Working Paper 2014-79

QCA, the Truth Table Analysis and Large-N Survey Data: The Benefits of Calibration and the Importance of Robustness Tests

Patrick Emmenegger, Dominik Schraff and André Walter

Department of Political Science
University of St. Gallen

Abstract: This paper argues that QCA is a suitable methodological choice for the analysis of a specific but widely used form of large-N data in the social sciences, namely survey data collected through computer-assisted telephone interviews or internet surveys. The reason is that the linguistic form of survey data often lends itself to a direct translation into fuzzy sets. Likert-scaled survey items let respondents make qualitative statements of agreement, disagreement and indifference. Fuzzy sets can capture these qualitative differences in a way that classical interval-scaled indicators cannot. Moreover, fuzzy algebra allows researchers to combine multiple sets in a simple and transparent manner, thereby giving QCA an important advantage over regression-based approaches. However, the analysis of large-N survey data removes one of the characteristic strengths of QCA: its case orientation. In the case of large-N survey data, the case orientation is typically weak and causal inference thus questionable. To remedy this shortcoming QCA methodologists have suggested robustness tests to enhance confidence in the proposed relationships. This paper shows how these robustness tests can be used in a large-N setting and suggests a new robustness test that is particularly suited for large-N survey data.

Keywords: Large-N, Surveys, Robustness Tests, Calibration, Truth Table Analysis, fsQCA

COMPASSS Working Paper 2014-79
Available at: http://www.compasss.org/wpseries/EmmeneggerSchraffWalter2014.pdf
COMPASSS Working Papers Series Managing Editor, Claude Rubinson

QCA, the Truth Table Analysis and Large-N Survey Data:

The Benefits of Calibration and the Importance of Robustness Tests

Abstract: This paper argues that QCA is a suitable methodological choice for the analysis of a specific but widely used form of large-N data in the social sciences, namely survey data collected through computer-assisted telephone interviews or internet surveys. The reason is that the linguistic form of survey data often lends itself to a direct translation into fuzzy sets. Likert-scaled survey items let respondents make qualitative statements of agreement, disagreement and indifference. Fuzzy sets can capture these qualitative differences in a way that classical interval-scaled indicators cannot. Moreover, fuzzy algebra allows researchers to combine multiple sets in a simple and transparent manner, thereby giving QCA an important advantage over regression-based approaches. However, the analysis of large-N survey data removes one of the characteristic strengths of QCA: its case orientation. In the case of large-N survey data, the case orientation is typically weak and causal inference thus questionable. To remedy this shortcoming QCA methodologists have suggested robustness tests to enhance confidence in the proposed relationships. This paper shows how these robustness tests can be used in a large-N setting and suggests a new robustness test that is particularly suited for large-N survey data.

Keywords: Large-N, Surveys, Robustness Tests, Calibration, fsQCA, Truth Table Analysis

Patrick Emmenegger (corresponding author)
Department of Political Science
University of St. Gallen
Rosenbergstrasse 51
CH-9000 St. Gallen
patrick.emmenegger@unisg.ch

Dominik Schraff
Department of Political Science
University of St. Gallen
Rosenbergstrasse 51
CH-9000 St. Gallen
dominik.schraff@unisg.ch

André Walter
Department of Political Science
University of St. Gallen
Rosenbergstrasse 51
CH-9000 St. Gallen
andre.walter@unisg.ch

Acknowledgements: For helpful comments on previous versions of this paper, we would like to thank Keith Banting, Andrew Bennett, Barry Cooper, Bernhard Ebbinghaus, Judith Glaesser, Johannes Meuer, Benoît Rihoux, Claude Rubinson, Carsten Q. Schneider, Wim van Oorschot, Claudius Wagemann, Oliver Westerwinter and two anonymous reviewers. Special thanks go to Alrik Thiem and Adrian Dusa for providing support on their QCA R package. We are also grateful to all participants of the ESPAnet doctoral workshop "Comparing Welfare States: Applying Quantitative and Qualitative Comparative Analysis in Social Policy Research" in Mannheim (2013), the NordWel & REASSESS International Summer School in Reykjavik (2013) and the QCA Expert Workshop in Zürich (2013).

Introduction

Qualitative Comparative Analysis (QCA) is typically used in case of small- and medium-sized datasets. However, in recent years scholars have begun to explore the potential of QCA for the analysis of large-N datasets (Cooper 2005; Ragin and Fiss 2008; Glaesser and Cooper 2010; Cooper and Glaesser 2011; Greckhamer et al. 2013). QCA has unique advantages over regression-based approaches (Ragin 2008; Vis 2012; Schneider and Wagemann 2013) and thus promises new insights in the analysis of large-N datasets. However, the suitability of QCA for the analysis of a large number of cases is still an open question. While Ragin (1987, 2000) originally envisioned QCA to be a medium-N data method, other authors have emphasized that QCA is the appropriate choice of method if theoretical arguments are formulated in set-theoretic terms, independent of the number of observations (Schneider and Wagemann 2012). Nevertheless, most methodological work on QCA has so far focused on medium-N datasets.

In this paper, we discuss the extent to which QCA is a suitable methodological choice for the analysis of a specific but widely used form of large-N data in the social sciences, namely survey data collected through computer-assisted telephone interviews or internet surveys. We argue that such large-N datasets raise two methodological issues that have not received sufficient attention in methodological debates yet: the potential of calibration and the importance of robustness tests.

First, we argue that the linguistic form of survey data often lends itself to a direct translation into fuzzy sets. Likert-scaled survey items let respondents make qualitative statements of agreement, disagreement and indifference. Fuzzy sets can reflect these qualitative differences in a way that classical interval-scaled indicators cannot. Moreover, fuzzy algebra allows researchers to combine multiple sets in a simple and transparent manner. As we demonstrate in the first part of

this paper, researchers have not yet taken sufficient advantage of this untapped potential of QCA for the analysis of large-N survey data. Contrarily, especially in large-N settings researchers still seem prone to use averages or other inductive procedures to calibrate sets. We argue that this is bad practice and illustrate how researcher can use theoretical knowledge to create more adequate sets.

Second, the analysis of large-N survey data removes one of the characteristic strengths of QCA: its case orientation. The results of the Truth Table Analysis are but a midpoint in a proper QCA analysis. Typically, the results of the Truth Table Analysis are, among others, complemented by a qualitative discussion of the cases that are covered by a solution term to show that the observed configurations indeed represent causal relationships. However, in the case of large-N survey data, this case orientation is often weak and causal inference thus questionable. To remedy this shortcoming QCA methodologists have suggested robustness tests to gauge causality. In the second part of this paper, we show how these robustness tests can be used and suggest a new test that we believe is particularly suited for large-N survey data.

To make these two methodological points, we use the extensive literature on opposition towards immigration as our empirical example. There is a large number of established theoretical arguments explaining opposition towards immigration and many researchers have used the 2002/03 wave of the European Social Survey, thus allowing us to compare our findings with numerous studies using regression-based approaches (e.g. Dustmann and Preston 2004; Sides and Citrin 2007; Finseraas 2008; Rydgren 2008; Herreros and Criado 2009; Meulemann et al. 2009; Senik et al. 2008; Emmenegger and Klemmensen 2013ab). In particular, we analyse whether preferences for cultural homogeneity, ethnic competition over scarce resources such as jobs,

heterogeneous social networks and education influence opposition towards immigration in Germany (Pettigrew 1998; Hainmueller and Hiscox 2007; Sides and Citrin 2007; van Oorschot and Uunk 2007; Rydgren 2008; Ceobanu and Escandell 2010). However, our main analytical goal is not to make a substantive contribution to the literature on preferences for immigration. Hence, we refrain from discussing the literature in any detail and refer readers to the cited sources for a discussion of the theoretical arguments.

The paper is structured as follows. In the next section, we discuss the untapped potential of survey data for the calibration of fuzzy sets. Subsequently, we demonstrate the importance of robustness tests in case of large-N survey data. In this section we also suggest a novel robustness test that is particularly suitable for large-N survey data. A final section concludes.

