Original Article

Sociological Methods & Research
1-33

DOI: 10.1177/0049124120914955
journals.sagepub.com/home/smr
SAGE

Applying and Assessing Large-N QCA: Causality and Robustness From a Critical Realist Perspective

Roel Rutten[1,2,3]

Abstract

Applying qualitative comparative analysis (QCA) to large *N*s relaxes researchers' case-based knowledge. This is problematic because causality in QCA is inferred from a dialogue between empirical, theoretical, and case-based knowledge. The lack of case-based knowledge may be remedied by various robustness tests. However, being a case-based method, QCA is designed to be sensitive to such tests, meaning that also large-*N* QCA robustness tests must be evaluated against substantive knowledge. This article connects QCA's substantive-interpretation approach of causality to critical realism. From that perspective, it identifies relevant robustness tests and applies them to a real-data large-*N* QCA study. Robustness test findings are visualized in a robustness table, and this article develops criteria to substantively interpret them. The robustness table is introduced as a tool to substantiate the validity of causal claims in large-*N* QCA studies.

Keywords

qualitative comparative analysis (QCA), large-*N*, robustness, validity, causality, critical realism

[1] Tilburg School of Social and Behavioral Sciences, Tilburg University, the Netherlands
[2] European Regional Affairs Consultants (ERAC), 's-Hertogenbosch, the Netherlands
[3] Northumbria University, Newcastle, United Kingdom

Corresponding Author:
Roel Rutten, Tilburg School of Social and Behavioral Sciences, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, the Netherlands.
Email: r.p.j.h.rutten@uvt.nl

Qualitative comparative analysis (QCA) was designed for midsize *N*s (10–50 cases) but is increasingly applied to (much) larger samples. However, large-*N* QCA is problematic because it defeats one of QCA's cornerstones: in-depth case knowledge that allows researchers to interpret empirical findings into causal mechanisms (Greckhamer, Misangyi, and Fiss 2013; Ragin and Fiss 2008). The truth-table analysis was never intended to produce causal statements as such. Cross-case regularities, expressed as consistent set relationships, must be interpreted and verbalized into causal mechanisms based on substantive knowledge (Ebbinghaus 2005; Ragin 2008; Schneider 2018; Schneider and Wagemann 2012). Lacking such knowledge, large-*N* QCA risks inferring causality from metrics alone (e.g., Cooper and Glaesser 2016). This is a bad idea because QCA has no universally agreed upon thresholds for these metrics; they must be derived from substantive and theoretical knowledge. To remedy the lack of in-depth case knowledge, the literature suggests various robustness tests for large-*N* QCA, such as changing consistency and frequency thresholds, calibrations, and randomly deleting cases (Emmenegger, Schraff, and Walter 2014; Maggetti and Levi-Faur 2013; Skaaning 2011). The solution is considered robust when it is also identified (or closely approximated) after the truth-table analysis has been redone with different thresholds, calibrations, and samples (Schneider and Wagemann 2012:285). However, these robustness tests ignore that QCA is designed to be sensitive to such changes (Emmenegger et al. 2014; Skaaning 2011). That being said, large-*N* QCA weakens the link between thresholds, calibrations, and cases, on the one hand, and theoretical and substantive knowledge, on the other hand, to the extent that multiple thresholds and calibrations are plausible and one would like to see robust findings for alternate settings. This article addresses the plausibility of set relationships from a substantive-interpretation approach. This approach emphasizes that (1) set relationships and the causal claims they support must be interpreted by "triangulating" empirical, substantive, and theoretical knowledge and that (2) the noisy nature of social reality makes that data are never perfect (because of limited diversity and unknown causes) and that methods must be designed to deal with that (Ragin 2000, 2008). Redundancy-free QCA, instead, follows a regularity approach to causality, benchmarked against error-free data and fully specified truth tables (Baumgartner 2015; Thiem, Baumgartner, and Bol 2016; Schneider 2018; Thomann and Maggetti 2017). A substantive-interpretation approach connects seamlessly to critical realism and its notion of causality as generative mechanisms. Generative mechanisms of the critical realist kind make outcomes possible but do not determine their occurrence. Generative mechanisms cannot be observed directly but must be

inferred from empirical observations (Bhaskar [1975] 2008:148; Byrne 2009:172; Collier 1994:27; Gerrits and Verweij 2013:105). This dovetails with QCA's position that empirically identified configurations must be interpreted into causal mechanisms and that inconsistent cases do not defeat causality as long as the consistency of the set relationship is high enough (Schneider and Wagemann 2012:127). Critical realism and the notion of generative mechanisms are used in this article to identify relevant robustness tests for large-*N* QCA.

This article aims to develop a tool to assess and interpret the robustness of large-*N* QCA studies. To that end, the article discusses existing robustness tests, introduces new ones, and develops criteria to interpret them. The article suggests robustness tests in large-*N* QCA as a process of "updating" or improving the (empirical) validity of the truth-table analysis in a way that mirrors the practice of "going back to the cases" of small-*N* QCA. The article applies robustness tests to an actual large-*N* QCA study and introduces the robustness table to report robustness test findings. Some would consider 108 cases a medium-size *N*, but the point is that 108 cases exclude the in-depth case knowledge on which small-*N* QCA relies. This article is not the first to suggest robustness tests for QCA. However, this article's critical realist and substantive-interpretation notion of causality take robustness beyond the mathematical tests that dominate the literature (cf. e.g., Baumgartner 2015; Hug 2013; Krogslund, Donghyun, and Poertner 2015; Rohlfing 2016). The literature correctly demands that (large-*N*) QCA, like any other method, subjects itself to robustness tests. However, meaningful robustness tests must be aligned with the ontological and epistemological assumptions underlying the method (Beach and Pedersen 2016:15).

