strategies described in Section 3. For practical reasons, we also restricted ourselves to English papers only. Additionally, we selected only those studies that provided valuable input for Table A1 in Appendix III (that is, for our 6 questions). Even though our review is not exhaustive, we are confident that because of the combination of different sources (Scopus, ISI Web of Science, snowballing) complemented with input from several QCA-experts, we covered quite a broad range of papers on the use of qualitative data for QCA. Thereby, we expect that the directions on how to address these challenges regarding using qualitative data in QCA contribute to best practices in QCA.

References

Aversa, Paolo, Santi Furnari, and Stefab Haefliger. 2015. “Business Model Configurations and Performance: A Qualitative Comparative Analysis in Formula One Racing, 2005–2013.” *Industrial and Corporate Change* 24(3): 655–76.

Basurto, Xavier. 2013. “Linking Multi-Level Governance to Local Common Pool Resource Theory Using Fuzzy-Set Qualitative Comparative Analysis: Insights from Twenty Years of Biodiversity Conservation in Costa Rica.” *Global Environmental Change* 23(3): 573–587.

Basurto, Xavier, and Johanna Speer. 2012. “Structuring the Calibration of Qualitative Data as Sets for Qualitative Comparative Analysis (QCA).” *Field Methods* 24(2): 155–74.

Baumgartner, Michael. 2015. “Parsimony and Causality.” *Quality & Quantity* 49(2): 839–56.

Baumgartner, Michael, and Alrik Thiem. 2015. “Model Ambiguities in Configurational Comparative Research.” *Sociological Methods & Research*.

———. 2017. “Often Trusted But Never (Properly) Tested: Evaluating Qualitative Comparative Analysis.” *Sociological Methods & Research*.

Chai, Ying, and Michael Schoon. 2016. “Institutions and Government Efficiency: Decentralized Irrigation Management in China.” *International Journal of the Commons* 10(1): 21–44.

Chatterley, Christie et al. 2014. “A Qualitative Comparative Analysis of Well-Managed School Sanitation in Bangladesh.” *BMC Public Health* 14(6): 1–14.

Chatterley, Christie, Karl G. Linden, and Amy Javernick-Will. 2013. “Identifying Pathways to Continued Maintenance of School Sanitation in Belize.” *Journal of Water, Sanitation and Hygiene for Development* 3(3): 411–22.

Crilly, Donal. 2011. “Predicting Stakeholder Orientation in the Multinational Enterprise: A Mid-Range Theory.” *Journal of International Business Studies* 42(5): 694–717.

Cronqvist, Lasse. 2016. “Tosmana.” http://www.tosmana.net.

Fischer, Manuel. 2014. “Coalition Structures and Policy Change in a Consensus Democracy.” *Policy Studies Journal* 42(3): 344–66.

———. 2015. “Institutions and Coalitions in Policy Processes: A Cross-Sectoral Comparison.” *Journal of Public Policy* 35(2): 245–68.

Fiss, Peer C. 2011. “Building Better Causal Theories: A Fuzzy Set Approach to Typologies in Organization Research.” *Academy of Management Journal* 54(2): 393–420.

Fiss, Peer C., Axel Marx, and Benoît Rihoux. 2014. “Comment: Getting QCA Right.” *Sociological Methodology* 44(1): 95–100.

Fiss, Peer C., Dmitry Sharapov, and Lasse Cronqvist. 2013. “Opposites Attract? Opportunities and Challenges for Integrating Large-N QCA and Econometric Analysis.” *Political Research Quarterly* 66(1): 191–98.

Gerring, John. 2012. *Social Science Methodology : A Unified Framework*. Cambridge: Cambridge University Press.

Goertz, Gary, and James Mahoney. 2012. *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton and Oxford: Princeton University Press.

van der Heijden, Jeroen. 2015. “What Roles Are There for Government in Voluntary Environmental Programs?” *Environmental Policy and Governance* 25(5): 303–15.

Henik, Erika. 2015. “Understanding Whistle-Blowing: A Set-Theoretic Approach.” *Journal of Business Research* 68(2): 442–450.

Hug, Simon. 2013. “Qualitative Comparative Analysis: How Inductive Use and Measurement Error Lead to Problematic Inference.” *Political Analysis* 21(2): 252–65.

