Addressing the Challenges Related to Transforming Qualitative into Quantitative Data in Qualitative Comparative Analysis

Accepted for publication in Journal of Mixed Methods Research

Debora de Block
Wageningen University & Research
E: debora.deblock@wur.nl

Prof. dr. Barbara Vis (*corresponding author*)
Utrecht University, Utrecht University School of Governance
Bijlhouwerstraat 6, 3511 ZC Utrecht
E: b.vis@uu.nl; Web: www.barbaravis.nl / www.highriskpolitics.org

Abstract

The use of qualitative data has so far received relatively little attention in methodological discussions on Qualitative Comparative Analysis (QCA). This paper addresses this lacuna by discussing the challenges researchers face when transforming qualitative into quantitative data in QCA. By reviewing 29 empirical studies using qualitative data for QCA, we explore common practices related to data calibration, presentation and sensitivity testing. Based on these three issues, we provide considerations when using qualitative data for QCA, which are relevant both for QCA-scholars working with qualitative data and the wider mixed methods research community involved in quantitizing.

Keywords

Mixed methods research; Qualitative Comparative Analysis; Qualitative data; Quantitizing; Calibration; Sensitivity testing; Data presentation

Qualitative Comparative Analysis (QCA) is an approach that combines quantitative and qualitative research (Ragin, 1987, 2008; Rihoux & Ragin, 2009). Its "hybrid" nature (Cragun et al., 2016) adheres to the definition of mixed methods research (MMR) by Johnson et al. (2007: 129) as 'an intellectual and practical synthesis based on qualitative and quantitative research (…)'. QCA is a set-theoretical approach that identifies the (minimally) necessary and (minimally) sufficient (combinations of) conditions for an outcome. It does so by using Boolean and/or fuzzy-set algebra to treat cases as configurations of causal conditions and an outcome and by analyzing whether a given (combination of) condition(s) stand(s) in a subset or superset relationship to the outcome (Schneider & Wagemann, 2012). To this end, a study's so-called raw data – either quantitative, qualitative or both – need to be transformed; a process called calibration in QCA (Ragin, 2008: chapters 4 & 5). Calibration of qualitative data resembles what in the mixed methods literature is known as *quantitizing*, that is 'the numerical translation, transformation, or conversion of qualitative data'; a process that 'has become a staple of mixed methods research' (Sandelowski, Voils, & Knafl, 2009: 208).

Thirty years after Ragin (1987) introduced the approach in the social sciences, QCA is becoming a "mainstream" approach in several fields, such as sociology and political science (Rihoux, Álamos-Concha, Bol, Marx, & Rezsöhazy, 2013); in other (sub-)fields, such as health services research (Summers Holtrop, Potworowski, Green, & Fetters, 2016), it remains relatively novel, however. As an approach, QCA is still in development. Several of the current methodological discussions relate to MMR, such as the discussion regarding the (in)compatibility of regression analysis and QCA (Fiss, Sharapov, & Cronqvist, 2013; Thiem, Baumgartner, & Bol, 2016; xxxx).

This paper's three objectives focus on a series of related issues that so far have received relatively little attention in methodological discussions about QCA, and that may be particularly relevant for readers of *JMMR*. Our *first objective* is to explore how researchers currently use qualitative data in QCA.[1] Hereby we focus on three key issues: (a) data calibration; (b) data presentation,

and (c) sensitivity tests. To achieve this first objective, we review 29 QCA studies that use various types of qualitative data. Appendix I details the selection procedure of the included studies. The *second objective* is to contribute to the standards of good practice in QCA (Schneider & Wagemann, 2010). Therefore, we critically examine how the 29 articles deal with the three key issues (i.e. calibration, sensitivity and presentation) and provide considerations for researchers using qualitative data in QCA. Our *third objective* is to place our findings in the context of MMR. We focus particularly on the discussion about quantitizing, showing that our considerations provide relevant lessons for the wider mixed-methods research community.

How to Calibrate Qualitative Data in QCA?

An important issue in QCA is the calibration of the raw data. When using crisp-set QCA, all cases are either "in" (1) or "out" (0) of the sets. In fuzzy-set QCA (fsQCA), the raw data are calibrated from "fully in" (1) and "fully out" (0) of the sets, with additional gradations of set-membership (e.g., "almost fully in" [.83] or "more out than in" [.40]). The 1 and the 0 are two of the so-called qualitative thresholds; the crossover point at 0.5 is the third.

The literature on calibration is mainly concentrated on *quantitative* data. For example, Ragin (2008: chapter 5) focuses exclusively on this issue, while providing no practical advice for researchers on how to calibrate *qualitative* data. The same holds for Schneider & Wagemann (2012: 32-41).

The only two studies offering explicit methodological advice on how to calibrate *qualitative* data in QCA are Basurto & Speer (2012) and Tóth et al. (2017) (see xxxx for a more extensive discussion). Basurto & Speer (2012) propose a stepwise procedure to calibrate qualitative (interview) data into qualitative classifications with associated fuzzy-set values (see Appendix II). Tóth et al. (2017) introduce the so-called Generic Membership Evaluation Template (GMET) to assign membership scores to conditions based on qualitative data.[2] (see Appendix III).

Yet although Basurto & Speer (2012) and Tóth et al. (2017) provide valuable guidelines

on how to calibrate qualitative data, some important questions remain. It remains uncertain, for example, how qualitative data can properly inform and justify the determination of the qualitative thresholds – especially regarding the crossover point. What is more, while filling in the GMET is rather straightforward, decisions about how to attribute the final fuzzy set score remain somewhat subjective.

Both Basurto & Speer (2012: 169) and Tóth et al. (2017:195) note that existing studies using qualitative data in QCA are typically unclear about how exactly they calibrated their data. The studies usually are not transparent about: (1) where they placed the thresholds for inclusion and exclusion of a set (respectively the 1 and the 0); and (2) how they established the degree to which a case is "in" ($0.5 < x \leq 1$) or "out" of the set ($0 \leq x < 0.5$), that is, the degree of set-membership. Since results of a QCA analysis can differ substantively depending on researchers' specific choices on these issues, such transparency is important.

Determining the Thresholds for Inclusion and Exclusion of a Set

How did the studies we reviewed determine where to place the thresholds for inclusion and exclusion of a set? Table 1 summarizes the five main strategies employed, while Table A1 in Appendix IV provides additional details for all reviewed studies.

--- Table 1 about here ---

A first strategy is developing a rubric or coding scheme to assign codes for the outcome and the conditions. Chatterley et al. (2014), for example, develop a rubric to assign codes based on their data from interviews, focus group and observations (see Table A2 in Appendix V for an overview of the type of qualitative data used in all the reviewed studies). 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 inclusion and exclusion of a set. Kirchherr et al. (2016) base the calibration of some fuzzy-set values on existing quantitative indices and of other values on an iterative process of multiple semi-structured expert interviews and an online survey. While the

thresholds for inclusion and exclusion of a set are rather straightforward for data based on indices (e.g., a ranking is used), it is unclear how Kirchherr et al. (2016) determined thresholds based on the qualitative data.

A second strategy is suggested by Basurto & Speer (2012, see Iannacci & Cornford, 2017 for an application). Basurto & Speer (2012) construct two imaginary ideal cases, one representing full membership in a set (1) and one representing full non-membership (0). The thresholds for inclusion and exclusion of the set, then, are put in between the two "extreme" values.