The remainder of this article is structured as follows. The next section introduces the study on openness values and regional innovation. For the full study, see Rutten (2019). Readers are referred to that paper to appreciate the context on which the robustness tests in this article rely. The following two sections discuss how causality and robustness are understood from QCA's substantive-interpretation perspective and how this connects to critical realism. From this perspective, the article explains what construct, internal, and external validity mean in a QCA context and suggests relevant robustness tests. The next sections discuss both analytical and empirical robustness issues for these validities and apply empirical robustness tests to the openness values and regional innovation study. The various robustness test findings are summarized and interpreted in the section on the robustness table. The final section draws the argument of the article together in a discussion and conclusion.

The Study: Openness Values and Regional Innovation

The economic geography literature suggests that openness values are important for regional innovation. "Tolerance" allows knowledge exchange between diverse communities in society, and "self-expression" encourages individuals to pursue new knowledge and ideas. That is, openness values allow regional innovation to benefit from a larger pool of knowledge and ideas. However, regional innovation is also explained without explicit reference to openness values. This suggests an equifinality problem where openness values sometimes contribute to regional innovation and sometimes they do not. In this study, openness values are conceptualized as "melting pot," the intersection of the sets of tolerant regions and ethnically diverse regions, and "self-expression," the intersection of the sets of personal achievement (i.e., postmaterialist values) regions and personal freedom and choice regions. The other three conditions are "analytic knowledge creation" (science), "synthetic knowledge creation" (application), and "economic diversity." The outcome is "regional innovation." Analytic and synthetic knowledge creation may connect differently to openness values; economic diversity reflects the exchange of knowledge and ideas between different sectors of the economy. Data were sourced from publicly available databases from the European Commission. The analysis was performed on 108 regions in North-West Europe (Sweden, Denmark, Germany, the Netherlands, Belgium, and the United Kingdom) because these regions differ in degree with one another on economic and institutional development but differ in kind on those factors from regions in Southern, Central, and Eastern Europe, thereby ruling out these factors as explanations for regional innovation and setting important scope conditions. Using the fuzzy set qualitative comparative analysis (fsQCA) Version 3.0 software (Ragin and Davey 2017), the solution produced four equifinal configurations, three of which include openness values (Figure 1). This solution was arrived at by going back and forth between empirical, substantive, and theoretical knowledge, which resulted in recalibrations and setting different thresholds in the truth table (Schneider and Wagemann 2012:281). In the remainder of the article, this solution is referred to as the "choice solution." All four configurations in the choice solution are corroborated by theoretical knowledge and by case studies available in the economic geography literature. Most of these case studies are from regions included in the sample. Some Australian cases suggest that findings may be generalized to other economically and institutionally developed regions.

	Choice Calibration					Continuous Calibration					Conservative Calibration					Lenient Calibration				
configuration	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V	I	II	III	IV	V
Analytical Knowledge		●	●	●			●	●	●	⊗		●	●				●	●	●	⊗
Synthetic Knowledge		●	●					●		⊗		●	●		⊗	●	●	●		⊗
Economic Diversity	●	●		●			●		●	●					●	●	●		●	●
Melting Pot	●		●					●								●		●	●	
Self-Expression				●						⊗					⊗				●	⊗
Solution coverage	0,789474					0,824586					0,874286					0,838983				
Solution consistency	0,944056					0,949719					0,905325					0,961165				
Frequency cut off	5 (78% of cases)					5 (92% of cases)					4 (77% of cases)					4 (83% of cases)				
Truth table rows	9 rows					12 rows					8 rows					11 rows				

Figure 1. Solutions in the choice and alternate calibrations.

Causality, Validity, and Robustness

In small-*N* QCA, validity of causal claims follows from in-depth case knowledge. It reduces measurement error (incorrect calibration), turns logical remainder rows into easy (plausible) and difficult (implausible) counterfactuals, and generally strengthens the plausibility of causal claims. In-depth case knowledge also sets scope conditions for causal claims (Goertz 2017:60; Maggetti and Levi-Faur 2013; Olsen 2014; Ragin 1992:221). That is why QCA researchers are adamant about "going back to the cases." It is not just a peculiarity of the method but its essence (Ebbinghaus 2005; Emmenegger et al. 2014; Greckhamer et al. 2013; Ragin 2008), an essence poorly understood by critics (e.g., Lucas and Szatrowski 2014; Seawright 2014) and little appreciated by large-*N* applications (e.g., Álvarez-Coque, Mas-Verdú, and Roig-Tierno 2017; Cooper and Glaesser 2016). Absent in-depth case knowledge, robustness tests must assess the empirical plausibility of set relationships and the causal claims they support.

Since robustness and causality are intimately linked, a discussion of robustness without explaining what causality means in (substantive-interpretation) QCA is problematic. Goertz and Mahoney (2012) distinguish between variable-based methods (effects-of-causes methods that start with the causes and predict the effects) and case-based methods (causes-of-effects methods that start with the outcome and detect the causes). QCA detects causes based on observed cross-case regularities and follows a difference-making method to eliminate redundant causes (Beach and Pedersen 2016:160; Wagemann and Schneider 2010). Critically, cross-case regularities (consistent set relationships) are not themselves causal mechanisms; they are empirical manifestations of underlying causal mechanisms that must be interpreted into statements of sufficiency by "going back to the cases." QCA cannot be reduced to a Boolean algebraic exercise in truth-table minimization (Ebbinghaus 2005; Rohlfing and Schneider 2018; Skaaning 2011). This means that QCA is not a regularity method. Regularities are indicative of but do not evidence causality. This dovetails with the critical realist position that causal mechanisms are (ontologically) real, irrespective of whether they are "triggered" or whether they produce outcomes. Consequently, the empirically observed implications of causal mechanisms (cross-case regularities in the form of configurations; i.e., epistemological statements) cannot be reduced to statements about the causal mechanisms themselves (i.e., ontological statements; Bhasker [1975] 2008:16-17; Byrne 2009:102-03, Collier 1994:36, 76; Gerrits and Verweij 2013:170).