Ide, Tobias. 2015. “Why Do Conflicts over Scarce Renewable Resources Turn Violent? A Qualitative Comparative Analysis.” *Global Environmental Change* 33(7): 61–70.

Kim, Yushim, and Stefan Verweij. 2016. “Two Effective Causal Paths That Explain the Adoption of US State Environmental Justice Policy.” *Policy Science*.

Kirchherr, Julian, Katrina J. Charles, and Matthew J. Walton. 2016. “Multi-

Causal Pathways of Public Opposition to Dam Projects in Asia: A Fuzzy Set Qualitative Comparative Analysis (fsQCA).” *Global Environmental Change* 41(November): 33–45.

Li, Yanwei, Joop Kopennjan, and Stefan Verweij. 2016. “Governing Environmental Conflicts in China: Under What Conditions Do Local Governments Compromise?” *Public Administration* 94(3): 806–822.

Lucas, Samuel R, and Alisa Szatrowski. 2014. “Qualitative Comparative Analysis in Critical Perspective.” *Sociological Methodology* 44(1): 1–79.

Maggetti, Martino, and David Levi-Faur. 2013. “Dealing with Errors in QCA.” *Political Research Quarterly* 66(1): 198–204.

Marx, Axel. 2010. “Crisp-Set Qualitative Comparative Analysis (csQCA) and Model Specification: Benchmarks for Future csQCA Applications.” *International Journal of Multiple Research Approaches* 4(2): 138–58.

Metelits, Claire M. 2009. “The Consequences of Rivalry: Explaining Insurgent Violence Using Fuzzy Sets.” *Political Research Quarterly* 62(4): 673–84.

Paine, Jack. 2016. “Set-Theoretic Comparative Methods: Less Distinctive Than Claimed.” *Comparative Political Studies* 49(6): 703–41.

Ragin, Charles C. 1987. *The Comparative Method. Moving beyond Qualitative and Quantative Strategies*. Berkeley: University of California Press.

———. 2000. *Fuzzy-Set Social Science*. Chicago and London: The University of Chicago Press.

———. 2008. *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago and London: University of Chicago Press.

———. 2013. “New Directions in the Logic of Social Inquiry.” *Political Research Quarterly* 66(1): 171–74.

———. 2014. “Lucas and Szatrowski in Critical Perspective.” *Sociological Methodology* 44(1): 80–94.

Rihoux, Benoît et al. 2013. “From Niche to Mainstream Method? A Comprehensive Mapping of QCA Applications in Journal Articles from

1984 to 2011.” *Political Research Quarterly* 66(1): 175–84.

Scheck McAlearney, Ann, Daniel Walker, Alexandra D. Moss, and Nina A. Bickell. 2016. “Using Qualitative Comparative Analysis of Key Informant Interviews in Health Services Research: Enhancing a Study of Adjuvant Therapy Use in Breast Cancer Care.” *Medical Care* 54(4): 400–405.

Schneider, Carsten Q, and Claudius Wagemann. 2012. *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. Cambridge: Cambridge University Press.

Skaaning, Svend Erik. 2011. “Assessing the Robustness of Crisp-Set and Fuzzy-Set QCA Results.” *Sociological Methods & Research* 40(2): 391–408.

Thiem, Alrik. 2014. “Membership Function Sensitivity of Descriptive Statistics in Fuzzy-Set Relations.” *International Journal of Social Research Methodology* 17(6): 625–42.

Thiem, Alrik, and Michael Baumgartner. 2016. “Back to Square One: A Reply to Munck, Paine, and Schneider.” *Comparative Political Studies* 49(6): 801–6.

Thiem, Alrik, Michael Baumgartner, and Damien Bol. 2016. “Still Lost in Translation! A Correction of Three Misunderstandings Between Configurational Comparativists and Regressional Analysts.” *Comparative Political Studies* 49(6): 742–74.

Thiem, Alrik, Reto Spöhel, and Adrian Duşa. 2016. “Enhancing Sensitivity Diagnostics for Qualitative Comparative Analysis: A Combinatorial Approach.” *Political Analysis* 24(1): 104–20.

Thomann, Eva. 2015. “Customizing Europe: Transposition as Bottom-up Implementation.” *Journal of European Public Policy* 22(10): 1368–87.