Untapped potential: The calibration of survey data

In some respects QCA is better suited to deal with survey data than regression-based approaches. Regression-based approaches typically rely on indicators created by means of inductive procedures and thus often sever the direct link between concepts and measures. In contrast, set-theoretic methods can translate survey items and Likert scales into (fuzzy) sets without any loss of information or conceptual clarity. In particular, the calibration of sets does not force researchers to turn differences-in-kind (opposition or no opposition) into differences-in-degree (more or less opposition). In the following section, we outline our argument in detail. In addition, we use the example of the "opposition towards immigration"-literature to illustrate our argument and demonstrate the advantages of QCA for survey data.

In regression-based approaches, researchers typically rely on quantitative, inductive techniques such as simple averages or factor analysis to create indicators. In a similar vein, QCA practitioners often propose and use such inductive techniques to calibrate sets (cf. Berg-Schlosser and De Meur 2009; Crilly et al. 2012; Greckhamer et al. 2008; Grendstad 2007; Schneider et al. 2009; Thygeson et al. 2012). However, such inductive approaches are generally seen in a critical light in set-theoretical research because they typically lead to concept-measure inconsistencies (Goertz 2006). Put simply, qualitative differences still reflected in concepts get lost once these concepts are operationalized as indicators (or as sets using inductive approaches such as averages). In contrast, sets, properly calibrated, are able to reflect these qualitative differences. Using external standards such as the researchers' theoretical and substantive knowledge, sets thus increase concept-measure consistency. But as the number of cases increases, it often becomes difficult to have the necessary substantive knowledge for assigning cases to sets. As a result, QCA scholars often revert to quantitative techniques such as simple averages of indicators to calibrate sets.

However, not all large-N data sets suffer from this problem. In particular, survey data, although typically large-N, can often be easily translated into sets. The reason for this is the survey questions' linguistic form. Words are inherently set-theoretical. As emphasized by the cognitive linguist Lakoff (1987), experiences, events and things are not perceived as unique but rather in terms of patterns and categories. To these categories we give names that convey meaning (concepts). For instance, when using the term 'democracy', we (and our social environment) recall a particular class of units that are all comparatively similar (summarized in the concept 'democracy'). To put it in set-theoretic terms, there is a set of democracies, in which countries have different levels of (non-)membership. Concepts such as democracy refer to the knowledge

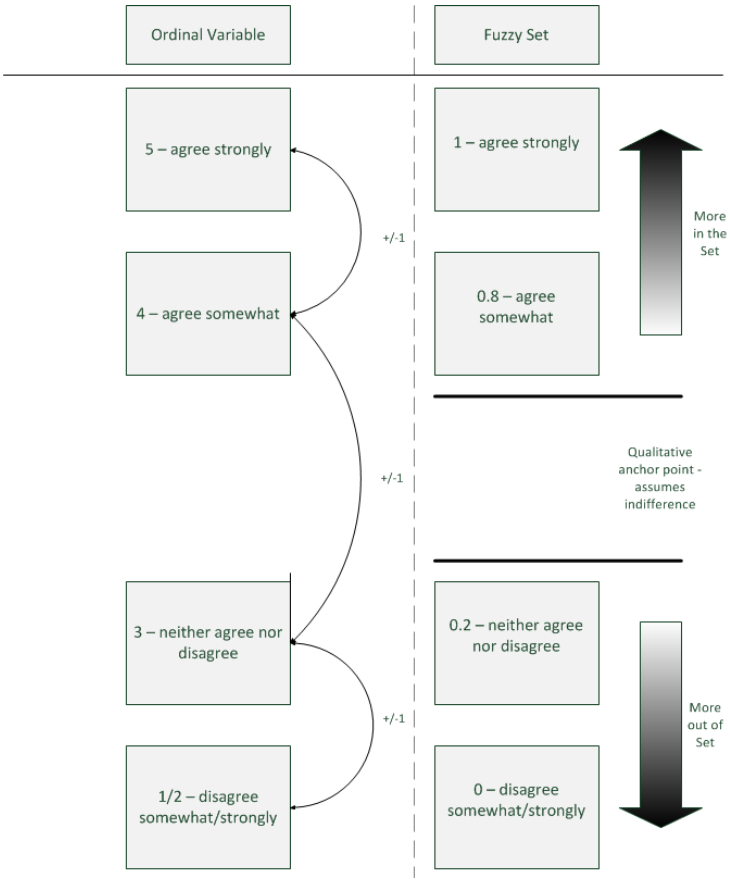
that we have about a category and therefore allow for linguistic action (Schedler 2012). They allow use to communicate what is (rather) part of the category and what is not.

Survey data, unlike conventional (numeric) indicators, is based on words. For instance, in case of Likert scales, respondents are asked whether they agree with a certain statement. In most research, respondents' answers are subsequently translated into ordinal scale variables that reflect how strong the respondents' agreement was with certain statements. However, such translation automatically leads to a significant loss of information because the qualitative dimension of the survey item is no longer considered in the indicator. While we are still able to observe whether a person has a higher level of agreement than another one, we are no longer able to observe whether somebody in fact agrees or disagrees with the statement. These differences-in-kind, as expressed in the survey question, have disappeared from the data set once the data have been turned into indicators. This loss of information, however, is not necessary. Fuzzy sets are able to capture all the relevant information, i.e. both differences-in-degree as well as differences-in-kind.

In our empirical example, we analyse the determinants of opposition towards immigration using fsQCA. One of the conditions that we use in the analysis captures respondents' preferences for cultural unity. This set is based on the survey statement that "it is better if almost everyone shares the same customs and traditions". Respondents were asked if they agree or disagree with the statement. The five answer options were "strongly agree", "agree somewhat", "neither agree nor disagree", "disagree somewhat" and "strongly disagree". We code respondents who strongly agree with the statement as fully in the set "preference for cultural unity", while respondents who somewhat agree are assigned to the value 0.8, reflecting the ambiguity in the formulation "somewhat". In contrast, all respondents who express disagreement with the statement are coded

as being fully out of the set "preference for cultural unity". Finally, indifferent respondents have some partial membership in the set "preference for cultural unity" although they are rather out of the set than in. By being indifferent, respondents rejected to agree with the statement. Hence, it would be incorrect to put indifferent respondents on the maximum point of ambiguity, 0.5. Instead, we assign them the value 0.2, which reflects the fact that these respondents also deliberately decided not to disagree with the statement. Figure 1 shows the calibration in detail.

Figure 1: Calibration of the Likert scale "cultural unity"



As a result of our calibration, only respondents who agree with the statement are in the set "cultural unity". In contrast, regression-based approaches do not consider this qualitative dimension of the survey item. Rather, they treat the difference between "strongly disagreeing"

and "somewhat disagreeing" exactly like the difference between "neither agreeing nor disagreeing" and "somewhat agreeing". The qualitative difference between agreement and disagreement has been lost in translation. As a consequence, set-theoretic approaches are particularly suited to maximize the information contained in Likert scales.

Unlike QCA, regression-based techniques use survey items as quasi-metric indicators, thereby ignoring the substantive meaning of indifference and qualitative differences between agreeing and disagreeing with certain statements.[1] However, following Goertz (2006), the graduation of a phenomenon makes sense only within and not across sets. The consideration of the qualitative dimension contained in survey items has also important implications for causal statements. For an investigation on the determinants of opposition towards immigration, it hardly makes sense to examine the determinants of support of immigration (which regression-based approaches implicitly do). The relationship between education and opposition to immigration clarifies this point: While many studies argue that higher education leads to more tolerance towards immigrants, no study employs the reverse argument, i.e. that lower levels of education cause intolerance towards immigration (Ceobanu and Escandell 2010: 319).

Set-theoretic approaches also have an untapped potential with regard to the combination of multiple survey items. When a concept cannot be measured by a single indicator, it is common procedure in studies using regression-based approaches (but oddly enough also in studies using

[1] Of course, regression-based approaches are also able to capture such qualitative differences. However, we could not find a single paper in the "opposition towards immigration" literature that considered such qualitative differences within survey items. On the difference between the methodologically possible and usual practice see also Goertz and Mahoney (2012).

set-theoretical methods) to simply use the average of different variables or to conduct a factor analysis of variables that are expected to be in a causal relationship with the latent concept (for QCA studies using such inductive procedures see Berg-Schlosser 2008; Cárdenas 2012; Cheng et al. 2013; Crowley 2013; Engeli 2012; Grendstad 2007; Vaisey 2007). However, using such an inductive approach to capture sets is problematic for at least two reasons.

