Working Paper 2019-92

QCA & the Robustness range of calibration thresholds: how sensitive are solution terms to changing calibrations?

Tore Hofstad

Abstract: Calibration procedures in Qualitative Comparative Analysis (QCA) rely on evaluation of cases, which necessarily entails a subjective element of interpretation. This questions the reliability of the inferences draw. To what extent are tests of sufficiency contingent on the researchers' choice of calibration? The concept robustness range is introduced to illuminate this question. Robustness range is tentatively defined as the calibration boundary within which solution terms and cases remain the same. Applying this concept in research practice increases transparency and adds to our methodological repertoire.

Keywords: **C**alibration, Research Transparency, Robustness

COMPASSS Working Paper 2019-92
Available at http://www.compasss.org/wpseries/Hofstad2019.pdf
COMPASSS Working Paper Series Managing Editor, Claude Rubinson

doi:xx.xxxx/xxxxx

QCA & the Robustness range of calibration thresholds: how sensitive are solution terms to changing calibrations?

TORE HOFSTAD *University of Oslo*

Calibration procedures in Qualitative Comparative Analysis (QCA) rely on evaluation of cases, which necessarily entails a subjective element of interpretation. This questions the reliability of the inferences drawn. To what extent are tests of sufficiency contingent on the researchers' choice of calibration? The concept robustness range is introduced to illuminate this question. Robustness range is tentatively defined as the calibration boundary within which solution terms and cases remain the same. Applying this concept in research practice increases transparency and adds to our methodological repertoire.

Word Count: 2089

BACKGROUND

Since its introduction by Charles Ragin (1987), QCA as a technique and approach has been portrayed as a middle way between qualitative and quantitative methods[1]. While Ragin's original procedure was limited to binary Boolean variables, his subsequent development and refinement of the method through the use of fuzzy logic, allows for more fine graded categorisation based on differences in degree (Ragin 2000, 2008). This procedure involves additional qualitative judgments on the part of the researcher. Correspondingly, criticism regarding the validity of QCA results has been voiced, particularly from scholars with a quantitative background (Lieberson 2004; Seawright 2005; Glaesser and Cooper 2014; Collier 2014; Krogslund, Choi, and Poertner 2015). These criticisms have helped the progress of QCA through clarifications of theoretical foundations (De Meur, Rihoux, and Yamasaki 2009; Goertz and Mahoney 2012; Rohlfing and Schneider 2014; Rohlfing 2016), and the development of new techniques and procedures (for an overview see Thomann and Maggetti 2017, p. 2).

One set of responses to these challenges are robustness tests, which have increasingly been recognised as an integral part of "best practice" (Wagemann and Schneider 2015), although consensus

Tore Hofstad is PhD fellow at the Center for Medical Ethics, University of Oslo, (tore.hofstad@medisin.uio.no)

AUTHOR'S NOTE: Thanks to Carsten Schneider for encouraging me to write this paper, to Claude Rubinson and the two anonymous reviewers from compasss.org for valuable feedback, and to Olav Schram Stokke for sharing his perspectives, which helped improve the paper considerably.

[1] I assume that the target audience is familiar with the basics of fuzzy set QCA. For an introduction, see Ragin (2008) or Schneider and Wagemann (2012). All analyses are performed in R version 3.5.2 (R Core Team 2018), using the "QCA"-package, version 3.4 (Duşa 2019). Plots are created using "ggplot2" (Wickham 2016).

has not been reached regarding which robustness tests are deemed relevant (Schneider, Vis, and Koivu 2019). The purpose of this paper is to outline an alternative approach to testing the robustness of calibrations, that could ideally form an integral part of the whole iterative process that is QCA.