Tóth, Zsófia, Stephan C. Henneberg, and Peter Naudé. 2016. “Addressing the ‘Qualitative’ in Fuzzy Set Qualitative Comparative Analysis: The Generic Membership Evaluation Template.” *Industrial Marketing Management*.

Vergne, Jean-Philippe, and Colette Depeyre. 2016. “How Do Firms Adapt? A Fuzzy-Set Analysis of the Role Cognition and Capabilities in U.S.

Defense Firms' Responses to 9/11." *Academy of Management Journal* 59(5): 1653–1680.

Verweij, Stefan. 2015. "Producing Satisfactory Outcomes in the Implementation Phase of PPP Infrastructure Projects: A Fuzzy Set Qualitative Comparative Analysis of 27 Road Constructions in the Netherlands." *International Journal of Project Management* 33(8): 1877–87.

Verweij, Stefan, and Lasse M Gerrits. 2015. "How Satisfaction Is Achieved in the Implementation Phase of Large Transportation Infrastructure Projects: A Qualitative Comparative Analysis Into the A2 Tunnel Project." *Public Works Management & Policy* 20(1): 5–28.

Verweij, Stefan, Erik-Hans Klijn, Jurian Edelenbos, and Arwin Van Buuren. 2013. "What Makes Governance Networks Work? A Fuzzy Set Qualitative Comparative Analysis of 14 Dutch Spatial Planning Projects." *Public Administration* 91(4): 1035–55.

Vis, Barbara. 2010. *Politics of Risk-Taking: Welfare State Reform in Advanced Democracies*. Amsterdam: Amsterdam University Press.

———. 2012. "The Comparative Advantages of fsQCA and Regression Analysis for Moderately Large-N Analyses." *Sociological Methods and Research* 41(1): 168–98.

Wang, Weijie. 2016. "Exploring the Determinants of Network Effectiveness: The Case of Neighborhood Governance Networks in Beijing." *Journal of Public Administration Research And Theory* 26(2): 375–388.

Appendix I

Own summary of Basurto and Speer's (2012) stepwise procedure for qualitative data calibration for QCA

Step 1: Operationalize the conditions and the outcome

Operationalize the theoretical concepts into a preliminary list of measures of the conditions and the outcome, based on standard-scientific practice and/ or the researchers' knowledge of the empirical context. An iterative process leads to a final list of conditions and outcome.

Step 2: Develop the qualitative thresholds (anchor points) and elaborate the qualitative interview guideline

Develop initial qualitative thresholds (i.e. 1 for full membership, 0.5 for the cross-over point and 0 for full non-membership) based on the researchers' theoretical knowledge. The thresholds are later on refined based on the case context.

The interview guideline contains separate sections for each condition and the outcome. Each section includes an introductory eliciting question, sub questions on each attribute and specifying questions.

Step 3: Conduct a content analysis of the raw interview data

Code the raw interview data using qualitative data analysis software taking the preliminary list of attributes of the conditions and outcome (see step 1) as a starting point.

Step 4: Summarize the coded qualitative data

Systematically analyse the coded qualitative data by 1) examining all quotations with the same code from all cases and all interviewees, 2) extracting the quotations for each code sorted by type of interviewee and 3) summarizing all interview quotations with the same code for each case in a qualitative classification.

Step 5: Determine the fuzzy-set scale and define the fuzzy-set values

Determine the degree of precision of the fuzzy sets and define each of their values based on theoretical and case and context knowledge. Construct an imaginary case both for full membership and non-membership.

Step 6: Assign and revise the fuzzy-set values of the conditions and outcome for each case

Assign values within the fuzzy sets to each case by matching the qualitative classifications derived in step 4 with the fuzzy-set values from step 5. Then revise and adjust the assigned fuzzy-set values for all cases and all measures by going through one measure across all cases. Finally, aggregate the fuzzy-set values of all measures into the condition to which they belong and create a summary table.