First, while indicators are typically numeric, concepts are constructed in terms of necessary and sufficient conditions (Goertz 2006). For instance, Canada is not a member of the category European democracies because Canada, although democratic, is not a European country. Hence, Canada's set membership is zero and not 0.5 as the average of the variables 'democratic' and 'European' might imply. Hence, the calibration of sets by means of linear algebra is highly susceptible to misclassification, while conceptual thinking implies that variables are combined in a logical fashion using AND/OR operations. If the conceptual structure of necessary and sufficient conditions is not reflected in the measurement process, the result is concept-measure inconsistency. In our empirical example, scholars typically relied on inductive approaches, thus leaving conceptualization underdeveloped and resulting in large number of empirical work which is conceptually only loosely connected.

Second, averaging different variables to capture a concept is based on the assumption that all indicators are equally important for a concept. For instance, opposing immigration from poor Asian or African countries has the same weight as the opposition towards immigration from rich, neighbouring European countries. However, this line of argumentation is hardly justifiable for a number of reasons, which we discuss below.

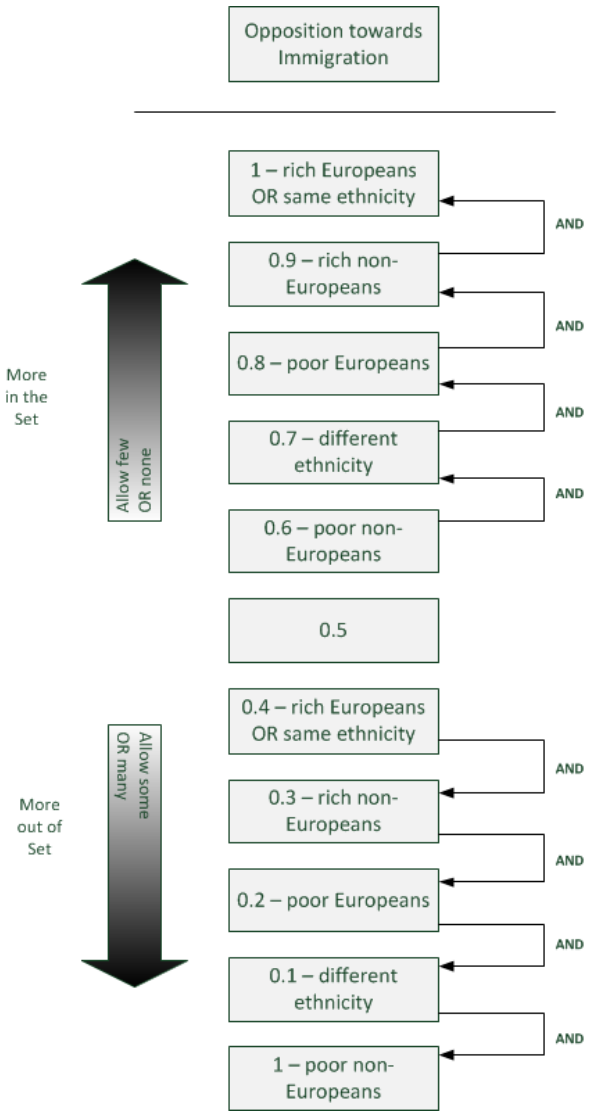
To clarify both points, we turn to the calibration of the outcome "opposition towards immigration" from our empirical example. Previous studies typically used a set of six survey items from the first round of European Social Survey. In the six questions, respondents were asked to what extent rich/poor immigrants from European/non-European countries and from the same/different race or ethnic group should be allowed to enter the country. The answer options were "allow many", "allow some", "allow few" and "allow none". The studies then often simply used the average of the respondents' answers to these six questions. In contrast, we combine these six survey items in a theoretically informed way to capture "opposition towards immigration" (see Figure 2).

To construct the set we distinguish between the geographical, ethnic and stratificatory dimensions of opposition towards immigration. Previous studies have shown that respondents assess immigration along these dimensions differently (Hainmueller and Hiscox 2007; Hainmueller and Hopkins 2013). In the first step, all respondents who oppose immigration (meaning answering "allow few" or "allow none") of poor, non-European immigrants are assigned to the score 0.6. The reason we assign these respondents only marginally above the 0.5 anchor point is that opposition to immigration is most common when it focuses on culturally and geographically more distant groups. In addition, previous studies provide evidence that people prefer high-skilled to low-skilled immigration, independent of the educational and occupational background of the respective respondent (Hainmueller and Hiscox 2007).[2] As a consequence,

[2] Hainmueller and Hiscox (2007) show that the education level of immigrants is strongly associated with the GDP per capita of the home country. Therefore, they argue, questions about immigration from poor countries can be interpreted as questions about low skilled immigration.

opposition towards immigration from poor, non-European countries is the most prevalent form of opposition towards immigration.

Figure 2: Calibration of the set "opposition towards immigration"



In a second step, we assign all respondents who *additionally* oppose immigration from different ethnic groups to 0.7. While the question about immigration of a different ethnic group suggests geographical and cultural distance, the stratificatory dimension is now missing. Therefore,

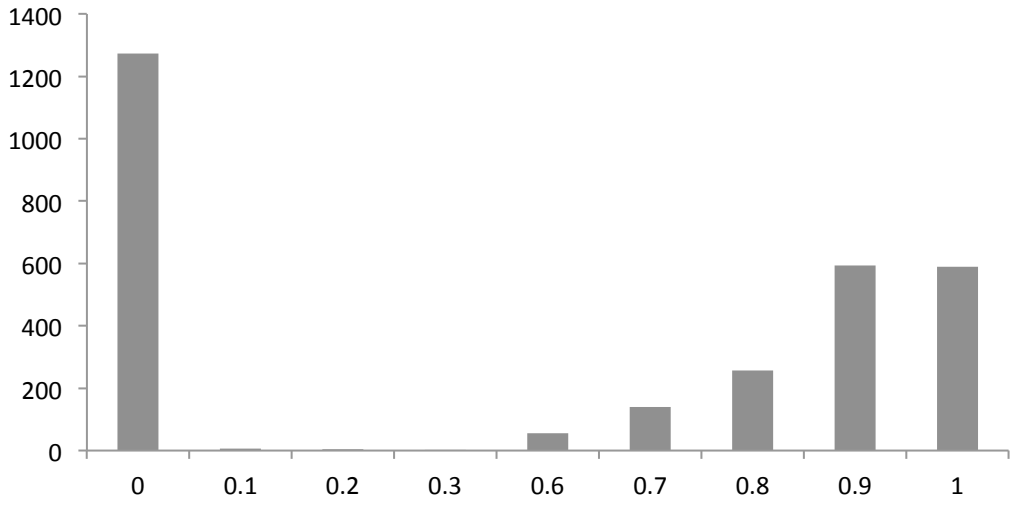
opposing immigration of a different ethnic group includes poor as well as rich individuals. Consequently, respondents have to oppose immigration in the current as well as in the previous question(s) to be assigned higher fuzzy set membership scores.[3] Using such AND-relationships between the questions, we can ensure that these respondents oppose immigration to a higher degree. The remaining set is constructed in the same manner. Respondents who oppose immigration from poor European countries (and poor non-European countries as well as from a different ethnic group) are assigned to 0.8. Opponents of immigration from rich non-European countries are assigned to 0.9, while respondents who oppose immigration even from rich European countries and from the same ethnic group are fully in the set (1.0).

We construct non-membership in the set in a similar fashion. We consider respondents to be out of the set if they answer "allow some" or "allow many" to the question, meaning they support immigration. To get assigned to the 0.4 point, respondents have to support immigration from the same ethnic group or rich European countries. Here, the intuition is also straightforward: Support for culturally and geographic proximate immigration as well as from rich European countries is more common compared to other forms of immigration. Respondents who also support immigration from rich, non-European countries are assigned to 0.3. Again, these respondents also have to support immigration from rich European countries as well as from the same ethnic group. Furthermore, respondents who support immigration from poor European countries are assigned to 0.2 while supporters of immigration of different ethnic groups are assigned to 0.1. Only

[3] We have examined the response behaviour by constructing dummy variables of opposition towards immigration ("allow few" and "allow none") and support for immigration ("allow some" and "allow many") for all survey items and cross-tabulated them. Overall, the response behaviour is very consistent with our construction of the outcome.

respondents who even support immigration from poor non-European countries and a different ethnicity are fully out of the set of "opposition towards immigration". Figure 3 shows the distribution of our outcome "opposition towards immigration". The distribution is almost perfectly U-shaped. A large part of the members of the set "opposition towards immigration" are fully in the set or assigned to the score 0.9. Furthermore, respondents who are out of the set are even more homogeneous with regard to their attitudes. The vast majority receives the score 0.0.