A third - inductive - strategy that several studies adopt is to set the thresholds using QCA-software, particularly the threshold setter in Tosmana (Cronqvist, 2016).[3] Exploring the possibly large gaps in the data is another inductive strategy. Note that these inductive strategies are applicable only when the raw data are already numerical. Yet for a study based exclusively on qualitative data, these strategies are not an option; as a result, researchers are restricted in these cases to applying Tóth et al.'s (2017) GMET or using one of the first two strategies listed above.

Establishing the Degree of Set-Membership

While the overview in Table A1 in Appendix IV shows that almost all fsQCA-studies are careful about establishing the degree of set-membership,[4] it also reveals that many studies are not fully transparent regarding how the qualitative data were used to this end.

For example, Verweij (2015) used both qualitative and quantitative data to calibrate the outcome and the conditions. As with studies using a similar approach (e.g., xxx), the quantitative material "dominated" the calibration (i.e., it was the benchmark that could be adjusted based on the qualitative material). One of Verweij's (2015) conditions was calibrated based on various qualitative documents, with codes assigned using qualitative data-analysis software. The few small coding contradictions were then re-calibrated in a final step (Verweij, 2015: 1883). While the latter is common practice in QCA – as well as in many qualitative studies – it is not clear exactly how and why this re-calibration was done. As a consequence, it is difficult to replicate studies that

employ this approach. The same holds for Verweij, Klijn, Edelenbos, & Van Buuren's (2013) study, which used various qualitative sources to calibrate their outcome and conditions. In line with good QCA-practice, Verweij et al. (2013) published their coding scheme and the resulting scores in an appendix, allowing other researchers to assess whether 'the observations meaningfully capture the ideas contained in the concepts' (Adcock & Collier, 2001: 529) and thereby are "valid" (i.e., that a given value makes sense given existing empirical and theoretical knowledge). Yet, these tables do not include the reasoning behind the coding decisions, and therefore cannot be reproduced fully.

Similarly, Van der Heijden (2015) used a systematic coding scheme and qualitative data analysis software to explore data on voluntary environmental programmes systematically and gain insights 'into the "repetitiveness" and "rarity" of experiences shared by the interviewees, and those reported in the existing information studied' (p. 581). However, Van der Heijden (2015) did not discuss *how* this information was subsequently used to code cases as "in" or "out" of the set. Other studies go over the coding decisions only briefly (e.g., Chatterley et al., 2014) or provide no information on how the interview data were translated into the fuzzy set values (e.g., Basurto, 2013). To varying degrees, this lack of transparency inhibits the studies' replicability.

Some studies use multiple coders to establish the degree of set-membership. In Henik (2015), for example, two coders applied a coding rubric on transcribed interviews, with the average of these coders' scores constituting the final set attribute. Henik (2015: 445) notes that the coders 'agreed within 0.25 set membership points on more than 90% of the 960 items (…)'. However, it is unclear how *qualitatively important* differences across coders were addressed, existing when one coder codes an item as being "in" the set and the other as "out". In this regard, 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).

Regarding which values to assign to qualitative data (i.e. the quantitizing), the studies we reviewed offer some suggestions. Table 2 lists the strategies, while Table A1 in Appendix IV provides a more comprehensive overview.

--- Table 2 about here ---

One strategy is to directly ask interviewees to provide answers on a Likert-type scale (or one based on other pre-determined options). This strategy is applied by Fischer (2014), who calibrated his outcome (policy change) by asking approximately 250 interviewees to rate their perception of policy change from 1 to 5. Next, Fischer (2014: 350-351) averaged the perceptions of actors and subsequently calibrated these data into fuzzy sets by rescaling the average value to a 0–1 scale.[5] Another strategy is adopted by Kirchherr et al. (2016), who used a 4-value and 2-value coding scheme to assign set-membership scores to the attributes. Subsequently, Kirchherr et al. (2016) averaged the calibrated values for the different attributes of the conditions. They addressed this strategy's potential weakness, as it 'could introduce misfits between the verbal meaning of a concept and its operationalization' (Kirchherr et al., 2016: 39), by reviewing all averaged calibrations of the conditions and changing or recalibrating the attributes when they found that the conditions' values did not correspond to their averaged operationalization. Alternatives for taking the average value are substitutability (i.e., taking their maximum value) or taking the weakest link (i.e., the minimum value of the attributes of the concept) (Ragin, 2000, see Chatterley et al., 2014 and Basurto & Speer, 2012 for examples).

The Meaning of a Zero

A third challenge relating to calibration concerns the zero (0). Conceptually, in QCA the meaning of a zero is clear: fully out of a set. However, discussions among QCA-scholars reveal a challenge when coding qualitative data: how can one differentiate between concepts that are *truly absent* (i.e., where the concept is indeed absent) and which should thus be coded 0, and those concepts that are simply *not mentioned* in, for example, an interview? This question relates to Sandelowski et al.'s (2009: 217) observation in the context of quantitizing in MMR that *absent* may refer to different things in interview data: '(…)"it" (a) did not come up; (b) was not seen by the analyst; (c) was forgotten as a factor by the participant; (d) was thought by the participant to be so understood as to

not require bringing it up; (e) was a factor, but the participant did not want to bring "it" up; (f) was not brought up because the conversation veered away from "it"; and (g) truly was not a dimension of experience'. This challenge holds not only for other types of qualitative data, such as existing documents or archive material, but also for quantitative data. If a concept is not mentioned in a document, does that mean that it is absent, or just that no information on it is included in the document? Data triangulation is one way to assess the likelihood of these two possibilities. In a QCA analysis, it will oftentimes be useful to explore the zeros in more detail to find out why the condition was absent or why the information was missing.

The large majority of the reviewed studies (n=25) do *not* discuss the meaning of the zero. There can be several reasons for this. First, sufficient information was available to assign "truly absent zeros" to cases. For example, Van der Heijden (2015) reported that he ensured sufficient information on all attributes by first gathering information from websites and reports and then filling in gaps using interview data (Crowley 2012 is another example). A second reason may be that researchers did not differentiate between "truly absent" and "not mentioned". For example, when calibrating their outcome "American states' levels of environmental justice policy", Kim & Verweij (2016) assigned a zero both to states with either "no action" or "no information", which is conceptually problematic.[6] Vergne & Depevre (2016) decided to ask people to not complete their survey when they were not knowledgeable enough, thus circumventing the problem of missing data; however, they also reported that they turned to additional databases when data about a specific attribute was missing, but also noted that sometimes, they did not find more information.

How to Present the Calibration Process and the Data?

To make studies replicable, the data sources and calibration process need to be presented transparently and comprehensively (Gerring, 2012). Ideally, this should also be done concisely, to make the material easily accessible. These goals – transparency and comprehensiveness versus conciseness – often conflict. What is more, even transparency and comprehensiveness may

conflict, as researchers aiming to be comprehensive risk burying their readers in details, thereby hindering transparency. How QCA scholars present the calibration process, and hence the actual possibility for replication, varies strongly across the reviewed studies. Table 3 summarizes the material from Table A1 in Appendix IV on this.

--- Table 3 about here ---

Table 3 demonstrates that most reviewed studies (n= 27) provide *some* information on the calibration procedure (Aversa, Furnari, & Haefliger, 2015; and Crowley, 2012 provide too little information). Numerous studies provide substantial information, but not all that would be required for full transparency.

Some studies' data calibration procedures make them easier to replicate than others. Kim & Verweij (2016), for example, included a table with the motivation of the assignment of US states to a specific category based on a combination of descriptions and secondary survey data. Fischer (2014) presented the calibration of outcome and conditions in tables in appendices. Both studies use a rather straightforward approach to calibration by respectively referring to survey results and directly asking interviewees to "score" their outcome and conditions, subsequently taking the average. Hence, replicating these findings is also rather straightforward.