Appendix II Own summary of the steps in Tóth et al.'s (2016) Generic Membership Evaluation Template (GMET) to calibrate qualitative data for QCA

The GMET is applied per condition or outcome, per case. So with say 10 cases, 3 conditions and 1 outcome, there are 40 GMETs to fill in. Tóth et al. (2016) do not discuss if and if so how these GMETs should be made available. In their paper, they include only one GMET as an example (which is the example we also use to indicate the different steps of the procedure below; note that these steps are not mentioned explicitly by Tóth et al. but are derived from their Table 2 on p. 6). We would strongly recommend to make the GMETs available, preferably through a data storage facility. Doing so will enable making full use of the possibilities this holds for, for instance, replication of the study's findings or using the calibrated qualitative data for other research projects. However, there may be ethical considerations because of which (some of) the GMETs cannot be made publicly available. In that case, we advise the researcher to indicate in general – but more specifically than is oftentimes done – how, for example, the qualitative data have been translated into the fuzzy-set scores. The researcher could, for instance, use rubrics and instead of "real" examples (from the interview data) use fictitious examples to illustrate how s/he went about.

Step 1: Overall case description from the perspective of the specific condition

Tóth et al. (2016) use the condition "Customer relations with good relational fit" as an example. Their illustrative example of an overall case description is the following: 'A sustainable but very difficult relationship with various problems at an inter-personal level (e.g. hidden agendas) as well as differences in corporate communication style (e.g. negotiations). The Customer's professional qualities are highly valued but power games around branding issues and ownership create a distrustful atmosphere with regular conflicts' (p. 6)

Step 2: List the dimensions or sub-measures (what we label following Goertz & Mahoney 2012 the attributes) of the condition

In the example of Tóth et al. (2016), these include for example "professional trust" and "frequent conflicts" (p. 6).

Step 3: For each of these attributes, provide the following information:

3a: A context-specific description, in the case of "professional trust" this is for example: 'there is trust in the abilities and skills of the customer' (p. 6);

3b: An indication of the direction/effect on membership (positive or negative);

3c: An indication of the relative intensity/relative importance (high, moderate or low);

3d: An illustrative quote.

Step 4: Provide supporting quantitative data (if applicable)

Step 5: Provide set membership score

Indicate in a note to the GMET what is the "verbal" meaning attached to the fuzzy-set membership scores.

Step 6: Summarize the argumentation for giving this set membership score

In the example of Tóth et al. (2016), the following argumentation is provided: ‘Various negative dimensions of the condition can be identified (some with articulate intense criticism, e.g. frequent conflicts) demonstrate that this case is “mostly but not fully out” of the set of “Good Relational Fit with the Customer”. Even though a positive dimension (professional trust) is present, this cannot balance the relative weight and importance of the dimensions with negative valence. The presence of this positive dimension is the reason why the fuzzy-set attribution score is not “fully out” in this specific case’. (p. 6).

Appendix III

Table A1. Overview of the studies using qualitative data for QCA based on our literature review

Author(s)	What qualitative data are used?	How is the threshold for in- and exclusion of a set determined?	How is the degree to which a case is “in” or “out” of a set, established?	How is differentiated between “truly absent” and “not mentioned” indicators?	Which sensitivity tests are conducted?	How is the calibration procedure presented?
Aversa et al.(2015)	Documents (e.g., industry reports, specialized and generalist press) and interviews	The authors use csQCA. They use their qualitative data to code the cases as being "in" (1.0) or "out" (0) of a set. However, despite the outcome, "high-performance firms", it is not immediately clear what are the sets (conditions) here. Probably, the one listed in Tables 5 and 6, but the calibration of these conditions is not discussed, nor are the calibrated data presented.	NA (csQCA)	NA	None	See column 3. Besides some information on the calibration of the outcome, this is not discussed.
Basurto (2013)	(Semi-)structured interviews conducted during field visits and archival records	Some conditions have continuous values based on percentages, others are dichotomous (presence/ absence or many/ few). The conditions with semi-continuous values have a five-point scale and the threshold lies between “more often than not” (0.6) and “less often than not” (0.4).	Either based on the assigned value or, in case of multiple measures constituting one condition, on averaging the measures.	NA	None	A table in the main text states the types of states per condition (e.g., 4 value) and defines them. No information is given on how the interview data translate into the values. The appendix contains tables with fuzzy-set values.
Basurto & Speer (2012)	Interviews, meeting minutes and information on municipal budgets	Full membership and non-membership are determined by constructing imaginary cases. The cross-over point is set in between. All values are based on theoretical and case knowledge.	The relevant interview codes for each case are matched with a predetermined qualitative classification and related (four-value) fuzzy-sets.	NA	None	The paper contains text boxes with examples on how the data are calibrated. The empirical data are merely used to illustrate the proposed calibration procedure.