Figure 3: Histogram of the outcome "opposition towards immigration"



To investigate the determinants of opposition towards immigration in the next section, we select five prominent arguments from the literature. First, the fuzzy set "cultural unity" accounts for identity-related arguments: Immigration threatens the national and cultural identity of natives (Sides and Citrin 2007: 479). Cultural unity is calibrated as an asymmetrical fuzzy set as discussed above.

The second fuzzy set is economic threat. According to self-interest theories, immigration leads to more competition on the labour market. As a result, natives develop hostile attitudes towards

immigrants (van Oorschot and Uunk 2007). To calibrate this condition we use different data sources: First, we capture the economic risk by calculating occupation-specific unemployment rates as well as the migrant share in all occupations using data from the ILO and OECD (using current or previous occupations). We use the 25, 50, and 75 percentiles for the 0, 0.5, and 1-anchor points to construct two separate fuzzy sets (for unemployment and migrant shares).[4] We then combine both sets by using the family resemblance strategy, meaning that we take the maximum score of both sets to assign the respondents to our set "economy threat". While this set can account for competition for wages and promotions, the crucial part of exclusion from the labour market is still missing. Hence, we use the survey question concerning the employment status of respondents. If respondents are currently unemployed, they are recoded as fully in the set of people facing an "economic threat" (independently of the unemployment rates and migrants shares in their former occupational groups). As this example demonstrates, fuzzy sets allow for the creative combination of multiple data sources.

The last three conditions are simply calibrated as crisp sets. According to the literature, education is one of the most important determinants for (or rather against) opposition towards immigration (cf. Ceobanu and Escandell 2010). We calibrate education by assigning all respondents who completed tertiary education to the set of highly educated respondents while all other respondents are out of the set. The second condition derives from interaction theory: Respondents who have a large number of immigrants in their social network develop sympathetic attitudes towards immigration (Pettigrew 1998; Sides and Citrin 2007). Our set differentiates between respondents

[4] Calibration with percentiles can be an adequate strategy for continuous variables. Our argument for a more sensible, set-theoretic calibration primarily applies to ordinal-scaled survey items often used in social science research.

with no immigrant friends (full membership) and respondents with many immigrant friends (full non-membership). The third crisp set captures gender, differentiating between men (full membership) and women (full non-membership). Previous studies provide evidence that men hold more negative attitudes toward immigration (Quillian 1995).

Questionable relationships: The importance of robustness tests

Even though set-theoretic approaches like QCA seem to have clear advantages over statistical approaches in utilizing information from large-N surveys, QCA faces difficulties in inferring causality from the data analysis. From a purely technical point of view, the Truth Table Analysis should work with large-N survey data just as well as with small-N data. However, QCA is a case-oriented method, which has important consequences within a large-N setting.

Going back to the cases is a crucial analytical step in QCA (Emmenegger et al. 2013). As Ragin (2000: 283) puts it, the Truth Table Analysis does not substitute for the study of cases "just as reading a detailed map is not a substitute for taking a hike in the mountains". Yet, this going back to the cases is often not possible with large-N survey data because we have only the limited information provided by the survey and no chance of contacting individual respondents afterwards. Even if we had the chance to contact individuals after a single survey, the mere number of cases would make it impossible for the researcher to manage the information (although researchers can of course take advantage of the additional information contained in a survey).

However, case-orientation is crucial since QCA (including the Truth Table Analysis) does not rest on strongly formalized, automatic procedures. The calibration of the data, the setting of consistency and frequency thresholds and the selection of conditions requires researchers to make qualitative decisions which influence the results. An investigation of single cases is important to validate that these decisions most adequately reflect the realities in the respective case universe.

It has correctly been pointed out that the Truth Table Analysis is very sensitive to single cases and measurement error (Hug 2013; Seawright 2005). This finding, however, is not very surprising for a case-oriented method. QCA depends on the qualitative position of its cases within a set. Hence, we want our analysis to react to changes in our data. With QCA a causal inference does not rest first and foremost on the robustness of the algorithm to data manipulations. Rather, causal inferences are identified by going back to the cases (Thiem and Ragin 2013). Yet, in a large-N setting we are no longer able to provide this crucial validation step. This is a challenge for QCA that does not hold true for statistical methods. Put differently, robustness concerns are specifically relevant for large-N QCA applications. Hence, in such settings the causal interpretation of QCA results seems problematic.[5]

In recent years, the literature has suggested a number of strategies to deal with such robustness concerns if going back to the cases is not a viable option (Skaaning 2011; Maggetti and Levi-Faur 2013; Schneider and Wagemann 2013: 284ff). Of course, issues such as measurement error, sensitivity to calibration decisions and the choice of thresholds are not problems exclusive to

[5] Of course, robustness does not imply causality. One could say that robustness is a necessary but not sufficient condition for causal inference. In practice, however, researchers typically use robustness tests to enhance their trust into the proposed causal relationship.

large-N settings. Also in small- and medium-N settings researchers have to explore the extent to which their results rest on particular decisions made in the analysis. Yet, we think that the issue of robustness is even more pronounced in large-N settings because of the loss of the case-orientation. Moreover, large-N survey data is especially prone to one specific error. With survey data *measurement error* is endemic (Hug 2013: 261) but we lack clear guidelines on how to deal with it. This problem is exacerbated by the fact that QCA papers are often rather weak with regard to the formulation of complex theoretical propositions in set-theoretical terms (Emmenegger et al. 2013; Hug 2013).

We now revisit some of the most prominent suggestions made to assess the robustness of QCA results and then use our large-N survey data example on opposition towards immigration to show how they can be implemented in a large-N setting. Following Skaaning (2011) and others, we examine the robustness of our findings by using different consistency and frequency thresholds and by changing the calibration. In addition, we propose a new robustness test that we believe is particularly suitable to deal with measurement error in case of large-N samples: the random deletion of shares of cases.

Table 1 presents the truth table of our five conditions: low education [LOWEDU], preference for cultural unity [UNITY], no immigrant friends [NOFRIEND], facing an economic threat [THREAT], gender [MAN] and the outcome opposition to immigration. As can be seen from the table, the consistency values of the rows are generally rather low.[6] The truth table reveals a problem not new to survey research. The abundance of information and the resulting

[6] Consistency levels are very similar for the negated outcome. Only the first five truth table rows have a consistency value of above 0.70.

heterogeneity in the response behaviour of survey participants reduces the consistency scores to a rather low level. A similar problem is often encountered in statistical research, where measures of explained variance (such as "R^2") usually become rather small when analysing individual-level data.