Calibration

Calibration is the procedure of categorising cases to sets, by assigning partial membership, based on relevant variation in the raw data. The direct method of calibration described by Ragin (2000) employs three qualitative anchors for establishing set membership. One threshold beyond which cases have *full membership* in the set, yielding the fuzzy value 1; another demarcating *full non-membership* which results in a fuzzy value of 0; and finally, the crossover point which establishes whether a case is more *in* than *out* of a set. The fuzzy value for the crossover point is 0.5 and represents maximum ambiguity or 'fuzziness'. Calibrated measures can be claimed to contain additional information, beyond the original raw data, as they not only allow us differentiate between "less versus more", but also "a little versus a lot" (Ragin 2008).

When calibrating we are advised to seek and provide theoretical justifications for the qualitative anchors, preferably on the basis of sources outside the empirical data to be analysed (Ragin 2008; Schneider and Wagemann 2012). Still, the researcher may arrive at several seemingly viable alternatives, by going back and forth between theory and cases, since QCA is an inherently iterative process. Interpretation and qualitative judgment is thus constitutive to the calibration process, which is why questions of robustness have repeatedly been raised (Skaaning 2011; Schneider and Wagemann 2012; Cooper and Glaesser 2016).

INTRODUCING THE ROBUSTNESS RANGE

Previous approaches to assessing the robustness of calibrations have applied various calibration thresholds, and examined to what extent results change as a consequence (Skaaning 2011; Schneider and Wagemann 2012; Thiem, Spöhel, and Duşa 2016; Cooper and Glaesser 2016). The procedure I suggest takes a somewhat different route and uses the minimised sufficiency solution as starting point[2]. By identifying the exact calibration thresholds where either solution terms or cases change, we uncover precisely how far the results hinge on a particular choice of calibration, as well as the location of the boundary where a qualitative difference is found. The robustness range can thus be defined as *the calibration boundary within which solution terms and cases remain the same*. In other words, the concept encompasses a region for calibration wherein minimised solutions remain identical and each

[2]It is not relevant for this particular exposition whether the researcher performs analyses of necessity prior to sufficiency (Schneider and Wagemann 2010), whether she prefers some kind of (enhanced) standard analysis (Schneider and Wagemann 2012), or only accepts the most parsimonious solution (Baumgartner and Thiem 2017). The robustness range can equally well be calculated for any solution formula.

term covers the exact same cases.

Knowledge of the width of the robustness range will bolster or reduce the researchers confidence in the results. If a solution term only appears within a narrow calibration range, the researcher would be advised to provide compelling reasons for that particular choice of calibration. On the other hand, if the solution term remains the same when a wide range of thresholds are chosen, the result can be considered robust to change in calibration and likely to be more trustworthy. This would be particularly reassuring in situations where a sound theoretical basis for calibration is missing (Thomann and Maggetti 2017).

METHOD

In order to calculate the robustness range I manually alter the calibration threshold for one condition at a time. I repeatedly increase the crossover point slightly, while rerunning the analysis of sufficiency, until the solution term changes, or one or more cases become uncovered by the solution term. I then lower the threshold until finding the exact upper boundary of the robustness range for the crossover point, with a precision of two or three decimals. The process is then repeated for the lower boundary for the crossover, and then for the thresholds for full membership and non-membership, for all conditions as well as the outcome.

EMPIRICAL EXAMPLE

To illustrate the procedure I replicate parts of the fsQCA of Cebotari and Vink (2013), who investigated variation in ethnic political protest (EP) in Europe. They apply five theoretically relevant conditions: territorial concentration (TC); level of democracy (LD); ethnic fractionalisation (EF); political discrimination (PD); and national pride (NP). The most parsimonious solution identifies four sufficient paths to strong protest:

$$\sim NP + (TC * LD * EF) + (TC * PD * EF) + (TC * PD * LD) \rightarrow EP \qquad (1)$$

The absence of national pride is alone sufficient for strong ethnic protest. Furthermore the following three configurations are considered sufficient for the outcome: a high degree of territorial concentration in combination with either high levels of democracy and fractionalisation; or discrimination and fractionalisation; or discrimination and democracy.