Author(s)	**What qualitative data are used?**	**How is the threshold for in- and exclusion of a set determined?**	**How is the degree to which a case is "in" or "out" of a set, established?**	**How is differentiated between "truly absent" and "not mentioned" indicators?**	**Which sensitivity tests are conducted?**	**How is the calibration procedure presented?**
Chai & Schoon (2016)	Statistical reports complemented with in-depth field interviews and case material	A software program is used to divide the outcome into four segments with related fuzzy values. The conditions are coded either present or absent, whereby the reasoning is at times not that straightforward. NB: With only crisp conditions, it makes no sense to have a fuzzy outcome as here.	"Absence" or "presence" of some conditions is determined based on interviews.	NA	None	The authors state they use the approach by Basurto and Speer (2012), but do not state how. The dichotomized data are presented in a table.
Chatterley et al. (2013)	Data from observations (e.g. photos), interviews and focus groups	A coding scheme is presented with qualitative descriptions (derived from literature) representing the membership (1) and non-membership (0) values for the conditions and the outcome.	"Absence" or "presence" of the conditions is determined based on observations and interviews.	NA	None	A detailed coding scheme is included in the main text. The dichotomized data are also presented, whereby some values are supported by quotes in the text.
Chatterley et al. (2014)	Semi-structured interviews, focus groups and data from observations (e.g. photos)	A coding rubric is presented with qualitative descriptions representing the 4 fuzzy-values for each condition and the outcome. No information is provided on how the thresholds are determined.	Values for outcome and conditions obtained by taking the minimum value of the sub-measures	NA	None	A table with the values for each condition and outcome per case is presented. Some of the values are supported by quotes in the text.
Crilly (2011)	Structured interviews and documents. The interview data are validated by having multiple interviews per case. Interviews were coded by at least two researchers using a clear coding scheme, with discrepancies being	The outcome's thresholds are based on the interviews, as are the values in-between (4-value fuzzy-set). The decision is explained clearly and illustrated with an example from the interviews.	The calibration of the 7 conditions in 4 fuzzy values is mostly done by using "external", typically quantitative or quantified standards (e.g., human development report, or the amount of corporate revenues). The author discusses clearly how these measures are "translated" into the fuzzy values. One condition (local government	NA	The author followed Epstein et al. (2008) and re-ran the analysis with a reduced consistency threshold of 0.85 (p. 712).	A table with the calibrated data per case is provided (fuzzy-set data table). For the calibration of the outcome, illustrative examples from the interviews are provided; the calibration of the conditions is explained clearly in main text (see column 3).

	resolved by a panel of three researchers.		influence) is calibrated based on the interview data, which is also clearly explained. NB: Not a best practice is that the conditions are void of a direction (e.g. strategic orientation or local government influence).			
Author(s)	**What qualitative data are used?**	**How is the threshold for in- and exclusion of a set determined?**	**How is the degree to which a case is "in" or "out" of a set, established?**	**How is differentiated between "truly absent" and "not mentioned" indicators?**	**Which sensitivity tests are conducted?**	**How is the calibration procedure presented?**
Fischer (2014)	Semi-structured interviews, with official documents providing supplementary information	For two of the three conditions, Fischer takes the observed maximum (1.0) and observed minimum values (0), and uses the median observed valued as crossover point (.5). For the third condition, he takes the theoretical maximum (1.0) and minimum (0).	Calibration of outcome and conditions by asking interviewees directly to express their perception using either a five-point scale or predetermined categories. Then converting the average of the actors involved (20-30 per process, that is a case) into a fuzzy-value using the direct method of calibration.	NA	None	Tables with how the outcome and conditions were calibrated and what were the resulting membership scores are presented in an appendix. The main text includes a table with the fuzzy set data.
Fischer (2015)	Semi-structured interviews, with official documents providing supplementary information.	The thresholds are determined using the substantive knowledge from the qualitative material. A coding rubric, including a description for determining the three thresholds, is presented in the appendix.	See also column 3. For the outcome of for one of the conditions, the author uses the direct method of calibration. For the other 2 conditions, he used a 7-value fuzzy-set, whereby he avoids assigning the score 0.5 to cases.	NA	None	Tables with how the outcome and conditions were calibrated and what were the resulting membership scores are presented in an appendix. The main text includes a table with the fuzzy set data.
Henik (2015)	In-depth, semi-structured interviews	A coding rubric is presented with qualitative descriptions representing the 4 or 2 fuzzy-values for each condition and the outcome. NB: The calibration scheme included 0.5, which is to be avoided.	Also through application of the coding rubric on the interview transcripts. This was done by 2 blind coders. The averages of their scores are the final set attribute. The author notes that the coders 'agreed within 0.25 set membership points on more than 90% of the 960 items (…)' (p. 445). In a few cases, this seemed to depend also on quantified measures (e.g., the anger scale for example). Note that no information is provided	NA	None	The coding rubric is included in the main text.