Arriving at similar results becomes more complicated when the data needed for a specific attribute cannot be directly derived from interviewees' answers. While journal space limitations often make the disclosure of all details of the calibration process challenging, using (online) appendices, an option available at a growing number of journals, is one way to give more insight in the argumentation of researchers (Basurto & Speer, 2012). This suggestion is taken up by a variety of the reviewed studies (Basurto, 2013; Fischer, 2014; Kirchherr et al., 2016; Thomann, 2015; Wang, 2016).

Which Sensitivity Tests to Conduct?

Testing findings' robustness by means of sensitivity analyses should be part of a good QCA study

(Schneider & Wagemann, 2012). The methodological literature on QCA pays increasing attention to sensitivity tests (Baumgartner & Thiem, 2017a; Marx, 2010; Skaaning, 2011; Thiem, 2014; Thiem, Baumgartner, et al., 2016), including how to deal with different types of errors (Maggetti & Levi-Faur, 2013). In addition, the literature criticizing QCA (e.g., Hug, 2013; Lucas & Szatrowski, 2014; Paine, 2016) regularly indicates that the alleged lack of findings' robustness is a key problem (but see Baumgartner & Thiem, 2017).

The QCA literature provides several suggestions on how to assess the robustness of QCA findings using sensitivity tests. A non-exhaustive list includes: (1) dropping or adding cases and conditions; (2) changing fuzzy-set membership functions; (3) altering consistency thresholds (Schneider & Wagemann, 2012; Thiem, 2014; Thiem, Spöhel, & Duşa, 2016); (4) changing the definitions of the set values; (5) using alternative measures for a concept (Basurto & Speer, 2012); (6) changing the calibration thresholds of raw data into set-membership; and (7) altering the frequency of cases linked to configurations (Skaaning, 2011). These suggestions are not specific to 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., 2017; Kim & Verweij, 2016). Similarly, changing the frequency of cases linked to the configuration can be done irrespective of the kind of data used. Still, the higher the number of cases, the more appropriate this sensitivity test becomes. Since studies using qualitative data often – though not always – have a relatively low number of cases, this will in many cases not be the most important sensitivity test to conduct. Some researchers conduct additional statistical analyses to assess the robustness of their findings, despite criticism about the comparability of the two methods (e.g., Thiem et al., 2016). For example, Hodson et al. (2006) investigated whether their QCA-generated configurations were associated with the outcome and whether the association was statistically significant. Hodson et al. (2006) also introduced multivariate controls by creating dummy variables specifying key configurations and including them in a linear model. Note that while combining QCA and statistical analyses might be of interest to the readership of

JMMR, we do not discuss this further since it is not specific to QCA studies using qualitative data.

Based on the reviewed literature, we selected those sensitivity tests that are relevant for QCA studies using qualitative data. We list these in Table 4. Table A1 in Appendix IV provides a more detailed overview for all reviewed studies.

--- Table 4 about here ---

First, the available qualitative data can be a strong motivator to decide which cases to drop or add in the sensitivity analysis. Dropping cases can be a useful way to assess findings' robustness. Kirchherr et al. (2016), for example, included an extensive section on robustness in which they motivate their choices to exclude certain cases based on case descriptions presented in an appendix. However, when dropping cases, it is important to make sure that the cases-to-conditions ratio is still acceptable – typically one condition to three cases (Marx, 2010). If this ratio becomes too low, the results become unreliable.

A second type of sensitivity test is conducted by altering the different attributes of the condition (Kirchherr et al., 2016), for example to base the membership score on only one attribute rather than multiple ones. Here as well, the motivation for such choices must be based on knowledge about case context (e.g., that the now omitted attributes introduced noise to the condition's operationalization). Another related option is to replace the condition by one of its attributes, a decision that can, for example, be based on the importance assigned to the specific attribute in the interviews, relevant documents or literature.

Another type of test, which we subsume here under the heading of sensitivity tests but which is technically a test to better determine which factors or mechanisms "drive" the outcome, is conducted by Tóth et al. (2017: 202), who follow Fiss (2011). A new outcome is introduced that is more extreme than the original (in Tóth et al., 2017: *very high* relational attractiveness of the customer [RAC]). The qualitative threshold (the "anchor point", in Tóth et al.'s (2017) terminology) 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 returning to the qualitative data and assigning appropriate (fuzzy) set values, where the calibration of the original outcome can be used as a starting point.

Considerations When Using Qualitative Data in QCA

Based on the studies we reviewed, we highlight five considerations for using qualitative data for QCA (summarized in Table A3 in Appendix VI). First, *QCA-researchers should be more explicit about how they arrive at certain thresholds for inclusion and exclusion of a set*. Depending on the type of data (to be) collected, these thresholds might, for example, be determined by constructing an imaginary ideal case, or be based on a classification of interview responses.

Second, researchers *should be more explicit about how they determined the degree of set-membership*. More specifically, the reasoning behind the coding of qualitative data and the subsequent translation of qualitative codes into fuzzy-set scores should be more clearly communicated in articles or (online) appendices (see also point four below). Qualitative data or codes can be linked to values on a Likert-type or other pre-determined numerical scale (potentially based on quantitative material) and subsequently translated into fuzzy-set values. Moreover, rubrics or coding schemes (e.g., with two or four values) or pre-determined qualitative classifications can be used as an intermediate step for assigning fuzzy-set values to qualitative data.

Third, *QCA researchers should pay more attention to the zeros in their calibrated data*. Crucially, they must be careful about distinguishing between 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) and those where outcomes are coded zero because they are "truly absent". To avoid this ambiguity when using interview data, researchers should attempt to construct their interview scheme such that all concepts are addressed during the interview (although Sandelowski et al.'s (2009) option – that the analyst did not see "it", even though it was there – would then still be a possibility). Creating a separate section for each condition and the outcome

in the interview guideline, as proposed by Basurto & Speer (2012), is one possibility to doing so. The same holds for Tóth et al.'s (2017) suggestion to draw up an initial template based on previous literature. When all concepts are addressed in an interview, a value of "0" would then be assigned only to attributes or conditions 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 attributes or conditions cannot always be avoided.

A similar data deficiency can also occur when analysing pre-existing data for QCA. We provide two options to deal with such data gaps. First, in cases where such an approach is possible, interviewees can be re-contacted about the attributes or conditions for which information is missing. This is the ideal solution, since it allows researchers to establish whether it was indeed absent, or whether it was just not mentioned in the initial interview. When it is not possible to go back to the interviewees, however – for example because of practical constraints –, a second-best option is to conduct sensitivity analyses. Three sensitivity analyses are particularly apt for addressing the zero-issue: (1) removing the conditions where this problem occurs and assessing the effect; (2) assigning the value "0.51" (i.e., just "in" the set) to cases of which the researcher is not sure whether the condition is "truly" absent to differentiate between the two findings; and (3) excluding the cases where the concept is "not mentioned" from the analysis.

Fourth, to increase a study's transparency and comprehensiveness, and hence its replicability, *QCA researchers should explicitly delineate the choices they made* (to the extent that this is possible given issues of, for example, confidentiality). We agree with Schneider & Wagemann's (2010) advice to publish the raw data matrix in addition to a detailed discussion of the calibration of the set membership scores. When a data set is too large to be published, the original data should be made available on the Internet or on demand. Large datasets, including transcribed interviews and reports, often exist when using qualitative data for QCA. In order to present 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.