Interpreting the robustness range

Table 1 shows the calibration thresholds used in Cebotari and Vink's article, along with the calculated robustness range for the most parsimonious solution. This information is also visualised in figures 1, 2 & 3, which shows the rank ordering of the raw data for each case. In all figures the solid horizontal lines show the calibration used in the article, encapsulated by the robustness range; green for full

TABLE 1. Calibration thresholds with robustness range for parsimonious solution

Set/Threshold	Calibration	Robustness range
Ethnopolitical protest		
Full Membership:	3	(1.5 - 5*)
Crossover Point:	1.5	(1.24 - 1.99)
Full Non-Membership:	0.5	(0.39 - 0.67)
Level of Democracy		
Full Membership:	9.5	(7 - 10*)
Crossover Point:	7	(6.97 - 7.19)
Full Non-Membership:	2	(0 - 6.99)
Ethnic fractionalisation		
Full Membership:	0.8	(0.495 - 1*)
Crossover Point:	0.495	(0.35 - 0.499)
Full Non-Membership:	0	(0 - 0.275)
Territorial concentration		
Full Membership:	3	(1.5 - 3*)
Crossover Point:	1.25	(1.01 - 1.96)
Full Non-Membership:	0	(0 - 0.72)
Political discrimination		
Full Membership:	3	(0.75 - 4*)
Crossover Point:	0.75	(0.001 - 0.99)
Full Non-Membership:	0	(0 - 0.49)
National pride		
Full Membership:	2.5	(1.5 - 3*)
Crossover Point:	1.5	(1.42 - 1.55)
Full Non-Membership:	0.5	(0 - 1.09)
Solution consistency	0.879	(0.736 - 0.902)
Solution coverage	0.811	(0.733 - 0.896)

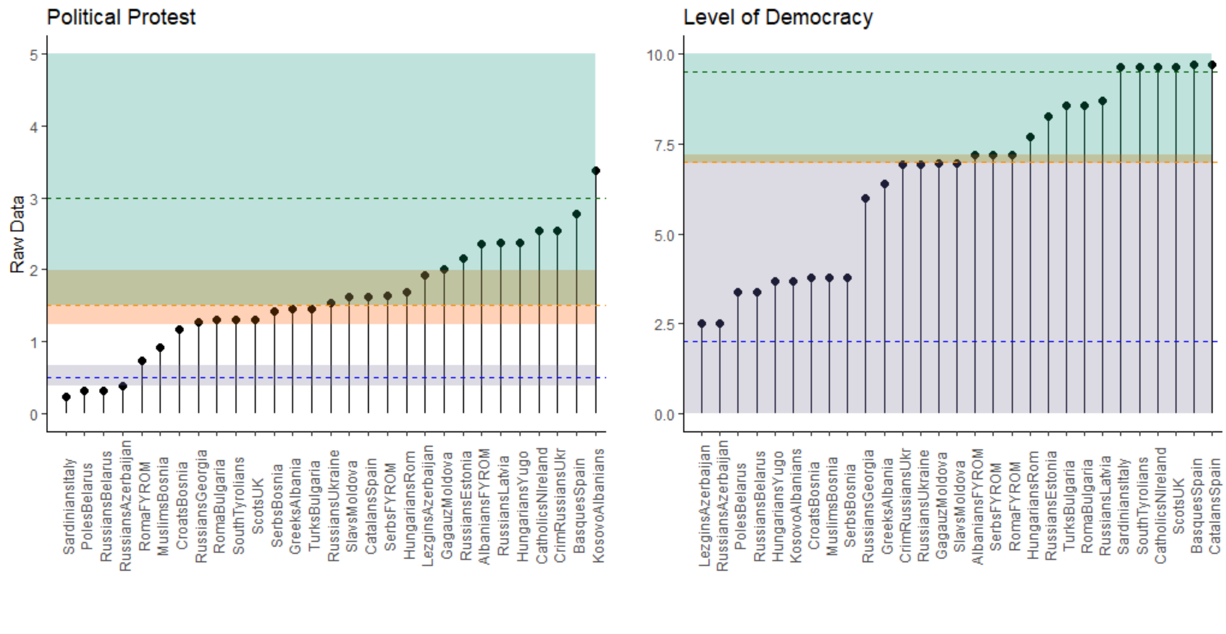
Note: * The upper range exceeds the scale.

membership, orange for the crossover, and purple for full non-membership.