			on *qualitatively* important differences across coders, namely when one codes an item as being "in" the set and the other as "out" of the set. This way, a discrepancy of . 15 (e.g., .45 vs. 0.6) can be more relevant than one of 0.3 (e.g., 0.6 vs 0.9). There is also no information provided on what the coders did, just what was the result hereof.			
Author(s)	**What qualitative data are used?**	**How is the threshold for in- and exclusion of a set determined?**	**How is the degree to which a case is "in" or "out" of a set, established?**	**How is differentiated between "truly absent" and "not mentioned" indicators?**	**Which sensitivity tests are conducted?**	**How is the calibration procedure presented?**
Li et al. (2016)	Secondary data from news reports, government documents, national laws and documents of environmental NGOs. Semi-structured interviews during empirical fieldwork. Attendance of 2 workshops.	Crisp-set QCA is used. The outcome is expressed as project relocations or cancellations (1) and project continuations (0). The threshold for the condition "scale of protests" is based on a big gap in the data (i.e. number of participants) combined with a value derived through cluster analysis using Tosmana QCA software.	Conditions are dichotomized: presence versus absence, occurrence versus non-occurrence, early versus late project stage.	NA	The authors make two comments about the robustness and validity. First, that a different cross-over point based on Tosmana cluster analysis does not influence the calibration. Second, that the 'symmetric nature of this finding strengthens the validity of the results of the respective analyses for the occurrence and non-occurrence of the outcome' (p. 14).	Calibration of the outcome and the conditions is presented in a table in the main text and the raw data are summarized in a table in the appendix. Justification for assigning the set membership scores can partially be derived from the case descriptions in another table in the main text.
Kim & Verweij (2016)	Reports determine the initial adoption level of environmental justice policies. These data are corroborated with survey results of other studies.	The qualitative anchors are determined based on existing indices and by using the Tosmana threshold setter (that is, cluster analysis).	Also mainly from existing indices and by using the Tosmana software.	NA	Sensitivity analysis based on different consistency cut-offs.	The three qualitative thresholds are presented in a table. The argumentation for these scores are discussed in the main text.
Kirchherr et al. (2016)	Peer-reviewed articles, press reports, online survey results and semi-structured interviews.	The authors used a 4-value and 2-value coding scheme to assign fuzzy-set values to either conditions or outcome, or to their attributes. Some of the fuzzy-set	The authors averaged the calibrated values for the condition's different sub-dimensions to derive at the fuzzy-set value of the condition. Subsequently, they reviewed all	NA	Three types of sensitivity analysis were conducted – dropping cases; introduction of additional conditions; and	How each condition and the outcome is calibrate is presented in the text, tables and an online appendix. The online