Finally, our review showed that although *conducting sensitivity tests in (qualitative) QCA* should be common practice, this is still not the case. Various tests are particularly suited to dealing with qualitative data, such as changing the number of cases, altering the conditions, or re-running the analysis with a more extreme outcome.

Transforming Qualitative into Quantitative Data in QCA: What Lessons for Mixed-Methods Research?

The considerations in the previous section are first and foremost meant for QCA-researchers using qualitative data. However, as Cragun et al. (2016) show, QCA's hybrid nature offers several advantages over other methods and is therefore interesting for mixed methods researchers more generally.

Our considerations regarding *calibration* specifically relate to the discussions in *JMMR* on quantitizing. Discussions have been held about 'the foundational assumptions, judgments, and compromises involved in converting qualitative into quantitative data (…)' (Sandelowski et al. 2009: 208), for example on what and how to count. Debates about how to quantitize qualitative data are not new to MMR (e.g., Boyatzis, 1998), and the topic is usually included in MMR text books (e.g., Miles, Huberman, & Saldaña, 2014). Typically, as in Teddlie & Tashakkori (2009: 270-271), examples are presented as to how qualitative data have been quantitized, or on how researchers have generally proceed, for instance by Sandelowski et al. (2009: 218): 'A common approach to quantitizing is to use the results of a prior quantitative analysis of quantitative data as the framework for the conversion of qualitative into quantitative data. This framework provides the decision rules for a directed form of content analysis whereby a priori codes are derived from a quantitative data set and applied to a qualitative data set (…)'. However, as with the studies reviewed above, the more detailed choices made by researchers frequently go undiscussed, alongside their underlying reasoning. Consequently, the methodological MMR literature provides little guidance for researchers seeking to quantitize their qualitative data. Since such choices may also

influence the substantive results of an MMR study, they must be clearly communicated. What is more, the transparency and hence replicability of MMR studies would increase if they were more explicit about the choices made and the reasoning underlying these choices regarding quantitizing.

Conversions from qualitative into quantitative data 'are by no means transparent and uncontentious' (Love, Pritchard, Maguire, McCarthy, & Paddock, 2005: 287 in Sandelowski et al., 2009). Our considerations regarding the *presentation* of the calibration process increase the transparency and replicability of studies where quantitization is used.

Given that quantitizing in MMR is to some extent subjective, it is relevant for MMR to conduct *sensitivity tests* to assess the robustness of the findings. Some of the sensitivity tests that we identified as relevant for QCA using qualitative data are also relevant for MMR that includes quantitizing; this is especially the case for studies in which the (in)dependent variables (conditions) include several sub-dimensions (attributes). Specifically, three of the sensitivity tests mentioned above are particularly appropriate to MMR: dropping or adding cases based on extensive case knowledge; altering the attributes of a condition based on knowledge of the case context; and replacing conditions by one of their attributes.

Considerations on quantitizing beyond the QCA literature

Although this paper focused on QCA studies, research using methodologies other than QCA also provide valuable insights about quantitization. This can be illustrated using examples from various scientific fields. In education research, the study of Gilmore, Maher, Feldon, & Timmerman (2014) quantitized data from 65 interviews to assess the relationship of participants' teaching experiences and teaching support systems with changes in their teaching orientation over time. They covered this longitudinal aspect by calculating the changes in coding scores between pre- and post-interviews. Moyer-Packenham et al. (2016) conducted pre- and post-assessments of quantitized video data when studying the role of affordances in children's learning performance.

As their study makes clear, using quantitized codes derived from sources based on different points in time is a useful consideration when investigating developments over time.

When considering on how to deal with zeros in the data, Gilmore, Maher, Feldon, & Timmerman (2014) suggest using multiple imputation procedures to fill the missing data. In the area of health research, Chang, Voils, Sandelowski, Hasselblad, & Crandell (2009) describe how qualitative labels for the number of respondents per specific finding on antiretroviral adherence – such as "few" or "many" – can be transformed in exact numbers – such as 2 or 50. They conducted an online survey at nursing school faculty to obtain lower and upper limits for specific verbal labels, and subsequently used the responses in regression analyses to estimate a plausible range of respondents in a given study. Sandelowski (2000), in turn, uses the study of Borkan, Quirk, & Sullivan (1991) as an example of quantitizing. In this study, the researchers use narrative analysis to determine the main categories of how elderly people viewed the hip fractures from which they suffered. A series of reliability tests were then conducted to ensure the consistency of the categories. Both studies provide additional insights on the issue of how to establish the degree of set-membership.

An example from economics comes from Vaitkevicius (2013), who suggests a systematic coding procedure based on hermeneutics to code qualitative data and subsequently analyze these data quantitatively. This procedure is, for instance, applicable to code and analyze closed-ended and open-ended questions. A final example also proposes a procedure for open-ended – qualitative – survey questions. Rohrer, Brummer, Schmukle, Goebel, & Wagner (2017) suggest the employment of tools from natural language processing to process and analyze potentially large numbers of answers to open ended questions. They demonstrate their procedure by analyzing the more than 35,000 answers to the question "What else are you worried about?" from the participants of a German socio-economic panel study. These examples can be used as a starting point for expanding the list of considerations to be reflected upon in mixed methods research.

Conclusion

This paper addressed the challenges that researchers face when using qualitative data in QCA, especially when it comes to transforming it into quantitative data. Although QCA training courses are offered worldwide and several textbooks and journal articles that include hands-on instructions have been published, specific guidance for the use of qualitative data in QCA has been largely absent. We addressed this lacuna by exploring the various ways in which researchers currently use qualitative data in QCA and by laying considerations on three key issues: (1) the calibration of qualitative data (known as quantitization in MMR); (2) the presentation of the calibration process and the data, and (3) sensitivity testing. Overall, our study demonstrates that many QCA-studies using qualitative data are not as transparent in their procedures as would be required to enable proper replicability.

We thus presented five main considerations for QCA researchers aiming to enhance their studies' transparency: first, researchers should be more explicit as to how they arrive at the thresholds for inclusion and exclusion of a set; second, they should be clear about how they determined the degree of set-membership; third, more attention should be paid to the "zeros" in the calibrated data; fourth, researchers should make more explicit and present clearly the choices they made during the calibration process; and finally, conducting sensitivity tests should become common practice. These considerations contribute to the methodological discussions on data calibration and quantitization. Moreover, our study provides QCA users, and readers of *JMMR* more generally, with ideas about how to transform qualitative data into quantitative form in their empirical studies. Which consideration(s) a given researcher ultimately takes into account will depend, among other things, on the specific research question, the type of data, and available time and resources.

Tables

Table 1. Different Strategies to Determine the Thresholds for Inclusion and Exclusion of a Set.

Strategy	Examples
Develop a rubric/coding scheme to assign codes to outcome and conditions.	(Chatterley et al., 2014; Chatterley, Linden, & Javernick-Will, 2013; Fischer, 2015; Henik, 2015; Iannacci & Cornford, 2017; Kirchherr et al., 2016)
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 inclusion and exclusion are then placed somewhere in-between these values.	(Basurto & Speer, 2012; Iannacci & Cornford, 2017)
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., 2017)
Conduct a cluster analysis by using, for example, Tosmana (Cronqvist, 2016).	(Kim & Verweij, 2016; Li, Kopennjan, & Verweij, 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)

Table 2. Different Strategies to Determine the Degree of Set Membership.