Table 1: Full truth table

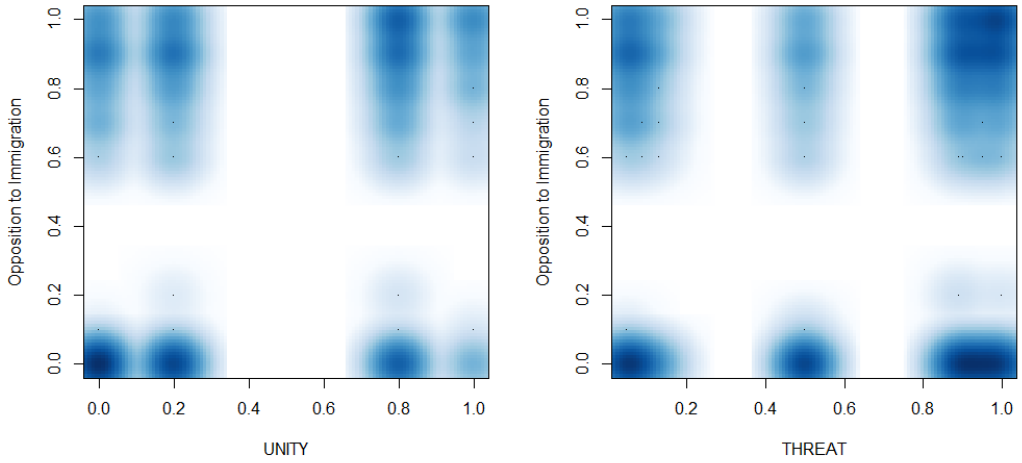
Row	[LOWEDU]	[UNITY]	[NOFRIEND]	[THREAT]	[MAN]	Outcome	N	Consistency
1	1	1	1	1	1	1	193	0.782
2	1	1	1	1	0	1	196	0.781
3	1	1	1	0	0	1	83	0.713
4	0	1	1	1	1	1	22	0.682
5	1	1	1	0	1	1	40	0.680
6	1	1	0	1	0	1	98	0.668
7	1	0	1	1	1	1	164	0.629
8	1	1	0	0	0	1	41	0.621
9	1	1	0	0	1	1	24	0.605
10	1	0	1	1	0	1	163	0.603
11	1	0	1	0	1	1	47	0.603
12	1	1	0	1	1	1	76	0.601
13	0	1	0	1	1	0	17	0.576
14	0	1	1	0	1	0	49	0.574
15	0	1	0	0	1	0	37	0.560
16	1	0	1	0	0	0	96	0.55
17	0	1	0	1	0	0	5	0.536
18	0	0	1	1	1	0	30	0.506
19	1	0	0	0	1	0	59	0.467
20	1	0	0	1	0	0	172	0.465
21	0	1	0	0	0	0	22	0.455
22	1	0	0	0	0	0	101	0.453
23	0	1	1	0	0	0	21	0.445
24	1	0	0	1	1	0	156	0.434
25	0	0	1	0	0	0	56	0.410
26	0	0	1	0	1	0	47	0.409
27	0	1	1	1	0	0	7	0.389
28	0	0	1	1	0	0	19	0.383
29	0	0	0	1	1	0	40	0.303
30	0	0	0	0	1	0	123	0.295
31	0	0	0	1	0	0	26	0.285
32	0	0	0	0	0	0	93	0.236

For QCA a consistency level of 0.75 is often used as a rough benchmark, even though it is emphasized that the threshold should depend on the specific research context. With a low number of cases consistency thresholds should be higher, while a high number of cases allows for lower

consistency values (Schneider and Wagemann 2013: 127f). In small- and medium-N settings, low consistency values are often prescribed to inadequate calibration or inadequate selection of conditions. Yet, we believe this interpretation can be disregarded in our case because we use data and conditions that found strong support in numerous previous studies (Dustmann and Preston 2004; Sides and Citrin 2007; Finseraas 2008; Rydgren 2008; Senik et al. 2008; Herreros and Criado 2009; Meuleman et al. 2009; Emmenegger and Klemmensen 2013ab). Hence, we are quite confident that the low consistency values are not due to a lack of association between the conditions and the outcome but rather due to data characteristics typical of large-N survey data.

The two plots of Figure 4 show how strong respondents cluster in the different corners of the property space given our two fuzzy conditions "UNITY" and "THREAT" as well as the outcome opposition to immigration. There are no clear visible patterns. Hence, given this heterogeneity in the response behaviour, the low consistency values displayed in Table 1 are not surprising at all.

Figure 4: Kernel density plots of fuzzy sets with the outcome



For the following analysis, we use a consistency threshold of 0.66 for two reasons. First, there is a clear drop of consistency between the sixth and seventh row in Table 1. Second, even though this consistency threshold is rather low, a large-N setting gives us more leverage to decrease the consistency value. A consistency threshold of 0.66 could be regarded as a degree of sufficiency described by Ragin (2000: 110) as "usually sufficient".[7]

We now turn to the analysis of sufficiency by logically minimizing the truth table.[8] We set a frequency threshold of N=25. By disregarding rare combinations of conditions, we already apply a first strategy against measurement error. Table 2 presents the intermediate solution, which reports two prime implicants (PI).[9] The solution consistency of 0.742 is rather good given the

[7] The 0.75-consistency threshold is a matter of convention rather than a fixed value. Schneider and Wagemann (2012: 128) highlight a number of contextual factors that should inform the choice of the consistency threshold. Most importantly, the authors emphasize the infeasibility of determining a single, universal consistency value. They therefore recommend varying the threshold to assess the sensitivity of the results.

[8] For reasons of space we do not display the results of the analysis of necessary conditions. We did not find a necessary condition in our analysis. The best result was received for LOW EDUCATION, having a consistency of 0.81 and coverage of 0.56. This is clearly below the recommended 0.90 consistency threshold. Also, the condition is likely to be trivial (Schneider and Wagemann Trivialness of Necessity Score is 0.45). Generally, our random deletion procedure should also work with necessity statements. Yet, we think that robustness might be less relevant since solutions are simple and researchers are more conservative with regards to the measures of fit. We thank Claude Rubinson for making us think about this question.

[9] To be short, we focus on the intermediate solution, as it is the most frequently used solution type. It is also a good starting point to investigate stability of the solutions given that the intermediate solution is a midpoint between the parsimonious and the complex solutions. For our baseline result in Table 2, the complex and intermediate solutions are the same, while the parsimonious solution misses LOW EDUCATION in the second solution term. Generally, all the robustness checks we discuss can be applied to all three solutions.

heterogeneity in the data. Also the two PIs from Table 2 both have good consistency values. Yet, the first PI has decisively more unique coverage. However, overall, the solution coverage suffers from the abundance of information induced by the individual-level data. The heterogeneity in the answer patterns makes it hard to identify regularities covering a large number of cases. What we can derive from such an analysis are relatively consistent sufficiency statements for comparatively small subsets of the data sample.

Table 2: Analysis of sufficiency: Intermediate solution

Analysis of the Outcome		**Consistency**	**Raw Coverage**	**Unique Coverage**
1. PI	LOWEDU * UNITY * NOFRIEND	0.758	0.312	0.183
2. PI	LOWEDU * UNITY * THREAT * man	0.741	0.188	0.059
	Solution	0.742	0.371	

Note: Consistency Threshold of 0.66; Frequency Threshold of N= 25; Cases Coded as In/Out: 610/1576

By employing the conventional robustness tests, we now assess how robust these conclusions are. Table 3 displays the results. In a first step, we vary the frequency threshold. The first panel of Table 3 shows the intermediate solution with a frequency threshold of N=5 (rather than 25). This effectively leads to the inclusion of row 4 in the minimization. The second panel displays the intermediate solution with a frequency threshold of N=50, which effectively leads to the exclusion of row 5. The results show that changes to the frequency threshold only lead to minor changes in the analysis. With a frequency threshold of five, we receive a third PI that has virtually no unique coverage. Overall, the solutions remain rather stable.

Varying the frequency threshold provided us with some first useful insights about the stability of the results. From this exercise we have won some confidence in the robustness of our solution. However, there are clear limits to extent to which the frequency thresholds can be varied (within

reasonable boundaries). The strategy thus provides only limited potential to assess the robustness of our results.

Table 3: Conventional robustness tests

Frequency Threshold = 5				
		Consistency	**Raw Coverage**	**Unique Coverage**
1. PI	LOWEDU * UNITY * NOFRIEND	0.758	0.312	0.060
2. PI	LOWEDU * UNITY * THREAT * man	0.741	0.188	0.059
3. PI	UNITY * NOFRIEND * THREAT * MAN	0.765	0.146	0.022
	Solution	0.738	0.394	
Frequency Threshold = 50				
		Consistency	**Raw Coverage**	**Unique Coverage**
1. PI	LOWEDU * UNITY * NOFRIEND	0.758	0.312	0.183
2. PI	LOWEDU * UNITY * THREAT * man	0.741	0.188	0.059
	Solution	0.742	0.371	
Consistency Threshold = 0.7				
		Consistency	**Raw Coverage**	**Unique Coverage**
1. PI	LOWEDU * UNITY * NOFRIEND * man	0.760	0.169	0.040
2. PI	LOWEDU * UNITY * NOFRIEND * THREAT	0.781	0.252	0.123
	Solution	0.769	0.292	
Consistency Threshold = 0.62				
		Consistency	**Raw Coverage**	**Unique Coverage**
1. PI	LOWEDU * UNITY * man	0.722	0.246	0.078
2. PI	LOWEDU * UNITY * NOFRIEND	0.758	0.312	0.020
3. PI	LOWEDU * NOFRIEND * THREAT * MAN	0.689	0.200	0.077
	Solution	0.703	0.467	
New Calibration of UNITY				
		Consistency	**Raw Coverage**	**Unique Coverage**
1. PI	LOWEDU * UNITY * NOFRIEND	0.757	0.304	0.179
2. PI	LOWEDU * UNITY * THREAT * man	0.736	0.184	0.059
	Solution	0.739	0.363	

We now turn to the consistency threshold. In case of large-N data, setting the consistency threshold might seem rather arbitrary. Yet, similar to the frequency threshold, the available options for reasonable thresholds are in fact rather limited. On the one hand, the consistency threshold should be clearly distinguishable from 0.5 and should deliver consistent results. On the other hand, it should also allow covering a substantial number of cases in the analysis.[10] That said, the results of the Truth Table Analysis should not be affected by slight variations of the consistency threshold.