		values were based on existing quantitative indices, whereas others were based on interview and survey data.	averaged calibrations of the conditions and changed or recalibrated the sub-dimensions when the conditions' values were not face valid		alternative measures for a concept –, yielding a total of 11 sensitivity analyses, which are explained both in writing and in a table	appendix also provides information on the raw data, sensitivity analysis and calibration of conditions using various qualitative data sources
Author(s)	**What qualitative data are used?**	**How is the threshold for in- and exclusion of a set determined?**	**How is the degree to which a case is "in" or "out" of a set, established?**	**How is differentiated between "truly absent" and "not mentioned" indicators?**	**Which sensitivity tests are conducted?**	**How is the calibration procedure presented?**
Metelits (2009)	Interviews	The interview material is used to establish the qualitative breakpoints, as well as the other values of the 6-value fuzzy set for the outcome and the 3 conditions. *How* exactly the author has used the interview material to this end is not spelled out.	By means of the interview material. The author discusses per case the fuzzy-set scores for the outcome and the conditions, even though it is not always clear *how* she has made this judgment (see also column 2).	NA	None	In the main text (see column 3). Tables with fuzzy values for insurgent groups are provided per group (i.e., 3 cases) and jointly (3x3 = 9 cases). Given typical space limitations, we would advise to present such information only once (in 1 table).
Thomann (2015)	Legal documents, policy documents, secondary literature, telephone interviews, and written questionnaires with relevant actors and stakeholders.	For the outcome, the author used the theoretical maximum of the developed customization index (1.0) and its theoretical minimum (0), with 1.5 (on a scale of 4) as crossover point (0.5). For the conditions, the author used a combination of existing indices that constituted the attributes of an index that was calibrated indirectly, and conditions that were calibrated using the qualitative material, typically the interviews. The author clearly states the reasoning behind the thresholds. For one condition, the thresholds were based on the sample range (1.0 and 0) and its mean (0.5), so as to avoid unrealistic scenarios.	See column 3 on the left.	NA	The author conducted an analysis of the negation of the outcome.	The calibration procedure is discussed in an appendix. This appendix also presents the raw data matrix and the fuzzy membership scores.

Author(s)	What qualitative data are used?	How is the threshold for in- and exclusion of a set determined?	How is the degree to which a case is 'in' or 'out' of a set, established?	How is differentiated between 'truly absent' and 'not mentioned' indicators?	Which sensitivity tests are conducted?	How is the calibration procedure presented?
Tóth et al. (2016)	In-depth interviews	These thresholds are based on the GMET (Generic Membership Evaluation Template). Full membership (1.0) is given when overall intense and various positive dimensions; full non-membership (0) is given when overall intense and various negative dimensions.	The value of each attribute is determined by both its intensity/relative importance and by the positive or negative direction on the membership (see Appendix II). The 'more in than out' category is characterized by mostly but not exclusively positive dimensions, whereas the 'more out than in' value is described by mostly but not exclusively negative dimensions in relation to the case's condition membership.	NA	Sensitivity analysis based on different consistency cut-offs.	The Generic Membership Evaluation Template (GMET) is used to assign fuzzy values to conditions and outcome. The GMET is filled in for one condition as an example. The GMET for the remaining conditions is neither presented in the paper nor in an appendix.
Van der Heijden (2015)	In-depth face-to-face interviews and existing documentation (e.g. websites, reports).	The author describes the assignment of the 3 thresholds for the outcomes and the conditions in the appendix. He has used the empirical material to inform this assignment, but does not discuss how exactly he has used the material to this end.	The author used a 4-value fuzzy set for the outcomes and conditions. In the appendix, he describes the assignment of these values (see column 3 on the left).	The author made sure to obtain enough information on all indicators to obtain a valid measurement. To this end, he started by using information from websites, existing reports and other sources. Novel data on the cases were subsequently obtained through a series of interviews to fill in gaps in the data from other sources.	None	The calibration of the data, including the setting of the thresholds, is discussed in an online appendix.
Vergne & Depeyre (2016)	Government reports, memos, books, smi-structured interviews, commentaries on media, >100 letters to shareholders including yearly firm level data (e.g. firms annual reports), databases and online survey. First, the coding	The threshold for the outcome is based on an expert survey giving answers on a scale from 1-7. Value 4 indicates the crossover point, and intended to capture the average (e.g., in adaptation).The threshold for one of the conditions is based on a clear gap in the data around the 0.5 qualitative anchor, allowing to use the raw measure of the condition.	For the outcome, the scores of 5 experts (see column 2) were averaged into the final set membership scores. The authors indicate that in 59% of the cases, experts were in agreement (p. 1662). Them using the average scoring "averages out" the qualitative differences across the experts (e.g., one scoring 3, which would be out of the set, and another scoring 5, that is out of the set), but	The option "I don't know" was deliberately excluded in the expert survey. When someone was insufficiently knowledgeable, the authors asked that person not to complete the survey at all (p. 1661, note 8). When data about a specific indicator were missing, the authors turned to additional databases for information (but	The authors conducted an additional analysis in which they did include directional expectations. Additionally, they conducted robustness analyses using alternative measures for one indicator and the outcome.	The calibrated sets are presented in a table in the main text. Further details about the calibration are presented in an appendix. Figures in the main text provide qualitative illustrations of set memberships based on the letters to shareholders.