Strategy	Examples
Use pre-determined options in an interview (e.g. Likert scale)	(Fischer, 2014)
Use a coding scheme (e.g., 4-value and 2-value fuzzy sets) to assign membership scores to attributes and subsequently: a. Average the calibrated values. b. Take the minimum value (when all attributes of a concept are necessary). c. Take the maximum value (when all attributes are sufficient).	(Kirchherr et al., 2016) (Chatterley, 2014) (Basurto & Speer, 2012)

Table 3. Different Strategies to Present the Calibration Procedure.

Approach	Examples
Table in main text, full information	(Kirchherr et al., 2016; Tóth et al., 2017 [for 1 GMET])
Table in main text, partial information	(Basurto, 2013; Chai & Schoon, 2016; Chatterley et al., 2014, 2013; Crilly, 2011; Hodson & Roscigno, 2004; Hodson, Roscigno, & Lopez, 2006; Iannacci & Cornford, 2017; Kim & Verweij, 2016; Li et al., 2016; Metelits, 2009; Summers Holtrop et al., 2016; Vergne & Depeyre, 2016; Verweij, 2015; Verweij & Gerrits, 2015; Verweij et al., 2013)
Text boxes	(Basurto & Speer, 2012; Mishra et al., 2017)
Discussed in words in main text, typically partial	(Chai & Schoon, 2016; Chatterley et al., 2013; Crilly, 2011; Henik, 2015; Iannacci & Cornford, 2017; Kim & Verweij, 2016; Kirchherr et al., 2016; Li et al., 2016; Verweij, 2015)
Discussed in words in appendix, typically partial	(Smilde, 2005; Vergne & Depeyre, 2016)
Table(s) in appendix, full information	(Fischer, 2014, 2015; Iannacci & Cornford, 2017; Kirchherr 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; Hodson & Roscigno, 2004)

Table 4. Relevant Sensitivity Tests for Assessing the Robustness of QCA-Findings 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).	(Fiss, 2011; Tóth et al., 2017)

Acknowledgments

Some of the ideas in this article have been presented at the 4th International QCA Expert Workshop in Zurich, Switzerland in 2016. We would like to thank all participants for their useful comments and suggestions. Additionally, we thank Claude Rubinson, Federico Iannacci, Eva Thomann, Zsofia Tóth and Peter Feindt for their constructive feedback. Barbara Vis' research was funded by a VIDI grant from the Netherlands Organization for Scientific Research (NWO, grant nr. 452-11-005).

References

Adcock, R., & Collier, D. (2001). Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *The American Political Science Review*, *95*(3), 529–546.

Aversa, P., Furnari, S., & Haefliger, S. (2015). Business Model Configurations and Performance: A qualitative Comparative Analysis in Formula One Racing, 2005–2013. *Industrial and Corporate Change*, *24*(3), 655–676.

Basurto, X. (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, X., & Speer, J. (2012). Structuring the Calibration of Qualitative Data as Sets for Qualitative Comparative Analysis (QCA). *Field Methods*, *24*(2), 155–174.

Baumgartner, M., & Thiem, A. (2017a). Model Ambiguities in Configurational Comparative Research. *Sociological Methods & Research*, *46*(4), 954–987.

Baumgartner, M., & Thiem, A. (2017b). Often Trusted But Never (Properly) Tested: Evaluating Qualitative Comparative Analysis. *Sociological Methods & Research*. http://doi.org/10.1177/0049124117701487

Borkan, J. M., Quirk, M., & Sullivan, M. (1991). Finding Meaning after the Fall: Injury Narratives from Elderly Hip Fracture Patients. *Social Science & Medicine*, *88*(8), 947–957.

Boyatzis, R. E. (1998). *Transforming Qualitative Information: Thematic Analysis and Code Development*. Thousand Oaks: Sage.

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

Chang, Y., Voils, C. I., Sandelowski, M., Hasselblad, V., & Crandell, J. L. (2009). Transforming

Verbal Counts in Reports of Qualitative Descriptive Studies Into Numbers. *Western Journal of Nursing Research*, *31*(7), 837–852.

Chatterley, C., Javernick-Will, A., Linden, K. G., Alam, K., Bottinelli, L., & Venkatesh, M. (2014). A Qualitative Comparative Analysis of Well-Managed School Sanitation in Bangladesh. *BMC Public Health*, *14*(6), 1–14.

Chatterley, C., Linden, K. G., & Javernick-Will, A. (2013). Identifying Pathways to Continued Maintenance of School Sanitation in Belize. *Journal of Water, Sanitation and Hygiene for Development*, *3*(3), 411–422.

Cragun, D., Pal, T., Vadaparampil, S. T., Baldwin, J., Hampel, H., & DeBate, R. D. (2016). Qualitative Comparative Analysis: A Hybrid Method for Identifying Factors Associated With Program Effectiveness. *Journal of Mixed Methods Research*, *10*(3), 251–272.

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

Cronqvist, L. (2016). Tosmana. Trier: University of Trier. Retrieved from http://www.tosmana.net.

Crowley, M. (2012). Control and Dignity in Professional, Manual and Service-Sector Employment. *Organization Studies*, *33*(10), 1383–1406.

Fischer, M. (2014). Coalition Structures and Policy Change in a Consensus Democracy. *Policy Studies Journal*, *42*(3), 344–366.

Fischer, M. (2015). Institutions and Coalitions in Policy Processes: A Cross-sectoral Comparison. *Journal of Public Policy*, *35*(2), 245–268.

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

Fiss, P. C., Sharapov, D., & Cronqvist, L. (2013). Opposites Attract? Opportunities and Challenges for Integrating Large-N QCA and Econometric Analysis. *Political Research Quarterly*, *66*(1), 191–198.

Gerring, J. (2012). *Social Science Methodology: A Unified Framework*. Cambridge: Cambridge University Press.

Gilmore, J., Maher, M. A., Feldon, D. F., & Timmerman, B. (2014). Exploration of Factors Related to the Development of Science, Technology, Engineering, and Mathematics Graduate Teaching Assistants' Teaching Orientations. *Studies in Higher Education*, *39*(10), 1910–1928.

Goertz, G., & Mahoney, J. (2012). *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton and Oxford: Princeton University Press.

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

Hodson, R., & Roscigno, V. J. (2004). The Organizational and Social Foundations of Worker Resistance. *American Sociological Review*, *69*(1), 14–39.

Hodson, R., Roscigno, V. J., & Lopez, S. H. (2006). Chaos and the Abuse of Power: Workplace Bullying in Organizational and Interactional Context. *Work and Occupations*, *33*(4), 382–416.

Hug, S. (2013). Qualitative Comparative Analysis: How Inductive Use and Measurement Error Lead to Problematic Inference. *Political Analysis*, *21*(2), 252–265.

Iannacci, F., & Cornford, T. (2017). Unravelling Causal and Temporal Influences Underpinning Monitoring Systems Success: A Typological Approach. *Information Systems Journal*, 1–24. http://doi.org/10.1111/isj.12145

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a Definition of Mixed Methods Research. *Journal of Mixed Methods Research*, *1*(2), 112–133.

Kim, Y., & Verweij, S. (2016). Two Effective Causal Paths that Explain the Adoption of US State Environmental Justice Policy. *Policy Science*, *49*(4), 505–523.

Kirchherr, J., Charles, K. J., & Walton, M. J. (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, Y., Kopennjan, J., & Verweij, S. (2016). Governing Environmental Conflicts in China: Under What Conditions do Local Governments Compromise? *Public Administration*, *94*(3), 806–822.

Lucas, S. R., & Szatrowski, A. (2014). Qualitative Comparative Analysis in Critical Perspective. *Sociological Methodology*, *44*(1), 1–79.