The left array of figure 1 shows that the threshold for full non-membership in the outcome is limited to the parameters between two specific cases, namely Russians in Azerbaijan and Roma in Macedonia, where the former are fully out of the set that display ethnopolitical protest, while the latter are more out than in the set. For all conditions, the threshold for full membership can be set considerably higher without changing solutions, although it would rarely make sense to place the anchor beyond the scale limit.

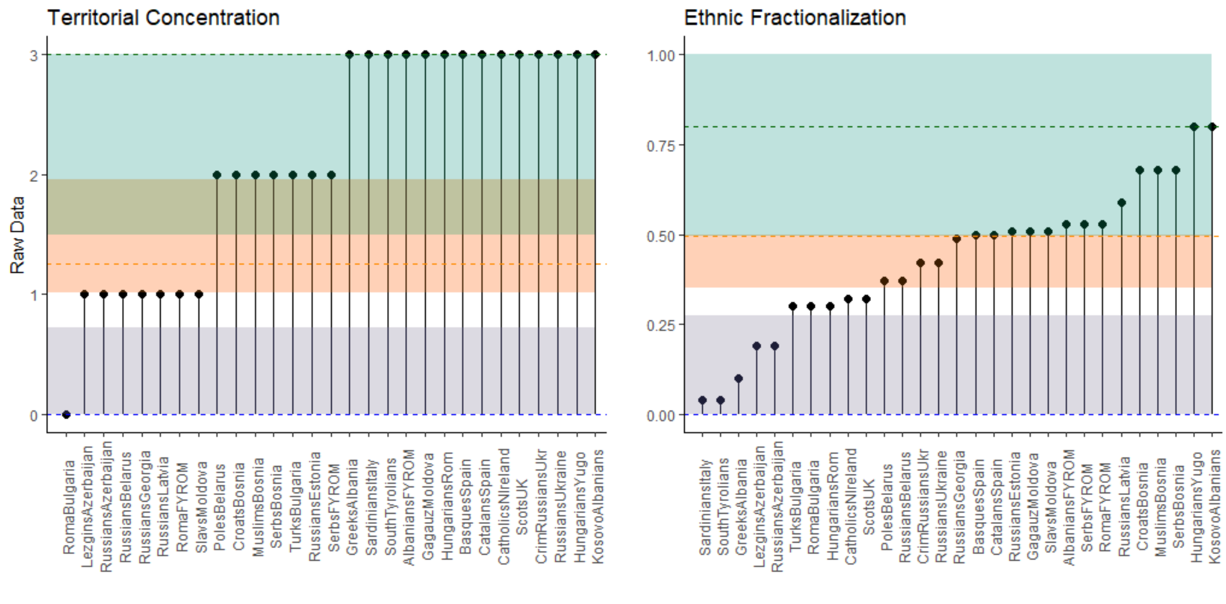
Concerning the crossover points, two of six robustness ranges are noteworthy narrow: level of democracy (figure 1) and national pride (figure 3). A marginal change in these calibrations will thus alter the solution formula and/or the cases covered by each term. For example, lowering the crossover