The literature is divided on the definition and nature of causal mechanisms (Goertz and Mahoney 2012:101; Hedström and Ylikoski 2010). An important common element in many definitions is an emphasis on different "elements" or "concepts" that work together to "trigger," "produce," or "generate" and outcome (Goertz 2017:33–34). That is, causal mechanisms imply "interaction" of multiple individual causes, which distinguishes causal mechanisms from causal effects of individual variables (Goertz and Mahoney 2012:43). Definitions disagree on the level of abstraction of a causal mechanism (Hedström and Ylikoski 2010:52). Process tracers (e.g., Beach and Pedersen 2016) suggest causal mechanisms as fine-grained "accounts" of how causes produce outcomes within cases. Goertz (2017) and Hedström and Swedberg (1998), on the other hand, suggest that causal mechanisms can have a higher level of abstraction. They are interpreted from a combination of within-case (proof of existence) and cross-case (proof of relevance) inferences (Goertz 2017:30). Causal mechanisms in QCA connect best to this analytical view. Empirically (epistemologically), a statement of sufficiency in QCA implies a cross-case regularity, substantively (ontologically) it means that the presence of conditions allows agents to act in ways that achieve the outcome. Put differently, causes make outcomes possible (Beach and Pedersen 2016:48; Goertz 2017:46, 154). Whether agents actually achieve the outcome is another matter. That is, causality is emergent, not deterministic (Bhasker [1975] 2008:50; Collier 1994:126–30). Unobserved causes may negate the willingness and/or ability of agents to achieve the outcome, even when sufficient conditions are present (i.e., inconsistent cases). The difference between substantive-interpretation and redundancy-free QCA lies in the question, what is sufficient? Redundancy-free QCA follows a purely mathematical logic whereas substantive-interpretation QCA may reject a mathematically consistent set relationship on the grounds that it is substantively uninterpretable (Schneider 2018; Thomann and Maggetti 2017). That is, substantive-interpretation QCA assesses the (substantive) plausibility of a set relationship, whereas redundancy-free QCA assumes that, absent explicit mathematical evidence to the contrary, all truth-table rows are sufficient for the outcome (Baumgartner 2015:583). This makes redundancy-free QCA a regularity method that does not distinguish between ontology and epistemology. Seeing causality as an "enabler," as conditions making outcomes possible, means that uncertainty in QCA is possibilistic in nature rather than probabilistic. In fact, the truth-table analysis models possibilistic uncertainty in that logical remainder rows, for which it is uncertain whether the outcome is possible, are turned into easy and difficult counterfactuals depending on their plausibility. The uncertainty of not knowing is of

a fundamentally different nature than the uncertainty that follows from randomness in the data (e.g., unsystematic measurement error) and, accordingly, should be modeled in a different way (Dubois 2006:48). Unsystematic measurement error affects the robustness of empirically observed regularities and, therefore, is a key concern for regularity methods. However, unsystematic measurement error is much less problematic for QCA where regularities are not decisive for causality. That is, for QCA, "analytical robustness" is ultimately more important than "empirical robustness."

Robustness tests applied to QCA focus on parameter sensitivity and measurement error (e.g., Hug 2013; Krogslund et al. 2015; Lucas and Szatrowski 2014; Rohlfing 2018; Skaaning 2011; Thiem, Spöhel, and Duşa 2016) and on model specification (e.g., Baumgartner 2015; Baumgartner and Thiem 2017; Braumoeller 2015; Krogslund et al. 2015; Lucas and Szatrowski 2014; Rohlfing 2016; Seawright 2014; Thiem, Baumgartner, and Bol 2016; Thiem, Spöhel, and Duşa 2016). Sensitivity tests have confirmed that QCA is sensitive to changes in parameters, most importantly consistency thresholds and calibrations, and to measurement error (cases calibrated on the wrong side of the crossover point). But that should surprise no one because, being a case-based method, QCA is case sensitive by design. Different consistency thresholds and calibrations (including wrong calibrations) change the nature of the cases and result in different cross-case regularities (Emmenegger et al. 2014). Substantive interpretation provides an effective, if not flawless, safeguard against calibration errors and "incorrectly" setting consistency thresholds (Maggetti and Levi-Faur 2013; Olsen 2014). The "arbitrariness" in setting QCA parameters (Krogslund et al. 2015:24) is by design, and their plausibility is based on substantive knowledge. We may disagree on parameter values, but they are set transparently and have predictable, systematic effects on the truth-table analysis. Moreover, Skaaning (2011) found that QCA is not very sensitive to small adjustments (p. 404). This is very important because only big adjustments (e.g., setting a substantially higher crossover point) calibrate semantically different sets, which should produce different cross-case regularities (Ragin 2000, 2008). As to unsystematic error (e.g., measurement error), sensitivity tests conflate ontological and epistemological determinism (Beach and Pedersen 2016:19). For example, Rohlfing (2018) is wrong to argue that "a perfect set-relationship [consistency is 1] can be taken as the benchmark for interpreting any empirical consistency value . . . [because] the empirical relationship between X and Y is deterministic on an ontological level, that is, the observed consistency would be unity with complete data measured without any error" (p. 74). However, ontological determinism does not imply perfect consistency in error-free

data. Inconsistent cases do not follow from measurement error (i.e., probabilistic uncertainty that can be remedied with error-free data) but from unknown causes that negate the mechanism implied by a consistent set relationship. That is, the implied mechanism is not activated (Mahoney 2001:580) because of unknown constraints. This is why critical realism ontologically speaks of causality in terms of generative mechanisms that make outcomes possible but do not determine them. Rohlfing (2018:74) thus is an example of the "epistemic fallacy," that is, collapsing statements about empirical regularities (epistemology) into statements about causal mechanisms (ontology; Bhaskar [1975] 2008:13, 16). On case level, outcomes (or their absence) have causes and things don't just happen (Beach and Pedersen 2016:17-18, 48; Mahoney 2008:420). The noisy nature of social reality will always produce inconsistent cases because we cannot know all causes (Ragin 2008:52; Ragin and Sonnett 2008:147-49). Moreover, even if we could know all causes, this would merely lead to infinite regress and result in highly complex configurations that cover single cases only (Hedström and Swedberg 1998:12; Schneider and Wagemann 2012:149). But this defeats QCA's aim to tease out idiosyncratic from generic factors when interpreting cross-case regularities into causal mechanisms (Goertz 2017:30). It also ignores the role of coverage in interpreting causality.