Maggetti, M., & Levi-Faur, D. (2013). Dealing with Errors in QCA. *Political Research Quarterly*, *66*(1), 198–204.

Marx, A. (2010). Crisp-set Qualitative Comparative Analysis (csQCA) and Model Specification: Benchmarks for Future csQCA Applications. *International Journal of Multiple Research Approaches*, *4*(2), 138–158.

Metelits, C. M. (2009). The Consequences of Rivalry: Explaining Insurgent Violence Using Fuzzy Sets. *Political Research Quarterly*, *62*(4), 673–684.

Miles, M. B., Huberman, M., & Saldaña, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook (3rd edition)*. Thousand Oaks: SAGE.

Mishra, A., Ghate, R., Maharjan, A., Gurung, J., Pathak, G., & Upraity, A. N. (2017). Building Ex Ante Resilience of Disaster-Exposed Mountain Communities: Drawing Insights from the Nepal Earthquake Recovery. *International Journal of Disaster Risk Reduction*, *22*, 167–178.

Moyer-Packenham, P. S., Bullock, E. K., Shumway, J. F., Tucker, S. I., Watts, C. M., Westenskow, A., … Jordan, K. (2016). The Role of Affordances in Children's Learning Performance and Efficiency When Using Virtual Manipulative Mathematics Touch-screen Apps. *Mathematics Education Research Journal*, *28*(1), 79–105.

Paine, J. (2016). Set-Theoretic Comparative Methods: Less Distinctive Than Claimed. *Comparative Political Studies*, *49*(6), 703–741.

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

Ragin, C. C. (2000). *Fuzzy-Set Social Science*. Chicago and London: The University of Chicago Press.

Ragin, C. C. (2008). *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago and London: University of Chicago Press.

Richards, L. (2005). *Handling Qualitative Data: A Practical Guide*. London etc.: SAGE Publications.

Rihoux, B., Álamos-Concha, P., Bol, D., Marx, A., & Rezsöhazy, I. (2013). From Niche to Mainstream Method? A Comprehensive Mapping of QCA Applications in Journal Articles from 1984 to 2011. *Political Research Quarterly*, *66*(1), 175–184.

Rihoux, B., & Ragin, C. C. eds. (2009). *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques*. Los Angeles etc.: Sage.

Rohrer, J. M., Brummer, M., Schmukle, S. C., Goebel, J., & Wagner, G. G. (2017). "What Else Are You Worried About?" Integrating Textual Responses into Quantitative Social Science Research. *PLoSONE*, *12*(7), e0182156.

Sandelowski, M. (2000). Combining Qualitative and Quantitative Sampling, Data Collection, and Analysis Techniques in Mixed-Method Studies. *Research in Nursing & Health*, *23*(3), 246–255.

Sandelowski, M., Voils, C. I., & Knafl, G. (2009). On Quantitizing. *Journal of Mixed Methods Research*, *3*(3), 208–222.

Schneider, C. Q., & Wagemann, C. (2012). *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. Cambridge: Cambridge University Press.

Skaaning, S. E. (2011). Assessing the Robustness of Crisp-Set and Fuzzy-Set QCA Results. *Sociological Methods & Research*, *40*(2), 391–408.

Smilde, D. (2005). A Qualitative Comparative Analysis of Conversion to Venezuelan Evangelicalism: How Networks Matter. *American Journal of Sociology*, *111*(3), 757–796.

Summers Holtrop, J., Potworowski, G., Green, L. A., & Fetters, M. (2016). Analysis of Novel Care Management Programs in Primary Care: An Example of Mixed Methods in Health Services Research. *Journal of Mixed Methods Research*.

http://doi.org/10.1177/1558689816668689

Teddlie, C., & Tashakkori, A. (2009). *Foundations of Mixed Methods Research: Integrating Qualitative and Quantitative Approaches in the Social and Behavioral Sciences*. Thousand Oaks: Sage.

Thiem, A. (2014). Membership Function Sensitivity of Descriptive Statistics in Fuzzy-set Relations. *International Journal of Social Research Methodology*, *17*(6), 625–642.

Thiem, A., Baumgartner, M., & Bol, D. (2016). Still Lost in Translation! A Correction of Three Misunderstandings Between Configurational Comparativists and Regressional Analysts. *Comparative Political Studies*, *49*(6), 742–774.

Thiem, A., Spöhel, R., & Duşa, A. (2016). Enhancing Sensitivity Diagnostics for Qualitative Comparative Analysis: A Combinatorial Approach. *Political Analysis*, *24*(1), 104–120.

Thomann, E. (2015). Customizing Europe: Transposition as Bottom-up Implementation. *Journal of European Public Policy*, *22*(10), 1368–1387.

Tóth, Z., Henneberg, S. C., & Naudé, P. (2017). Addressing the "Qualitative" in fuzzy set Qualitative Comparative Analysis: The Generic Membership Evaluation Template. *Industrial Marketing Management*, *63*(May), 192–204.

Vaitkevicius, S. (2013). Rethinking the Applicability of Hermeneutic Systems for Coding and Statistical Analysis of Authorial Intentions in Economics. *Inzinerine Ekonomika-Engineering Economics*, *24*(5), 415–423.

Van der Heijden, J. (2015). What Roles Are There for Government in Voluntary Environmental Programs? *Environmental Policy and Governance*, *25*(5), 303–315.

Vergne, J.-P., & Depeyre, C. (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, S. (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), 1877–1887.

Verweij, S., & Gerrits, L. M. (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, S., Klijn, E.-H., Edelenbos, J., & Van Buuren, A. (2013). What Makes Governance Networks Work? A Fuzzy Set Qualitative Comparative Analysis of 14 Dutch Spatial Planning Projects. *Public Administration*, *91*(4), 1035–1055.

Wang, W. (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.

NB: 3 References removed for review purposes.

Appendix I:

Selection Procedure of Studies Included in the Review

Our criteria for selecting QCA studies using qualitative data were: applying a QCA analysis, using qualitative data, refereed journal articles, English language. To find the studies that meet these criteria, we used a variety of search strategies. We consulted the bibliography on the COMPASSS website, which is a worldwide network of scholars and practitioners working with QCA (www.compasss.org, last accessed November 2016). We examined the articles' potential relevance based on mentioning the use of qualitative data in the titles and/or abstracts. When considered relevant, we read the methods section to see whether qualitative data had been used. This search process led to the selection of three papers. Additionally, we used Scopus to find articles that referenced one of the few methodological studies on how to use qualitative data in QCA: Basurto & Speer (2012) (n=10, accessed on October 20, 2016). We selected four relevant ones, using the same strategy as with the COMPASSS bibliography. A similar search on ISI Web of Science yielded no additional articles. We further determined the relevance of the seven articles discussed by the other methodological study on how to use qualitative data in QCA: Tóth, Henneberg, & Naudé (2017). This resulted in three additional papers. Finally, we derived 19 papers based on references in already selected papers (i.e., snowballing) and through suggestions for relevant articles from our network. This search process resulted in a total of 29 articles.

Appendix II

Own summary of Basurto & 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 III

Own summary of the steps in Tóth et al.'s (2017) 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. (2017) 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. (2017) but are derived from their Table 2 on p. 197). We would strongly recommend making 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. (2017) 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. 197)

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. (2017), these include for example "professional trust" and "frequent conflicts" (p. 197).

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. 197);

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. (2017), 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. 197).