FIGURE 1. Calibration thresholds with robustness range



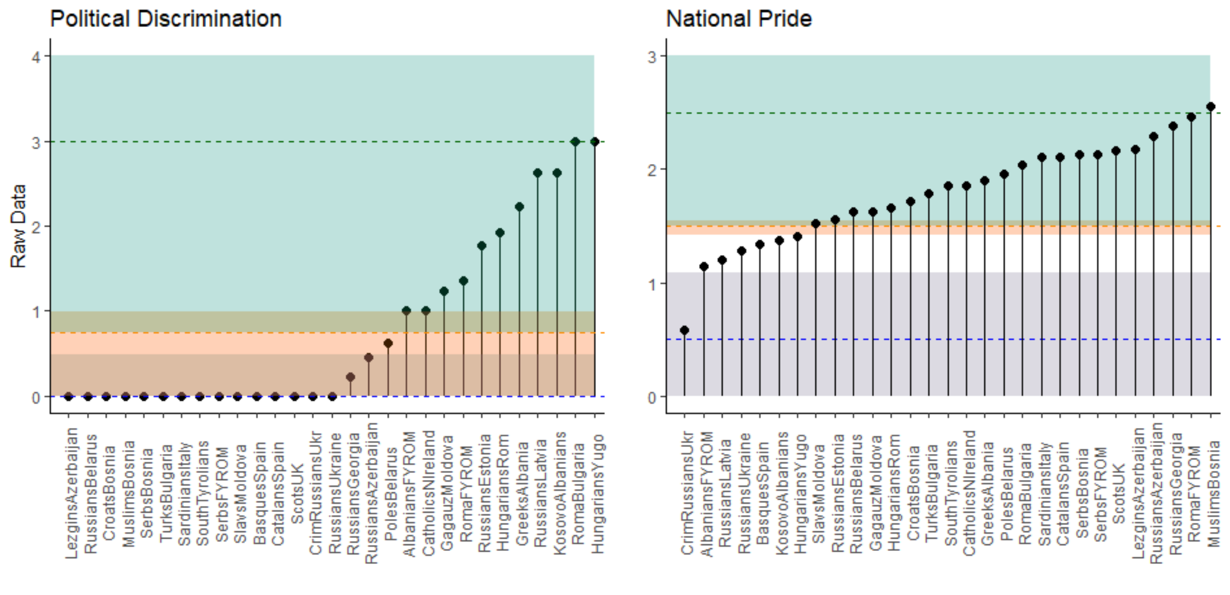
Note: Horizontal lines represent calibrations used by Cebotari & Vink.

FIGURE 2. Calibration thresholds with robustness range



threshold for level of democracy from 7 to 6.95, will produce a substantially different solution formula. Consequently the findings from the article are contingent on the evaluation that Moldova belongs more out than in the group of countries with a high level of democracy at the time of the study, and that there was a relevant difference between the state of democracy in Moldova and Macedonia.

FIGURE 3. Calibration thresholds with robustness range



The democracy condition is particularly interesting because the robustness ranges for full membership and non-membership span the entire range. I reran the analysis with level of democracy calibrated as a crisp condition with crossover at 7, which resulted in the same solution, with only slightly lower consistency and coverage. This suggests that differences in *level* of democracy do not matter for facilitating ethnopolitical protest, according to this solution; only whether it is present or absent.

DISCUSSION

The primary aim of introducing the robustness range for calibration is to provide a new tool to the armoury of set-theoretic methods. By calculating the robustness range after a satisfactory sufficiency solution has been found, we discover just how (in-) vulnerable the solution is to precise changes in calibration. Furthermore this may enhance our conceptual understanding of the set-theoretic relations involved. Consequently the calculation of the robustness range should not be considered an endpoint, but can meaningfully be included as part of the overall iterative process of QCA.

A narrow robustness range signifies that a minor adjustment is enough to cause a difference in solution terms or cases covered. This may not necessarily alter substantial interpretation - if so, all the better. On the other hand, if a minor change significantly alters our conceptual understanding of the set relations, the researcher is encouraged to reconsider the choice, or provide an explicit description that the solution rests on a highly particular understanding of the concept.

In situations where the robustness range is strikingly *wide*, as we observed with level of democracy in the empirical example, it might turn out that differences in degree are irrelevant, and only differences in kind matter. This should in turn inform the substantial interpretation.