The model specification tests (above references) are problematic because they either take fully specified truth tables as benchmarks or include random or redundant conditions in the truth-table analysis or both. Fully specified truth tables are problematic because some truth-table rows are logically impossible and will always be void of cases, while the noisy nature of social reality means we will never observe all logically possible rows (Schneider 2018; Thomann and Maggetti 2017; Wagemann and Schneider 2010:389). "Ideal" truth tables thus contain limited diversity. By implication, methods and robustness tests must be designed to deal with imperfect data. Next, including redundant or random conditions is problematic because a truth-table analysis is meaningful only with causally interpretable conditions. The truth-table analysis is agnostic about causality; it merely identifies cross-cases regularities, even in random data. After all, there is regularity in randomness. Probability theory distinguishes between genuine and spurious regularities based on the statistical significance of the probability of their occurrence (Taylor 1978). However, the coverages of patterns from random data will be low, meaning that coverage is a QCA safeguard against spurious set relationships, particularly in large-*N* QCA—one completely ignored by model specification tests. Since causality is derived from interpretation and close engagement with the cases (Rohlfing and Schneider 2014:28), the

notion of "wrong model specification" is problematic. Baumgartner's (2015) observation that "if a Boolean expression contains a redundant element it does not reflect a causal structure" (p. 844) is valid from an effects-of-causes perspective. However, when redundant conditions are added to the truth table, all the "wrong" models will always be subsets of the "true" causal structure (Rohlfing 2018:88). From a possibilistic (causes-of-effects) perspective, conjunctions with a redundant condition are unproblematic because they contain the "true" causal structure. That is, conditions present make it possible for agents to achieve the outcome. However, the causal mechanism is unnecessarily limited to a subset of the cases that have membership in the "true" causal structure.

In sum, absent a "context of interpretation," robustness tests hit the target but miss the point. They merely demonstrate how the method works mathematically, but QCA is case sensitive by design. QCA is an iterative process (going back to the cases) that substantively interprets mathematical findings into causal mechanisms. Robustness tests for QCA must be designed around this principle, and empirical robustness comes second to analytical robustness.

A Critical Realist Take on Robustness

Robustness tests must be informed by an explicit understanding of what causality means. QCA's complex causality approach and its explicitly distinguishing between empirically observed cross-case regularities (the truth-table analysis findings) and causal mechanisms (the substantive interpretation of these findings) connect extremely well to critical realism (Byrne 2009:103–05; Gerrits and Verweij 2013:177–79). Both QCA and critical realism argue that:

1. Constant conjunctions between causes and outcomes are emergent, not generative. They are empirical manifestations of underlying causal mechanisms. Constant conjunctions suggest but do not evidence causality (Bhaskar [1975] 2008:184; Collier 1994:128; Rohlfing and Schneider 2018:38). This means that the empirical robustness of constant conjunctions, or lack thereof, while important, is not decisive in establishing the validity of a causal claim.
2. Constant conjunctions must be interpreted into a causal mechanism by answering the question: How and why does the presence of a cause make the occurrence of the outcome possible? (Bhaskar [1975] 2008:51; Collier 1994:22; Ragin 2008:149). This exercise in

substantive interpretation serves to develop epistemological statements (about constant conjunctions) into ontological statements (about the underlying causal mechanisms; Bhaskar [1975] 2008:50; Byrne 2009:104; Collier 1994:76). This means that empirical robustness tests are only relevant in so far as they are analytically meaningful.

3. Causality is an enabler, not a forcing. The presence of causes makes the occurrence of the outcome possible but does not determine that it occurs. Put differently, underlying causal mechanisms are ontologically real also when they do not "produce" the outcome. Consequently, an empirically robust constant conjunction between causes and outcomes is not necessary for causality (Bhaskar [1975] 2008:48-50; Collier 1994:44; Schneider and Wagemann 2012:291-94). This means that empirical robustness tests must identify an empirical range within which the same substantive interpretation is plausible (analytical robustness) rather than identify a threshold below which a causal claim is no longer considered valid.

These considerations inform the meaning of construct, internal, and external validity of a QCA study. *Construct validity* is about capturing the "causal power" of conditions (Bhaskar [1975] 2008:50; Byrne 2009:104; Collier 1994:62). Analytically, this means that conditions must be defined in such a way that it is obvious how and why they contribute to the outcome. It must be clear that their presence allows agents (the case) to achieve the outcome. Thus, construct validity is about calibration; it is about answering the questions: What does a condition (or outcome) mean (ontology) and how can we know that a case has membership in it (epistemology)? Empirically, therefore, construct validity is about alternate calibrations. *Internal validity* is about the causal interpretability of empirical findings. That is, it addresses how statements about empirically observed constant conjunctions (epistemology) are interpreted into causal mechanisms (ontology). Analytically, this pertains to a focus on "causal" rather than "context" conditions (Goertz 2017:59) and on limiting the number of conditions in a QCA study to avoid difficult-to-interpret empirical findings (Schneider and Wagemann 2012:276). Empirically, internal validity pertains to measurement error, as this may cast doubt on the validity of observed constant conjunctions. *External validity* is about the relevance of the identified causal mechanisms beyond the cases under observation. Analytically, this refers to the comparability of cases and the scope conditions of the study. These factors define the nature of the cases under observation and, hence, that the same causal

mechanisms are "at work" within them (Goertz 2017:60, 82). Empirically, external validity pertains to the case sensitivity of the solution and the empirical similarity of cases on confounding conditions. These factors may suggest that the observed outcome was the result of different mechanisms in different cases.