Appendix IV

Table A1. Overview of the Studies Using Qualitative Data for QCA Based on Our Literature Review

Author(s)	CALIBRATION			PRESENTATION	SENSITIVITY
	How is the threshold for inclusion and exclusion of a set determined?	**How is the degree of set-membership established?**	**How is differentiated between "truly absent" and "not mentioned" indicators?**	**How is the calibration procedure presented?**	**Which sensitivity tests are conducted?**
Aversa et al.(2015)	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, it is not clear what are the sets (conditions). Probably the ones listed in Table 5 and 6, but the calibration of these conditions is not discussed.	NA (csQCA)	Not discussed	Besides some information on the calibration of the outcome, calibration of the conditions is not discussed and/ or presented.	None
Basurto (2013)	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 (expressed in percentages) or, in case of multiple measures constituting one condition, on averaging the measures.	Not discussed	A table in the main text states the types of states per condition (e.g., four-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 of all conditions and the outcome.	None
Basurto & Speer (2012)	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. Values for one condition obtained by taking the maximum value of the sub-measures, since they are substitutable.	Not discussed	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.	None
Chai & Schoon (2016)	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, there is no point in having a fuzzy outcome .	NA (csQCA)	Not discussed	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 in the main text.	None
Chatterley et al. (2013)	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.	NA (csQCA)	Not discussed	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	None

				text.	
Chatterley et al. (2014)	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.	Not discussed	A coding rubric is presented with qualitative descriptions representing the four-value fuzzy-set for each condition and the outcome. Some of the values are supported by quotes in the text.	None
Crilly (2011)	The outcome's thresholds are based on interviews, as are the values in-between (four-value fuzzy-set). The decision is explained clearly and illustrated with an example from the interviews.	The calibration of the seven conditions in four 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 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).	Not discussed	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 the main text.	The author followed Epstein et al. (2008) and reran the analysis with a reduced consistency threshold of 0.85 (p. 712).
Crowley (2012)	A coding instrument for the workplace ethnographies is developed by four researchers and adjusted on the basis of eight (out of 154) workplace ethnographies. The codes for the >10 conditions and outcomes are displayed in a table and include Likert scales (1-none 2-little 3-average 4-high 5-very high) and present/absent scoring. How the thresholds have been established for the dichotomous conditions and outcomes is clear; for the Likert-ones, there is no information.	It is not clear what type of QCA has been used. In any case, there is no information on how the Likert-conditions and outcomes (see previous column) have been calibrated into fuzzy or crisp sets. A link to additional information is mentioned in a note, but this link does not work.	NA (information on all conditions and the outcome available)	There is no information on this in the main text or in an appendix.	None
Fischer (2014)	For two of the three conditions, Fischer takes the observed maximum (1.0) and observed minimum values (0), and uses the median observed value 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 into a fuzzy-value using the direct method of calibration.	Not discussed	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.	None
Fischer (2015)	The thresholds are determined using the substantive knowledge from the	For assigning values to one of the conditions, the author uses the direct method of	Not discussed	Tables with how the outcome and conditions were calibrated and	None

	qualitative material. A coding rubric, including a description for determining the three thresholds, is presented in the appendix.	calibration. For the other two conditions, he uses a 7-value fuzzy-set, whereby he avoids assigning the score 0.5 to cases.		what were the resulting membership scores are presented in an appendix. The main text includes a table with the fuzzy set data.	
Henik (2015)	A coding rubric is presented with qualitative descriptions representing the four or two- fuzzy-values for each condition and the outcome. NB: The calibration scheme includes 0.5, which should be avoided.	A coding rubric is applied on the interview transcripts by 2 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).	Not discussed	The coding rubric is included in the main text.	None
Hodson & Roscigno (2004)	The authors use csQCA. They use their qualitative data to code the cases as being "in" (1.0) or "out" (0) of a set using different categories for the concepts, e.g. average or less versus more than average, or no versus yes.	NA (csQCA)	Not discussed	The binary coding categories for the concepts and outcome are presented in tables in the main text and appendix. A footnote in the main text indicates that the code sheet, coding protocol and data are available on a website, but this link is not/ no longer valid. As such, it is unclear how the qualitative data are assigned to the categories.	The authors only discuss how they addressed the sensitivity of the coding exercise (i.e. by recoding a 10% sample of cases as a reliability check) and not the QCA analysis.
Hodson et al. (2006)	The authors use csQCA. They use their qualitative data to code the cases as being "in" (1.0) or "out" (0) of a set using different categories for the concepts and outcome, e.g. adequate or less versus good or exceptional.	NA (csQCA)	Not discussed	The binary coding categories for the concepts and outcome are presented in tables in the main text. A note indicates that the codesheet, coding protocol and data are available on a website, but this link is not/ no longer valid. As such, it is unclear how the qualitative data are assigned to the categories.	The authors ask whether each QCA-generated configuration is associated with the outcome and whether the association is statistically significant. They also introduce multivariate controls. This is accomplished by creating dummy variables specifying key configurations and entering these into a general linear model with appropriate controls.

					This multivariate analysis assists in evaluating the robustness of the configurational findings and also in evaluating possible alternative explanations for the outcome.
Iannacci & Cornford (2017)	Using the proposal by Basurto & Speer (2012) to formulate ideal cases or types at the extremes, to determine the (1) and the (0).	As "deviations" from the ideal type (see column 2), based on a coding rubric and summary statements based on the collected material.	Not discussed.	Calibration of the outcome and the conditions is presented in the main text (both in words and in tables) and in the appendix in a comprehensive fashion (e.g., coding rubric, summary statements, coding examplars). The fuzzy set data are presented in a table in the main text.	None. Still, by applying also process tracing, the authors can assess the relevance of the QCA-findings.
Li et al. (2016)	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.	NA (csQCA)	Not discussed	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.	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).
Kim & Verweij (2016)	The qualitative anchors are determined based on existing indices and by using the Tosmana threshold setter (that is, cluster	Mainly from existing indices and by using the Tosmana software.	For calibrating their outcome, the authors assigned a zero both to "no action" or "no information",	The three qualitative thresholds are presented in a table. The argumentation for these scores are	Sensitivity analysis based on different consistency cut-

	analysis).		which is conceptually not fully clear.	discussed in the main text.	offs.
Kirchherr et al. (2016)	The authors used a four-value and two-value coding scheme to assign fuzzy-set values to either conditions or outcome, or to their attributes. Some of the fuzzy-set values were based on existing quantitative indices, whereas others were based on interview and survey data.	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 averaged calibrations of the conditions and changed or recalibrated the sub-dimensions when the conditions' values were not face valid.	Not discussed	Calibration of each condition and the outcome is presented in the text, tables and an online appendix. The online appendix also provides information on the raw data, sensitivity analysis and calibration of conditions using various qualitative data sources.	Three types of sensitivity analyses were conducted – dropping cases; introduction of additional conditions; and alternative measures for a concept –, yielding a total of 11 sensitivity analyses, which are explained both in writing and in a table.
Metelits (2009)	The interview material is used to establish the qualitative breakpoints, as well as the other values of the six-value fuzzy set for the outcome and the three 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.	Not discussed	Tables with fuzzy values for the nine cases are provided per group of cases (i.e., 3 groups) and jointly in the main text.	None
Mishra et al. (2017)	The authors use four-value fuzzy sets. Most of their data is qualitative. They develop coding schemes for the conditions (or their sub-measures) and the outcome and illustrate this for one of their conditions in a table.	See also column 2. The authors finalized their calibration process with a final triangulation of the scores with field notes/observations and secondary data.	Not discussed	The authors present an example in a table in the main text. The calibrated data are not included in the paper or in an appendix.	None
Smilde (2005)	The author uses csQCA. He uses his qualitative (life-history interview) data to code the cases as being "in" (1.0) or "out" (0) of a set (such as the condition "life problems"). He discusses his coding rules in an appendix and offers examples of cases that would be coded out of a set and that would be coded in of a set.	NA (csQCA)	Not discussed	The author discusses his coding rules in an appendix and offers examples. The calibrated data are not included in the paper or in an appendix.	None
Summers Holtrop et al. 2016	No information is provided on how the thresholds are determined. Moreover, the qualitative descriptions representing the fuzzy-values for each condition (Table 6) sometimes span values both 'in' and 'out'	First, a scoring system was created using a 1-5 Likert-type scale to assign values to a list of attributes, based on qualitative information. The resulting quantitative scores were then analysed using basic descriptive statistics to determine which attributes	Not discussed	A table in the main text describes two case examples of how the qualitative information and the quantitative Likert-type assessment ratings informed the fsQCA	None