The robustness range might prove particularly beneficial in intermediate to large-*N* scenarios, where in-depth case-knowledge is not always attainable or relevant (Emmenegger, Schraff, and Walter 2014). Armed with impeccable case knowledge, a narrow robustness range is likely to cause less concern, since the researcher would be more likely to have compelling reasons for her choice. On the other hand, lacking relevant case-knowledge or a sound theoretical basis for anchoring, a narrow robustness range makes for a less persuasive narrative. Although possibly less valuable, calculating the robustness range can still be relevant in small-*N* research, since it reveals to the researcher and the audience the exact extent to which a particular solution hinges on a specific choice of calibration threshold.

Uncertainty is constitutive to science, so we are generally well-advised to articulate the trustworthiness of our findings[3]. The robustness range increases transparency, and can substantiate a particular choice of threshold. If a case is removed from a solution term as a consequence of change in threshold, the researcher should apply her case knowledge to the particular case and evaluate whether the case actually merits full or partial (non-) membership in each set.

The robustness range can also have implications for generalisability, when applying the same calibration thresholds to a different population. Perhaps situating the anchor in the middle of the robustness range will be more representative than choosing a value in the outskirts?

Measures of solution consistency and coverage vary within the robustness range, as can be seen in table 1. The highest solution consistency is found when the crossover threshold for the outcome is anchored at the lower boundary of the robustness range. Since the measures of consistency and coverage are inverse, the highest coverage is obtained at the upper boundary of the robustness range for the outcome. This variation provides us with a range of measures rather than a point estimate, yielding additional information beyond that which is conventionally reported. This feature could be useful if the researcher, for any reason, desires to maximise either consistency or coverage values for the overall formula. Arguably this does not change anything about the cases involved or their set-relations, but it can influence how we think about them.

Since crisp set QCA is a special case of fsQCA (Schneider and Wagemann 2012), the robustness range can meaningfully be applied here as well, although naturally limited to the crossover anchor. Similarly, robustness ranges can also be calculated for fuzzy sets calibrated by the indirect method.

A limitation of the approach sketched here is that it only considers threshold change in one condition at a time. This means that the robustness range for each condition is only valid all other things being equal. Moreover, the replication was performed to illustrate the possible usefulness of the procedure, and is not meant to question the authors' choice of calibrations. For that reason I have not replicated the QCA for the negation of the outcome. Since this is meant as an introduction to the concept, the definition is tentative. The explorative researcher may decide to calculate an alternative, less strict robustness range, which allows for cases to fluctuate, as long as the sufficiency solution remains the same. Time will tell which of the definitions prove most useful for applied research.

[3]Note that this procedure only concerns calibration thresholds, thus separate tests of robustness for choice of frequency and consistency thresholds should also be run (Skaaning 2011).

Finally, I have not attempted to systematically investigate whether there are recurring characteristics of situations where we find a wide or a narrow robustness range, or whether a particular distribution of cases, such as skewed sets, has a predictable impact on the robustness range. This may prove a fruitful avenue for future research. As a further suggestion, it would be laudable if any developers of set-theoretic packages for R would take it upon themselves to automate the process of calculating the robustness range.

CONCLUSION

The robustness range for calibrations displays the exact limits for qualitative anchors within which sufficiency solutions and cases remain identical. The calculation is performed after the truth table analysis and forms part of the whole iterative process that is QCA. Increased transparency provides a better platform for evaluating trustworthiness, for both researcher and audience, particularly when a strong theoretical basis for calibration is lacking. A range of potential values allows for maximisation of either consistency or coverage of the solution formula. By broadening, or limiting, the conceptual boundaries, the robustness range can extend or reduce the way we think about explicit set-relations.

REFERENCES

Baumgartner, Michael, and Alrik Thiem. 2017. “Often trusted but never (properly) tested: Evaluating Qualitative Comparative Analysis”. *Sociological Methods & Research*: 0049124117701487.

Cebotari, Victor, and Maarten P Vink. 2013. “A configurational analysis of ethnic protest in Europe”. *International Journal of Comparative Sociology* 54 (4): 298–324.