The third and fourth panels of Table 3 present the intermediate solutions with consistency thresholds of 0.7 and 0.62, respectively (instead of 0.66). With the higher consistency value we observe important changes. The first PI of the original solution is now joined by "man", while the second PI now contains "NOFRIEND" instead of "man". This would change the substantive conclusions we draw from the analysis. Also, unique coverage now assigns more importance to the second PI. The results also change substantively if we lower the consistency value to 0.62. Here the second PI is identical with the first PI in the original solution. However, the other two PIs are different. Unique coverage is now low for all three paths. In addition, solution coverage increased while consistency slightly decreased.

From varying the consistency threshold we conclude that the results are not entirely robust. Even though we identify parallels between the solutions, the changes are likely to challenge the substantive interpretations we draw from the solutions.

[10] This trade-off between consistency and coverage is not unique to large-N QCA but a general characteristic of set-theoretic methods (Ragin 2008: 55; Schneider and Wagemann 2013: 148ff)

As a third robustness test, we now change the calibration of our condition "UNITY". Taking seriously our suggestions about theoretically informed calibration from the first part of this article, we want to highlight that the usefulness of this strategy is limited as well because the set membership scores are determined by theoretical reasoning. Yet, for instance for "UNITY" the set membership values of 0.8 ("somewhat agree") and 0.2 ("neither agree nor disagree" are assigned rather arbitrarily. For the purpose of robustness we change these values to 0.7 and 0.3 respectively. The last panel in Table 3 shows that these calibration changes do not affect our findings. One interesting consequence of this exercise is that it shows that one should not put too much meaning into fuzzy set scores. As long as scores remain on the same side of the point of maximum ambiguity, the results of the Truth Table Analysis seem to be rather robust. This is important to note for scholars criticising the often unavoidable degree of arbitrariness in specific fuzzy set scores.

So far we have implemented the state-of-the-art strategies to assess the robustness of QCA solutions. The findings, however, are somewhat inconclusive. Solutions change in response to our tests, yet we could also clearly identify patterns of stability. What is more, our implementation of these robustness tests shows that the possibilities to do reasonable re-adjustments to the specifications are in fact quite limited.[11] Especially with regard to the frequency threshold, we have often limited possibilities to investigate reasonable alternative specifications. Yet, the frequency threshold is thought of being especially useful to address case-sensitivity and measurement error in large-N QCA (Ragin 2008: 133). This type of error can be referred to as random error and is typically considered to be pervasive in large-N survey datasets

[11] This is not bad news as it implies that in QCA researchers should often converge to the same choices with regard to calibration and thresholds even in the absence of 'objective' criteria.

(Maggetti and Levi-Faur 2013: 202). Hence we are particularly dissatisfied with our robustness tests regarding random error. Therefore, we use the remainder of this paper to propose a new strategy for assessing random measurement error in large-N QCA.

Our strategy takes advantage of the possibilities offered by large-N data with a random data-generating process. This setting allows us to randomly delete a proportion of cases from the data set and re-run the minimization. We think it is a useful endeavour to assess how sensitive a solution is to random deletion of cases for two reasons. First, this strategy allows us to investigate how much a solution depends on single or small groups of cases. Second, by reducing the sample size through random deletion we can model the effect of randomly missing data. This is a crucial extension for our robustness test section because the survey data we are using is very likely to suffer from random error and the previous strategies to account for it (e.g. frequency threshold) provide only a limited test.

However, two caveats have to be mentioned prior the implementation of this strategy. First, random deletion does not just model random missing data; it also changes limited diversity and the relative position of the consistency threshold.[12] These are all important parameters to vary for the purpose of robustness, which makes this strategy even more interesting. Yet we cannot clearly disentangle them with our method. Second, random deletion of cases does not tackle a number of other error sources such as systematic error or conditioning error (Maggetti and Levi-Faur 2013).

[12] We thank Barry Cooper and Judith Glaesser for this remark.

Using the QCA package for R (Dusa and Thiem 2014), we simulate a large number of subsamples from the original data and assess the stability of the minimization results over these subsamples. We randomly simulate subsamples containing only 90 per cent of the observations (N=2091). Effectively, we randomly delete 10 per cent of the observations from the sample. Step-by-step the simulation resamples, re-runs the QCA minimization, saves the results and starts over again (in total 999 times). This strategy has some similarities with Hug's (2013) 'Monte Carlo' simulations that drop single cases from the QCA analysis. Yet, this strategy was implemented with a small-N dataset. As Ragin and Thiem (2013) show, this is a questionable strategy for data that is not generated at random. Accordingly, it is not surprising that a case-oriented method using macro-comparative data is case-sensitive. However, in our large-N survey data framework, a random deletion of cases is useful because respondents were randomly sampled.

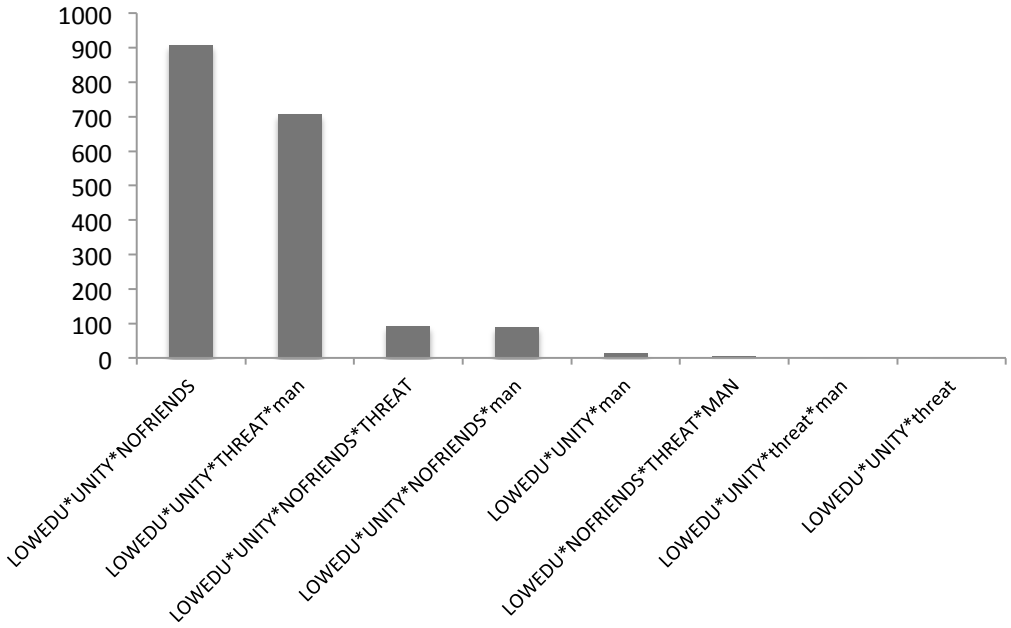
Our simulation procedure results in 999 QCA solutions. How can we now assess robustness using these results? Potentially, researchers might be interested in a number of the solution terms' characteristics such as the type and number of prime implicants (PIs) or consistency and coverage scores. In a first step we propose an easy-to-implement strategy that allows researchers to identify potentially important configurations not reported by the original solution due to random error.