Construct Validity

Construct validity captures the semantical (analytical) meaning of conditions (and outcomes) and how they may be measured. That is, it captures the ontological and epistemological aspects of calibration; that is, the notion that set membership values reflect semantical meanings (Ragin 2000:80). It is important to keep in mind that not all conditions are "causal," as in actually contributing to the outcome (Goertz 2017:59–60; Schneider and Wagemann 2012:253–54). For example, Cooper and Glaesser (2016) use the condition (social) "class" in relation to educational achievement of individuals, but "class" does not actually cause individuals to obtain an educational degree, that is, it is a context rather than a causal condition. Given the notion of causality underlying QCA, "class" puts in place the conditions that may enable individuals to obtain a higher educational degree such as parents being able to afford the best education and providing their children with a safe environment to grow up in. "Going back to the cases," these factors rather than "class" explain obtaining an educational degree. The condition "melting pot" is a relevant example from the study on openness values and regional innovation. Correlational studies connect sociocultural diversity, measured as the proportion of nonnationals in the population, to regional innovation. However, nonnationals do not cause innovation. What is relevant about sociocultural diversity is that it allows knowledge-creating individuals to access a wider and more diverse pool of knowledge and ideas. But sociocultural diversity does not contribute to innovation when different communities live on their own "islands" without much exchange happening between them. That requires tolerance for different cultures, lifestyles, and religions. "Melting pot" may thus be conceptualized as the intersection of the sets of socioculturally diverse and tolerant regions. That is, "melting pot" reflects why knowledge and ideas may be exchanged in a region and thus facilitates causal interpretation of configurations including this condition. The point is, large-*N* QCA researchers must carefully consider how they can combine individual database observations into semantically meaningful sets and conditions (Greckhamer et al. 2013; Emmenegger et al. 2014).

Using database observations, an obvious (but not the only) way to calibrate cases is Ragin's (2008) direct method of calibration. The key question is how to set the critical values in the absence of in-depth case knowledge? Cooper and Glaesser (2016) correctly observe that, lacking such knowledge, there may be no obvious way to set the critical values. Ragin (2008) suggests identifying "convenient gaps" in the data, which begs the question why one gap should be more convenient than another? The solution of Cooper and Glaesser (2016) is to try a number of alternate critical values and to see which one provides the most interesting findings (p. 448). That sounds uncomfortably like stacking the cards in one's favor, nor does it meet the criterion that calibrations should be based on external knowledge. The openness values and regional innovation study investigate 108 cases, but data are available on 249 cases (nearly all European Union regions). Ranking cases such that the highest raw data value region gets rank 249 and the lowest raw data value region gets rank 1 allows making an *X* (raw data)–*Y* (rankings) plot. If that plot produces an *S*-curve, the bends in the *S*-curve identify where the convenient gaps in the data may be found (Figure 2). In this study, "ranking" works because using data from all 249 cases makes the *S*-curve an external argument to the 108 cases under observation. Moreover, substantive knowledge comes into play to identify gaps between regions that are generally recognized as different in kind (crossover point) considered "typical" regions (threshold of full membership) and "unrepresentative" regions (threshold of full nonmembership). Lacking such knowledge may question the legitimacy of doing a comparative case study in the first place. Other convenient gaps may be used as alternate critical values for robustness tests. One should set more "lenient" and more "conservative" critical values for each set and rerun the truth-table analysis. Alternate critical values should be set close to the original ones lest the semantical meaning of the set changes.

Ragin's direct method of calibration produces continuous set membership values. At first sight, continuous or fine-grained set membership values seem attractive and they suggest accuracy. However, they have important limitations that thus far have received no attention in the literature.

– Set membership values are verbal constructs, and they attribute a numerical value to qualitative statements based on substantive knowledge of cases (Ragin 2000:154, 2008:30). Semantically, therefore, fine-grained set membership values can only be substantiated based on in-depth case knowledge that is lacking in large-*N* QCA. This makes fine-grained set membership values for database observations somewhat of a contradiction in terms because there is no substantive