Collier, D. 2014. “Problematic Tools: Introduction to Symposium on Set Theory in Social Science”. *Qualitative & Multi-method Research, 12 (1).*

Cooper, Barry, and Judith Glaesser. 2016. “Exploring the robustness of set theoretic findings from a large n fsQCA: an illustration from the sociology of education”. *International Journal of Social Research Methodology* 19 (4): 445–459.

De Meur, Gisèle, Benoît Rihoux, and Sakura Yamasaki. 2009. “Addressing the critiques of QCA”. In *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques*, ed. by Benoît Rihoux and Charles Ragin, 147–165. Sage Thousand Oaks, CA.

Duşa, Adrian. 2019. *QCA with R. A comprehensive resource*. Springer International Publishing.

Emmenegger, Patrick, Dominik Schraff, and André Walter. 2014. “QCA, the truth table analysis and large-N survey data: The benefits of calibration and the importance of robustness tests”. In *COMPASSS Working Papers Series 2014-79*.

Glaesser, Judith, and Barry Cooper. 2014. “Exploring the consequences of a recalibration of causal conditions when assessing sufficiency with fuzzy set QCA”. 17 (4): 387–401.

Goertz, Gary, and James Mahoney. 2012. *A tale of two cultures: Qualitative and quantitative research in the social sciences*. Princeton University Press.

Krogslund, Chris, Donghyun Danny Choi, and Mathias Poertner. 2015. "Fuzzy sets on shaky ground: Parameter sensitivity and confirmation bias in fsQCA". *Political Analysis* 23 (1): 21–41.

Lieberson, Stanley. 2004. "Comments on the Use and Utility of QCA". 2 (2): 13–14.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ragin, Charles C. 1987. *The comparative method: Moving beyond qualitative and quantitative strategies*. Univ of California Press.

— . 2000. *Fuzzy-set social science*. Univ of California Press.

— . 2008. *Redesigning social inquiry: Fuzzy sets and beyond*. Univ of Chicago Press: Chicago / London.

Rohlfing, Ingo. 2016. "Why simulations are appropriate for evaluating Qualitative Comparative Analysis". *Quality & quantity* 50 (5): 2073–2084.

Rohlfing, Ingo, and Carsten Q Schneider. 2014. "Clarifying Misunderstandings, Moving Forward: Towards Standards and Tools for Set-Theoretic Methods". *Qualitative and Multi-Method Research* 12 (2): 27–33.

Schneider, Carsten Q., Barbara Vis, and Kendra Koivu. 2019. *Set-Analytic Approaches, Especially Qualitative Comparative Analysis (QCA)*. Tech. rep. American Political Science Association Organized Section for Qualitative and Multi-Method Research, Qualitative Transparency Deliberations, Working Group Final Reports, Report III, 4.

Schneider, Carsten Q., and Claudius Wagemann. 2010. "Standards of good practice in qualitative comparative analysis (QCA) and fuzzy-sets". *Comparative Sociology* 9 (3): 397–418.

— . 2012. *Set-Theoretic Methods for the Social Sciences*. Cambridge University Press.

Seawright, Jason. 2005. "Qualitative Comparative Analysis vis-à-vis Regression". 40 (1): 3–26.

Skaaning, Svend-Erik. 2011. "Assessing the robustness of crisp-set and fuzzy-set QCA results". *Sociological Methods & Research* 40 (2): 391–408.

Thiem, Alrik, Reto Spöhel, and Adrian Duşa. 2016. "Enhancing sensitivity diagnostics for qualitative comparative analysis: A combinatorial approach". *Political Analysis* 24 (1): 104–120.

Thomann, Eva, and Martino Maggetti. 2017. "Designing Research With Qualitative Comparative Analysis (QCA) Approaches, Challenges, and Tools". *Sociological Methods & Research*: 0049124117729700.

Wagemann, Claudius, and Carsten Q. Schneider. 2015. "Transparency standards in qualitative comparative analysis". *Qualitative & Multi-Method Research* 13 (1): 38–42.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. http://ggplot2.org.