Rather intuitively, a solution seems to be robust if it is composed of the same PIs over most of the simulated results. Hence, we simply count the number of times a PI appears in the 999 simulated solutions. Evidence for robustness is created if the most frequent PIs are the ones reported in the original solution. Ideally, we would like to see that the simulations do not report any PIs different from the ones we have in the original solution. Robustness suffers if we receive a great number of

PIs, which appear in the solutions at similar rates or if we have new, different PIs as the most frequent ones.

Figure 5 presents the frequency of all PIs appearing in the simulated solutions using the specifications from our original result in Table 2. The figure shows that our result is very stable against the random deletion of cases. The first PI from Table 2 appears in 908 of the 999 simulated solutions; the second PI 705 times. The third and fourth PIs appear 91 and 89 times and are subsets of the original solutions. The remaining PIs are new, but appear very rarely.

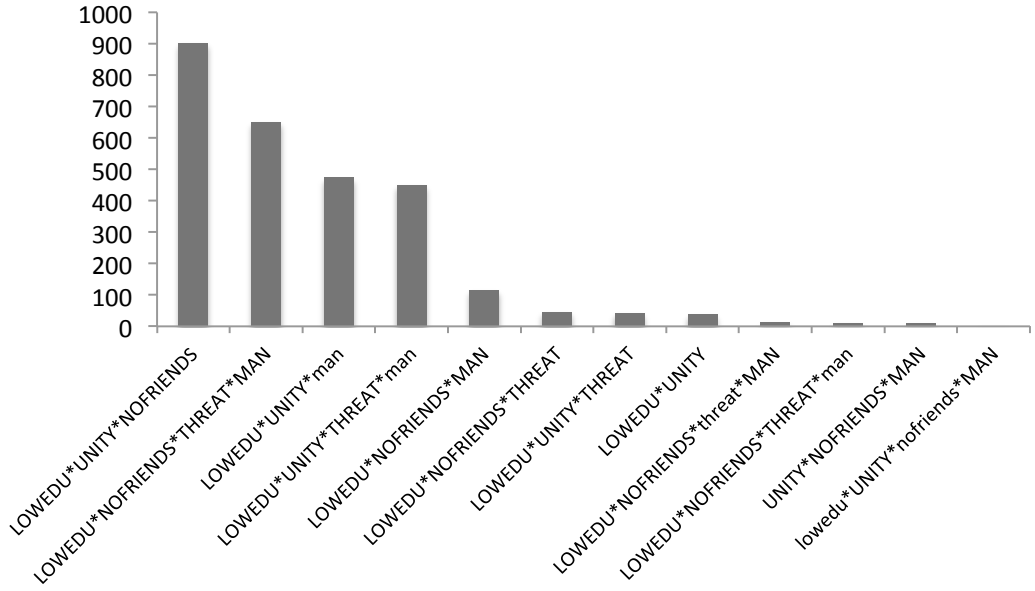
Figure 5: Frequency of PIs over simulated solutions (consistency threshold = 0.66)



While our original solution seems to be rather robust against random missing data and measurement error, Figure 6 presents an example of a less robust solution. Here, we simulated the results for the intermediate solution with the consistency threshold 0.62. Reflecting this lack of robustness, Figure 6 displays a comparatively large number of PIs with some frequently

occurring PIs such as "LOWEDU*UNITY*man" (appearing nearly 500 times in Figure 6) not being present in the original solution displayed in Table 3.

Figure 6: Frequency of PIs over simulated solutions (consistency threshold = 0.62)



Our simulation-based approach also allows for the analysis of the effect of random error on the distribution of PIs' measures of fit. Table 4 presents means, medians and standard deviations of the consistency, raw coverage and unique coverage scores of the PI "LOWEDU*UNITY*NOFRIEND" that was part of both the original solution (see Table 2) and the solution with the lowered consistency threshold (see Table 3). The first panel is based on the simulation results from Figure 5 (consistency threshold 0.66) and the second panel is based on the results from Figure 6 (consistency threshold 0.62). The measures of fit are all normally distributed and – except one – have low standard deviations. From this we can infer that our consistency and coverage values from Table 2 are rather robust against random error (see first panel of Table 4). Yet, the decreased stability of the solution using the 0.62 consistency threshold

is also reflected in the measures of fit: Unique coverage of the PI "LOWEDU*UNITY*NOFRIENDS" has now a rather high standard deviation of 0.053 (see lower panel in Table 4). By comparison, the standard deviation of this PI in the original solution (consistency threshold 0.66) is much smaller (see first panel).

Table 4: Summary statistics for simulated measures of fit

Consistency Threshold = 0.66			
	LOWEDU * UNITY * NOFRIEND		
	Consistency	Raw Coverage	Unique Coverage
Mean	0.758	0.312	0.182
Median	0.758	0.312	0.183
Std. Dev.	0.005	0.004	0.010
Consistency Threshold = 0.62			
	LOWEDU * UNITY * NOFRIEND		
	Consistency	Raw Coverage	Unique Coverage
Mean	0.758	0.312	0.064
Median	0.758	0.312	0.058
Std. Dev.	0.005	0.004	0.053

What can we conclude from our random deletion strategy? With respect to random error, our solution from Table 2 is rather robust. This finding is confirmed by the continuous presence of the two PIs in a vast majority of the simulated solutions and the concentrated distribution of the measures of fit. In contrast, the results we have obtained using the consistency threshold 0.62 are considerably less robust. Therefore, this finding seems to be contingent on the specific rows included under the lower consistency value. The consistency threshold 0.66 allows for robust findings, while the consistency threshold 0.62 does not.

To summarize the general insights from our random deletion procedure, we find that the procedure allows us to assess the stability of a solution with regard to random errors and that it helps us to discover frequent and substantially relevant PIs not reported by the original solution. A comparison of different simulation results allows us to identify the most robust solutions. Moreover, investigating the distribution of the measures of fit enables us to test whether the conclusions we draw (for example regarding unique coverage) are robust against random error.

Conclusion

This article has advanced the methodological discussion on large-N QCA on two important fronts. First, we have emphasized the unexploited potential of fsQCA in utilizing survey items in a theoretically informed way. In particular, we have argued that fuzzy sets are often able to represent the information included in survey items more adequately than ordinal- or interval-scaled variables. To our knowledge, nobody has pointed out the crucial differences between different types of large-N data and the advantages that come along with it for QCA. Previous methodological discussions about large-N QCAs focused mainly on the calibration-related problems of quantitative indicators. In contrast, we have demonstrated that it is necessary for future applications of large-N QCA to be aware of different types of large-N data and the opportunities they offer. In addition, we argue that previous strategies of set calibration leave room for improvement. Instead of simply using averages or factor analysis to combine indicators or survey items, we have presented examples of how sets can be calibrated in a theoretically more meaningful way. More precisely, we have demonstrated that (1) the calibration of a set should be more theoretically guided and (2) the combination of large-N data sources has to reflect the theoretical structure of the set, mainly by means of Boolean algebra.

Second, we have shown that large-N QCA struggles to establish causality because a crucial validation step – going back to the cases – is often not possible with anonymous large-N survey data. The main alternative to going back to the cases is robustness tests. Yet, our empirical example has shown that a more sophisticated tool-kit is needed to assess the robustness of QCA solutions. Hence, we have proposed a random deletion procedure that we believe is particularly suitable for robustness tests in case of large-N data generated through random sampling. Given the limited potential of standard robustness tests in case of large-N data, the random-deletion-of-cases strategy allows us to more adequately test a QCA solutions' robustness against random error. This is crucial for a QCA using large-N survey data, since here random error is likely to be pervasive.

Yet, our argumentation has to be set within clear boundaries. The analytical challenges highlighted in this paper are not a general problem of survey data per se. In case of small- and medium-N survey data, QCA researchers can often benefit from the advantages of QCA over regression-based approaches (e.g. the calibration of statements) without scarifying case orientation. Only in a research context using anonymous large-N information (e.g. computer-assisted telephone interviews or internet surveys), validation becomes difficult. Yet, we do not think that this should stop researchers from conducting large-N QCAs with survey data because with regard to calibration and measurement, respectively, QCA is better suited to deal with survey data than regression-based approaches. Therefore, we hope that our discussion can provide some guidance for QCA researchers on how to approach this critical issue.

References

Berg-Schlosser, Dirk. 2008. “Determinants of Democratic Successes and Failures in Afirca.” *European Journal of Political Research* 47(3): 269-306.