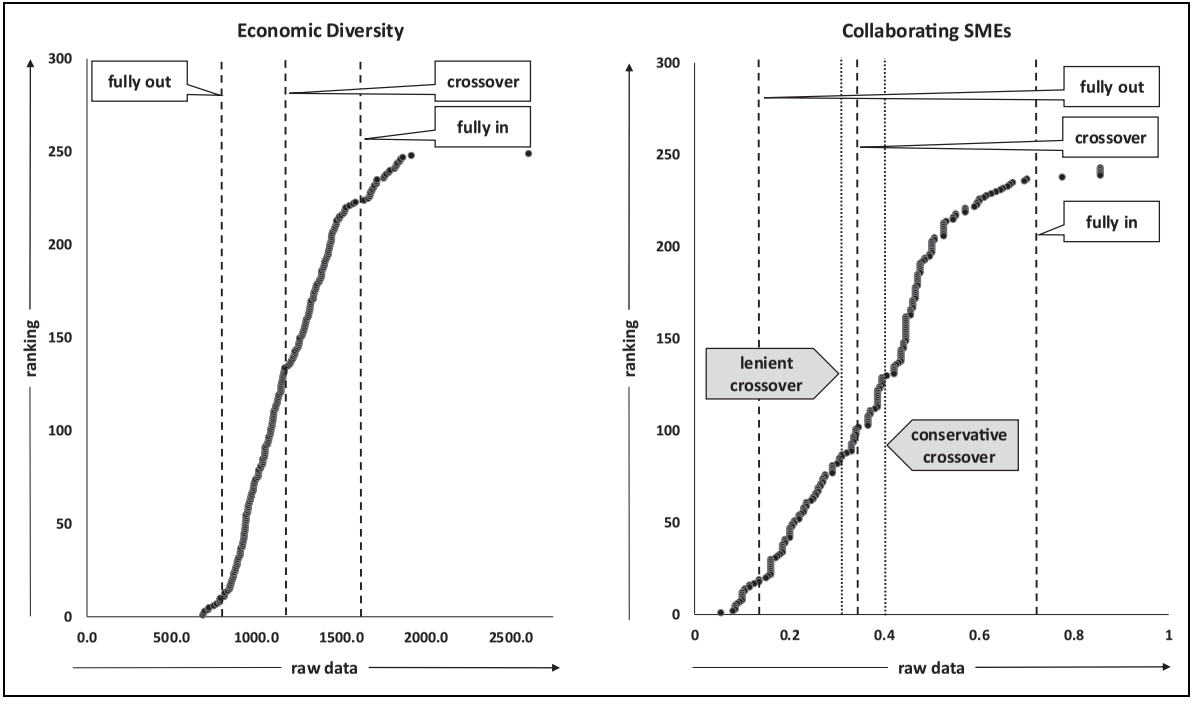

Figure 2. Calibrating by ranking.

knowledge to suggest that small raw data differences between cases are semantically meaningful.

– More importantly, fine-grained set membership values may actually contribute to producing cases inconsistent in degree. The more fine-grained the set membership values, the higher the likelihood that, for a number of cases, X will be higher than Y, but Y is still higher than 0.5. Since cases inconsistent in degree do not violate an if-then statement, QCA researchers rarely, if ever, attribute consequences to them. This raises two problems. First, if cases inconsistent in degree are ignored, fine-grained set membership values become an exercise in spurious specificity. Second, if consequences are attached to cases inconsistent in degree, the question is how many such cases are allowed before a statement of sufficiency ($X \leq Y$ in fsQCA) is violated? It introduces a layer of arbitrariness and ambiguity in fsQCA that defeats one of the very reasons for using fuzzy sets (see Ragin [2008] and Schneider and Wagemann [2012] for the rationale behind fuzzy sets).
– Furthermore, using different set membership levels for different sets (e.g., when using both continuous database observations and Likert-type scale survey data) runs the risk of using the same semantical label for different numerical values. "Mostly in the set" may be numerated as .8 for continuous database observations and as .67 for five-level Likert-type scale data, which causes interpretation problems (Smithson and Verkuilen 2006:27). It may also create more cases inconsistent in degree.

These arguments dovetail to dismiss the use of fine-grained set membership values in large-N QCA. Semantically, there may be little point going beyond a six-value scale: 0, 0.2, 0.4, 0.6, 0.8, and 1.

Running the truth-table analysis in the choice calibration returns a solution of four configurations: (i) diversity, (ii) technology transfer, (iii) cosmopolitan environment, and (iv) creativity configurations. All the four configurations are perfectly interpretable and are corroborated by theoretical knowledge and by substantive knowledge in the form case studies available in the literature. Also, the solution coverage and consistency are very good (Figure 1). That is, from a substantive-interpretation perspective, the choice calibration is completely unproblematic. Running the truth-table analysis also for the continuous, conservative, and lenient calibrations (Figure 1) shows all the three alternate calibrations identifying a fifth configuration also. However, this configuration is uninterpretable because it is unclear how the *absence* of knowledge creation and the *absence* of self-expression could

be interpreted into a mechanism responsible for the *presence* of innovation. Consequently, configuration V is best ignored as an artifact of (noisy) data. Otherwise, the alternate calibrations do what one expects them to do; they produce supersets and subsets of the configurations in the choice solution.

Figure 1 suggests that the alternate calibrations produce quite different truth tables in terms of cutoff frequencies and the number of rows and the percentage of cases included in the truth-table analysis. This is what one expects the method to do: changing the nature of the cases (by changing the calibration) one expects to find different cross-case patterns (Rohlfing and Schneider 2014). Finally, comparing the choice (coarse) calibration and the continuous calibration for cases inconsistent in degree confirms the above suspicion (Table 1). Considering the continuous calibration exclusive configuration V shows that both solutions perform highly similar in terms of cases inconsistent in kind and of the cases they cover. So for this study, fine-grained set membership values have no added analytical value. Moreover, the coarse calibration has no cases inconsistent in degree but the continuous calibration does. Particularly, configuration II is problematic because half of its cases are inconsistent in one way or another, which suggests a validity issue if one believes that sufficiency means $X \leq Y$ in fsQCA. Whether other large-*N* QCA studies have similar issues is a matter of further investigation. The broader points are that calibration must never become a mere mathematical exercise and that, in fsQCA, cases inconsistent in degree also defeat causality.

Internal Validity

Internal validity in QCA is about the causal interpretability of set relationships. Following the substantive-interpretation approach, a consistent set relationship on itself does not make a statement of sufficiency or necessity (Ebbinghaus 2005; Schneider 2018). Conceptually, internal validity in large-*N* QCA is about the conditions being causally connected to the outcome. Internal validity thus has an important theoretical and substantive component. The conceptual aspect of internal validity is also directly related to agency. A causal mechanism implies agency, which means that conditions relevant for explaining the outcome should affect (enable or constrain) agency (Bhaskar [1975] 2008:184; Collier 1994:130; Goertz 2017:154). That is, "causal conditions" enable agents to achieve the outcome. Context conditions may be better dealt with as selection criteria for cases or in a two-step QCA rather than in the principle truth-table analysis. By implication, the tighter the connection between condition (configuration) and agency, the

Table 1. Consistent and Inconsistent Cases.