	procedures among the 2 authors were aligned. Then, the text fragments were independently coded using qualitative data analysis software NVivo 9. Agreement between the coders was high (90%) and the remaining ambiguities were resolved through discussion.		this may not result in a valid measurement. Calibration of one condition was based on letters to shareholders. Based on these letters, four values were given to each case (e.g. 0 indicating "not paying any attention" and 0.33 indicating "paying some attention").	report that they sometimes did not find more information) (p. 1679).		
Author(s)	**What qualitative data are used?**	**How is the threshold for in- and exclusion of a set determined?**	**How is the degree to which a case is 'in' or 'out' of a set, established?**	**How is differentiated between 'truly absent' and 'not mentioned' indicators?**	**Which sensitivity tests are conducted?**	**How is the calibration procedure presented?**
Verweij (2015)	General management summaries, stakeholder environment reports and market cooperation reports.	The qualitative anchors are determined based on existing indicators (such as project size), qualitative data (such as summaries by managers) and by using the Tosmana threshold setter (that is, cluster analysis).	To establish the degree of membership in the 4-value fuzzy sets, the author uses mainly existing indicators (such as project size), qualitative data (such as management summaries) and the Tosmana software.	NA	The author also conducted an analysis of the negation of the outcome.	The "raw" data plus the membership score are provided in a table. The reasoning behind this is discussed in the main text.
Verweij & Gerrits (2015)	Open interviews, multiple site visits, observations of project meetings, project documents and project websites.	The qualitative data are used to determine the multi-value scores (0, 1 or 2) and the Boolean ones (0 and 1). These scores were recalibrated in a second round because they yielded too many logical contradictions.	The conditions are broken down into categories. A value is assigned to each category which is then used for the mvQCA analysis. Summaries in a table provide some justification for why specific values are assigned to certain categories.	NA	None	Three tables in the main text respectively present a qualitative description of each case, the category assigned to each case, and the value assigned to each category as part of the mvQCA.
Verweij et al. (2013)	Semi-structured interviews, documents and participant observation.	The qualitative anchors are determined based on existing indicators (such as the number of actors involved) and by the interview and secondary data.	Quantitative and/ or qualitative case description for each condition are translated into fuzzy-set scores. The authors first scored the cases individually. A subsequent iterative dialogue of several rounds between researchers' theoretical and substantive case knowledge was used to amend each other's scores. This resulted in the assignment of case membership scores on each condition (based on averaging the indicators).	NA	None	The scores on each separate indicator are presented in tables in the appendices. Some scores are based on quantitative data (e.g. number of actors involved). A qualitative description with corresponding qualitative scores (e.g. high-moderate-low) is given for the other indicators.

Author(s)	**What qualitative data are used?**	**How is the threshold for in- and exclusion of a set determined?**	**How is the degree to which a case is 'in' or 'out' of a set, established?**	**How is differentiated between 'truly absent' and 'not mentioned' indicators?**	**Which sensitivity tests are conducted?**	**How is the calibration procedure presented?**
Wang (2016)	Semi-structured in-depth interviews, photos, survey evidence	Based on the existing "raw" data (see column 4), whereby the coding decision is not explained very clearly (e.g., why are neighbourhoods below the 27% percentile clearly poorly governed, i.e. fuzzy value 0)?	The author discusses in much detail how he measured the outcome and the causal conditions. The result hereof are the "raw" data, which were also used in a network analysis and in a linear regression. How these "raw" data are translated into fuzzy values is discussed in an appendix. Some choices are explained well, but others less so (see also column 3). NB: The score of 0.5 is given, which is problematic.	NA	Alternative specifications of the calibration thresholds, specifically – following Fiss (2011) – the specification of two new crossover points for the fuzzy conditions. The new crossover points are provided in a table, as are the changes (or lack therefore) in the causal paths and the biggest change in coverage or consistency.	In an appendix. There is no table summarizing the calibration procedure.

Note: NA means not applicable.