Berg-Schlosser, Dirk and Gisèle De Meur. 2009. “Comparative Research Design: Case and Variable Selection.” In *Configurational Comparative Methods. Qualitative Comparative Analysis (QCA) and Related Techniques*, edited by Benoît Rihoux and Charles Ragin. Los Angeles: SAGE, 19–32.

Cárdenas, Julián. 2012. “Varieties of Corporate Networks: Network Analysis and fsQCA.“ *International Journal of Comparative Sociology* 53(4): 298-322.

Ceobanu, Alin and Xavier Escandell. 2010. “Comparative Analyses of Public Attitudes Toward Immigrants and Immigration Using Multinational Survey Data: A Review of Theories and Research.” *Annual Review of Sociology* 36: 309–328.

Cheng, Cheng-Feng, Man-Ling Chang and Chu-Shiu Li. 2013. “Configural Paths to Successful Product Innovation.“ *Journal of Business Research* 66(12): 2561-2573.

Cooper, Barry. 2005. “Applying Ragin's Crisp and Fuzzy Set QCA to Large Datasets : Social Class and Educational Achievement in the National Child Development Study.” *Sociological Research Online* 10(2).

Cooper, Barry and Judith Glaesser. 2011. “Using Case-Based Approaches to Analyse Large Datasets: A Comparison of Ragin’s fsQCA and Fuzzy Cluster Analysis.” *International Journal of Social Research Methodology* 14(1): 31–48.

Crilly, Donal, Morten Hansen, Esben Pedersen, and Francesco Perrini. 2012. “Faking It or Muddling Through? Understanding Decoupling in Response to Stakeholder Pressures.” *Academy of Management Journal* 55(6): 1429–1449.

Crowley, Martha. 2013. “Class, Control, and Relational Indignity: Labor Process Foundations for Workplace Humiliation, Conflict, and Shame.” *American Behavioral Scientist* 58(3): 416-434.

Dustmann, Christian and Ian Preston. 2004. “Is Immigration Good or Bad for the Economy? Analysis of Attitudinal Responses.” *Research in Labor Economics* 24: 3-34.

Emmenegger, Patrick and Robert Klemmensen. 2013a. “What Motivates You? The Relationship between Preferences for Redistribution and Attitudes toward Immigration.” *Comparative Politics* 45(2): 227–246.

———. 2013b. “Immigration and Redistribution Revisited: How Different Motivations Can Offset Each Other.” *Journal of European Social Policy* 23(4): 406–422.

Emmenegger, Patrick, Jon Kvist, and Svend-Erik Skaaning. 2013. “Making the Most of Configurational Comparative Analysis: An Assessment of QCA Applications in Comparative Welfare State Research.” *Political Research Quarterly* 66(1): 185–190.

Engeli, Isabelle. 2012. “Policy Struggle on Reproduction: Doctors, Women, and Christians.“ *Political Research Quarterly* 65(2): 330-345.

Finseraas, Henning. 2008. “Immigration and Preferences for Redistribution: An Empirical Analysis of European Survey Data.” *Comparative European Politics* 6(3): 407–431.

Glaesser, Judith and Barry Cooper. 2010. “Selectivity and Flexibility in the German Secondary School System: A Configurational Analysis of Recent Data from the German Socio-Economic Panel.” *European Sociological Review* 27(5): 570–585.

Goertz, Gary. 2006. *Social Science Concepts: A User’s Guide*. Princeton: Princeton University Press.

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

Greckhamer, Thomas, Vilmos Misangyi, Heather Elms, and Rodney Lacey. 2008. “Using Qualitative Comparative Analysis in Strategic Management Research: An Examination of Combinations of Industry, Corporate, and Business-Unit Effects.” *Organizational Research Methods* 11(4): 695–726.

Greckhamer, Thomas, Vilmos Misangyi, and Peer Fiss. 2013. “The Two QCAs: From a Small-N To A Large-N Set Theoretic Approach.” In *Configurational Theory and Methods in Organizational Research*, edited by Peer Fiss, Bart Cambré, and Axel Marx. Emerald Group Publishing Limited, 49–75.

Grendstad, Gunnar. 2007. “Causal Complexity and Party Preference.” *European Journal of Political Research* 46(1): 121–149.

Hainmueller, Jens and Michael Hiscox. 2007. “Educated Preferences: Explaining Attitudes Toward Immigration in Europe.” *International Organization* 61(2): 399–442.

Hainmueller, Jens and Daniel Hopkins. 2013. “Public Attitudes Toward Immigration.” *Annual Review of Political Science* 17(1): 1–25.

Herreros, Francisco and Henar Criado. 2009. “Social Trust, Social Capital and Perceptions of Immigration.” *Political Studies* 57(2): 337–355.

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

Lakoff, George. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: University of Chicago Press.

Maggetti, Martino and David Levi-Faur. 2013. “Measurement Error in Qualitative Comparative Analysis.” *Political Research Quarterly* 66(1): 199–204.

Meuleman, Bart, Eldad Davidov, and Jaak Billiet. 2009. “Changing Attitudes toward Immigration in Europe, 2002–2007: A Dynamic Group Conflict Theory Approach.” *Social Science Research* 38(2): 352–365.

Pettigrew, Thomas. 1998. “Intergroup Contact Theory.” *Annual Review of Psychology* 49(1): 65–85.

Quillian, Lincoln. 1995. “Prejudice as a Response to Perceived Group Threat: Population Composition and Anti- Immigrant and Racial Prejudice in Europe.” *American Sociological Review* 60(4): 586–611.

Ragin, Charles. 1987. *The Comparative Method Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press.

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

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

Ragin, Charles and Peer Fiss. 2008. “Net Effects versus Configurations: An Empirical Demonstration.” In *Redesigning Social Inquiry. Fuzzy Sets and Beyond*, edited by Charles Ragin. Chicago: University of Chicago Press, 190–212.

Rydgren, Jens. 2008. “Immigration Sceptics, Xenophobes or Racists? Radical Right-Wing Voting in Six West European Countries.” *European Journal of Political Research* 47(6): 737–765.

Schedler, Andreas. 2012. “Judgment and Measurement in Political Science.” *Perspectives on Politics* 10(1): 21–36.

Schneider, Carsten and Claudius Wagemann. 2013. *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. Cambridge: Cambridge University Press.

Schneider, Martin, Conrad Schulze-Bentrop, and Mihai Paunescu. 2009. “Mapping the Institutional Capital of High-Tech Firms: A Fuzzy-Set Analysis of Capitalist Variety and Export Performance.” *Journal of International Business Studies* 41(2): 246–266.

Seawright, Jason. 2005. “Qualitative Comparative Analysis vis-à-vis Regression.” *Studies in Comparative International Development* 40(1): 3–26.

Senik, Claudia, Holger Stichnoth, and Karine Straeten. 2008. “Immigration and Natives’ Attitudes towards the Welfare State: Evidence from the European Social Survey.” *Social Indicators Research* 91(3): 345–370.

Sides, John and Jack Citrin. 2007. “European Opinion About Immigration: The Role of Identities, Interests and Information.” *British Journal of Political Science* 37(3): 477-504.

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

Thiem, Alrik and Charles Ragin. 2013. “Mill's Methods, Induction, Missing Data and Measurement Error in Qualitative Comparative Analysis.” mimeo.

Thygeson, Nels Marcus, Leif Solberg, Stephen Asche, Patricia Fontaine, Leonard Pawlson, and Sarah Hudson Scholle. 2012. “Using Fuzzy Set Qualitative Comparative Analysis (fs/QCA) to Explore the Relationship between Medical ‘Homeness’ and Quality.” *Health Services Research* 47(1): 22–45.

Vaisey, Stephen. 2007. “Structure, Culture, and Community: The Search for Belonging in 50 Urban Communes.” *American Sociological Review* 72(6): 851-873.

Van Oorschot, Wim and Wilfred Uunk. 2007. “European Countries Welfare Spending and the Public’s Concern for Immigrants: Multilevel Evidence for Eighteen European Countries.” *Comparative Politics* 40(1): 63–82.

Vis, Barbara. 2012. “The Comparative Advantages of fsQCA and Regression Analysis for Moderately Large-N Analyses.” *Sociological Methods & Research* 41(1): 168–198.