	Choice Six-value Calibration			Continuous Calibration			Continuous Calibration Exclusive Configuration V		
		Inconsistent			Inconsistent			Inconsistent	
Cases	Consistent	Degree	Kind	Consistent	Degree	Kind	Consistent	Degree	Kind
Configuration I	36	0	9	—	—	—	—	—	—
Configuration II	24	0	8	42	10	12	42	10	12
Configuration III	33	0	6	31	2	6	31	2	6
Configuration IV	34	0	8	—	—	—	—	—	—
Configuration V	—	—	—	21	3	7	—	—	—
Total *Y* cases		89			89			89	
Total covered cases		82			100			79	
Cases inconsistent in kind		13			18			13	
Uncovered cases		20			7			23	
Consistent cases		69			89			66	
Uniquely covered cases		46			66			55	

more plausible the causal mechanism that is interpreted from the condition (configuration; Abbott 1998).

Conceptually, internal validity also pertains to the number of conditions in a large-*N* QCA study. Too many conditions complicate internal validity for two reasons. First, with too many conditions, the QCA algorithm may no longer be able to distinguish between conditions, which compromises the method's difference-making logic (Beach and Pedersen 2016:241; Marx, Cambré, and Rihoux 2013:37; Schneider and Wagemann 2012:152). Second, too many conditions increase limited diversity which, in turn, makes solutions needlessly complex because the algorithm can no longer minimize redundant conditions (Ragin 2008; Schneider and Wagemann 2012). Also in large-*N* QCA, limited diversity may be considerable. In this study, only 20 of the 32 truth-table rows are populated with cases, but 8 of them are too sparsely populated and must be declared logical remainders (Emmenegger et al. 2014; Greckhamer et al. 2013). This leaves a maximum of 12 truth-table rows (37.5 percent) for analysis (Online Annex 1 [which can be found at http://smr.sagepub.com/supplemental/]). Expanding the number of conditions to seven results in only 33 of 128 truth-table rows being populated and just 14 rows with three or more cases (11 percent; Online Annex 2 [which can be found at http://smr.sagepub.com/supplemental/]). Of course, QCA is designed to deal with limited diversity, but that does not take away the fact that large numbers of conditions (almost) always lead to complex and difficult-to-interpret configurations (Schneider and Wagemann 2012; Skaaning 2011; Ragin 2008). The best remedy against these problems is to limit the number of conditions to "a few" causally relevant ones and to use other (context) conditions as selection criteria and scope conditions. Any attempt to quantify "a few" is arbitrary, but Greckhamer et al.'s (2013) suggestion to use 6–12 conditions (p. 54) seems very generous. Interpreting any configuration of six or more conditions will always be a daunting task. Being serious about substantive interpretation implies taking seriously the limits of interpretability.

Empirically, internal validity pertains to measurement error and to consistency and frequency thresholds to deal with it. This aspect is well-developed in the (large-*N*) QCA literature. QCA seems to have converged on a minimum consistency threshold of .8 (Rubinson et al. 2019). This is also the default setting in the fsQCA Version 3.0 software. Given that large-*N* QCA lacks in-depth case knowledge to assess inconsistent cases, it seems reasonable to set higher consistency thresholds, for example, $\geq$.85 (Greckhamer et al. 2013:54). Higher consistency thresholds also reduce the number of inconsistent cases, which is helpful because the absolute number of

inconsistent cases affects consistency by defeating the implied cross-case regularity (Hug 2013:263). Setting different consistency thresholds is not possible in the openness values and regional innovation example because all truth-table rows have a consistency of $\geq .9$ (Online Annex 1 [which can be found at http://smr.sagepub.com/supplemental/]).

The frequency threshold is an important tool for internal validity in large-*N* QCA because it rules out sparsely populated truth-table rows. Those rows may reflect exceptional cases or measurement errors and should not be used to identify and generalize cross-case patterns. No formal rules exist for setting frequency thresholds, and this process has to balance two competing considerations. First, a high-frequency threshold eliminates measurement error and exceptional cases from the truth-table analysis. Second, a low frequency includes as many as possible of the (purposefully) selected cases. Ragin and Fiss (2008:197) and Greckhamer et al. (2013:67) suggest that large-*N* QCA should set frequency thresholds to include at least 80 percent of the cases but to always exclude sparsely populated truth-table rows. Setting frequency thresholds is also an effective way to address the case sensitivity of QCA as it only includes empirically relevant rows in the truth-table analysis (Emmenegger et al. 2014; Maggetti and Levi-Faur 2013).

In the present example, two frequency cutoff values are plausible: four and five. A frequency cutoff of four is technically better because it allows more truth-table rows to enter the minimization process (12 vs. 9 rows). Moreover, a frequency cutoff of four includes 89 percent of all cases in the minimization process compared to 78 percent for a frequency cutoff of five. Running the truth-table analysis for both frequency cutoffs produces set theoretically similar but substantively different findings (Figure 3). Frequency cutoff five identifies the choice solution. Frequency cutoff four eliminates "self-expression" from configuration IV, making this configuration a superset of configuration II and thus eliminating "synthetic knowledge" from it. Set theoretically, this is what one expects the method to do, and it produces an interpretable solution. The problem is that the configuration (analytical knowledge AND economic diversity) has a relatively low proportional reduction in inconsistency (PRI) consistency ($<.8$) and that it identifies a basic configuration that can be further specified by connecting it to (self-expression OR synthetic knowledge). Substantive interpretation suggests that these complex configurations reflect very different underlying causal mechanisms. This means that a technically inferior frequency threshold may yield a better solution in terms of causal interpretability.