	of the set (e.g. 0.2-0.8), which is confusing.	would be used for the QCA analysis. Then, the quotations and codes were taken together to determine how the interviewees discussed each selected attribute, resulting in a score from 1-5. These were then converted into fuzzy-set scores, which were based on 'in-depth analysis and thematic analysis of features and context' (p. 20).		values for the five conditions and the outcome. A second table in the main text presents the list of conditions, an overall explanation of each condition, and the calibrated score given for each condition with an explanation of the various categories for the calibration values.	
Thomann (2015)	For the outcome, the author uses 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 uses a combination of existing indices that constitute 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 are based on the sample range (1.0 and 0) and its mean (0.5), so as to avoid unrealistic scenarios.	See column 2.	Not discussed	The calibration procedure is discussed in an appendix. This appendix also presents the raw data matrix and the fuzzy membership scores.	The author conducted an analysis of the negation of the outcome.
Tóth et al. (2017)	The 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 IV). 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.	Not discussed	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.	Sensitivity analysis based on different consistency cut-offs.
Van der Heijden (2015)	The author describes the assignment of the three 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 uses a four-value fuzzy set for the outcomes and conditions.	The author makes sure he receives enough information on all indicators to obtain a valid measurement. To this end, he starts by using information from websites, existing reports and other sources. Novel data on the cases are	The calibration of the data, including the setting of the thresholds, is discussed in an online appendix.	None

			subsequently obtained through a series of interviews to fill in gaps in the data from other sources.		
Vergne & Depeyre (2016)	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. 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) are 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 this may not result in a valid measurement. Calibration of one condition is based on letters to shareholders. Based on these letters, four values are given to each case (e.g. 0 indicating "not paying any attention" and 0.33 indicating "paying some attention").	The option "I don't know" is deliberately excluded in the expert survey. When someone was insufficiently knowledgeable, the authors ask that person not to complete the survey at all (p. 1661, note 8). When data about a specific indicator are missing, the authors turn to additional databases for information (but report that they sometimes did not find more information) (p. 1679).	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.	The authors conduct an additional analysis in which they did include directional expectations. Additionally, they conduct robustness analyses using alternative measures for one indicator and the outcome.
Verweij (2015)	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.	Not discussed	The "raw" data and membership scores are provided in a table in the main text. The reasoning behind this is discussed in the main text.	The author also conducts an analysis of the negation of the outcome.
Verweij & Gerrits (2015)	The qualitative data are used to determine the multi-value scores (0, 1 or 2) and the Boolean ones (0 and 1). These scores are recalibrated in a second round because they yield 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.	Not discussed	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.	None
Verweij et al. (2013)	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 score the cases individually. A subsequent iterative dialogue of several rounds between researchers' theoretical and substantive case knowledge is used to amend each other's scores. This results in the assignment of case membership scores on each condition (based on averaging the indicators).	Not discussed	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.	None

Wang (2016)	Based on the existing "raw" data (see column 3), 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 are 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 2). NB: The score of 0.5 is given, which is problematic.	Not discussed	In an appendix. There is no table summarizing the calibration procedure.	Alternative specifications of the calibration thresholds, specifically –following Fiss (2011)–, 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.

Note: NA means not applicable.

Appendix V

Table A2. Summary of the Types of Qualitative Data Used.

Interviews	(Basurto, 2013; Basurto & Speer, 2012; Chai & Schoon, 2016; Chatterley et al., 2014, 2013; Crilly, 2011; Fischer, 2014; Henik, 2015; Iannacci & Cornford, 2017; Kirchherr et al., 2016; Li et al., 2016; Metelits, 2009; Smilde, 2005; Summers Holtrop et al., 2016; Tóth et al., 2017; Van der Heijden, 2015; Vergne & Depeyre, 2016; Verweij, 2015; Verweij & Gerrits, 2015; Wang, 2016)
Existing documents/ archive material; ethno-graphies	(Basurto, 2013; Basurto & Speer, 2012; Chai & Schoon, 2016; Crilly, 2011; Crowley, 2012; Fischer, 2014; Hodson & Roscigno, 2004; Hodson et al., 2006; Iannacci & Cornford, 2017; 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., 2014, 2013; Mishra et al., 2017; Summers Holtrop et al., 2016; Verweij & Gerrits, 2015; Wang, 2016)
Focus groups	(Chatterley et al., 2014, 2013; Mishra et al., 2017)
Participant observation	(Smilde, 2005; Verweij & Gerrits, 2015; Verweij et al., 2013)

Appendix VI

Table A3. Checklist of considerations when using qualitative data in QCA

	Consideration	**Examples**
1	Be more explicit about how the thresholds for inclusion and exclusion of a set are established	a) Determine the threshold by constructing an imaginary ideal case b) Base the thresholds on a classification of interview responses
2	Be more explicit about how the degree of set-membership is established	a) Link qualitative data or codes to values on a Likert-type or other pre-determined numerical scale and subsequently translate it into fuzzy-set values b) Use rubrics, coding schemes or pre-determined qualitative classifications to assign fuzzy-set values
3	Pay more attention to the zeros in calibrated date	a) Construct the interview scheme such that all concepts are addressed during the interview b) Approach interviewees again with questions about the missing data c) Conduct sensitivity analyses
4	Explicitly delineate the choices made and present them clearly	a) Publish the raw data matrix b) Make large datasets available on the Internet or on demand c) Use a combination of explanations, illustrations and tables in the main text and (online) appendices
5	Conduct sensitivity analyses	a) Change the number of cases b) Change the conditions c) Re-run the analysis with a more extreme outcome

Notes

[1] We understand qualitative data as 'records of observation or interaction that are complex and contexted, and that are not easily reduced immediately (or, sometimes, ever), to numbers' (Richards, 2005: 34). Note that in this paper, we assume that researchers have already collected their qualitative data.

[2] We follow Goertz and Mahoney's (2012) terminology for qualitative research, hence using the terms concepts, attributes and data instead of variables and indicators (and sub-measures).

[3] The threshold setter should never be used mechanically. Researchers should check whether the thresholds set make sense – e.g., whether, for instance, qualitatively similar cases are all either "in" or "out" of the set –, and preferably complement this approach by another strategy.

[4] The exception here is Crowley (2012). However, the website to which Crowley refers is not accessible.

[5] Taking the average is mechanistic and provides a valid fuzzy-set only if the average adequately reflects actors' perceptions. If, however, the standard deviation is high, taking the average fails to result in a valid fuzzy-set.

[6] Assigning a zero in Kim & Verweij's (2016) study would influence the results only when the "no information" would be coded as "in" the set (i.e., >.5) if the information had been available, because this would bring the case from "out" to "in" the set.