	Frequency cut off 5 (choice) 78% of cases, 9 truth-table rows				Frequency cut off 4 (alternate) 89% of cases, 12 truth-table rows			
configuration	I	II	III	IV	I	II	III	IV
Analytical knowledge		●	●	●		●	●	●
Synthetic knowledge		●	●				●	
Economic diversity	●	●		●	●	●		●
Melting pot	●		●		●		●	
Self-expression				●				
Raw consistency	0,944163	0,947369	0,962366	0,958333	0,944163	0,922131	0,962366	0,922131
PRI consistency	0,845070	0,803922	0,887097	0,863636	0,845070	0,793478	0,887097	0,793478
Raw coverage	0,543860	0,526316	0,523392	0,605264	0,543860	0,657895	0,523392	0,657895
Unique coverage	0,061403	0,008772	0,099145	0,043860	0,049707	0,163743	0,099415	0,163743
Solution coverage	0,789474				0,807018			
Solution consistency	0,944056				0,926175			

Figure 3. Solutions for alternate frequency cutoffs.

External Validity

Analytically, external validity in QCA pertains to the relevance of the identified causal mechanisms beyond the cases under investigation and chiefly depends on case selection and scope conditions (Goertz 2017:81). This pertains to both the analytical comparability of cases and the relevance of the identified mechanisms for other contexts (Ebbinghaus 2005; Goertz and Mahoney 2012). Empirically, external validity in large-*N* QCA has two aspects: case sensitivity and the differences and similarities between cases that have membership in the identified conjunctions. An accepted way to test case sensitivity is to rerun the truth-table analysis minus randomly deleted cases. If repeated such reruns produce similar solutions, the choice solution is argued to be robust (Maggetti and Levi-Faur 2013). The problem of this procedure is that it affects limited diversity and the relative positioning of the consistency thresholds and that randomly deleting cases may be a questionable strategy for data that are not randomly generated (Emmenegger et al. 2014:26–27). This means that one should expect this exercise to return subsets and supersets of the choice solution rather than exact replications (Emmenegger et al. 2014; Ragin 2008; Schneider and Wagemann 2012).

Using the website random.org allows to randomly delete 10 cases from the 108 cases under investigation, to rerun the truth-table analysis minus those cases, and to repeat the exercise 10 times. The result is shown in Figure 4. For example, configuration I was exactly replicated in 7 of the 10 reruns, reproduced as a subset in 2 of the 10, while 1 of the 10 reruns failed to identify configuration I. Expecting this method to return subsets and supersets of the choice solution suggests that seven (exact replications) + two (subsets) = nine of the reruns were accurate, meaning that configuration I has an accuracy of 0.9 (9 of 10). The accuracies of configurations II, III, and IV are 0.7, 1, and 0.8, respectively. The accuracy of the solution is 0.85, but that may not be very relevant as the focus is on the individual configurations and their implied causal mechanisms. The interpretation of the numbers is discussed in the next section.

External validity also pertains to similarities and differences between cases having membership in a configuration. First and foremost, cases must be selected so as to be analytically comparable, to be empirical examples of the same analytical concept (Ebbinghaus 2005; Ragin 1992). This allows mechanisms to be generalized to other contexts if those cases are analytically similar to the ones under investigation (Beach and Pedersen 2016:54). Selection criteria may be set theoretically relevant for the outcome but can be made logically redundant by selecting cases so that they differ only in degree

Identified as	Configuration I	Configuration II	Configuration III	Configuration IV	Accuracy
	7 times	2 times	6 times	3 times	18/40 = 0.45
	--	4 times	--	4 times	8/40 = 0.20
	2 times	1 time	4 times	1 time	8/40 = 0.20
	1 time	3 times	--	2 times	6/40 = 0.15
Accuracy	9/10 = 0.90	7/10 = 0.70	10/10 = 1.00	8/10 = 0.80	(9+7+10+8)/40 = 0.85

Configuration:  replicated  identified as superset  identified as subset  not identified

Figure 4. Analysis of accuracy.

Table 2. Confounder Condition.

	Population Density Regions		∼Population Density Regions		
Solution	Number	Proportion	Number	Proportion	Heterogeneity
Configuration I	32	.89	4	.11	(.89 × .11)/.25 = .3916
Configuration II	20	.83	4	.17	(.83 × .27)/.25 = .5644
Configuration III	24	.73	9	.27	(.73 × .17)/.25 = .7884
Configuration IV	27	.79	7	.21	(.79 × .21)/.25 = .6636
Solution	51	.74	18	.26	(.74 × .26)/.25 = .7696

on them. All other characteristics of the cases should be logically redundant if the observed mechanism is to be generalizable. There is very little literature on how similar or different cases should be on those other characteristics to allow generalizing their configuration to other contexts. However, the more different the cases, the better generalizable their configuration because it suggests that all the observed empirical differences between cases are logically redundant (no constant association). The more cases share a particular characteristic, the more likely it is that this characteristic may, in fact, be a causal insufficient but necessary condition of a unnecessary but sufficient conjunction (INUS) condition. To check whether this is actually the case, large-*N* QCA researchers can identify theoretically and/or substantively important "confounding conditions" (Goertz 2017:107). They can calibrate these conditions into crisp sets (confounder and ∼confounder) and then assess whether they are logically redundant. Researchers assess the proportion of "confounder cases" and "∼confounder cases" in each individual configuration. A very high or very low proportion of "confounder cases" suggests that the presence or absence of the confounder condition is set theoretically relevant in the configuration because there is a (near-)constant association with the outcome.

An obvious "confounder" in the openness values and regional innovation study is population density. It indicates agglomeration, which is connected to both economic development and innovation. Calibrating the set of population density regions using the above ranking method shows that 51 of 69 consistent cases in the choice solution (Table 2) are density regions while the remaining 18 are ∼density regions. The proportions of density and ∼density regions thus are 0.74 and 0.26, respectively. This means a substantial minority of cases are ∼density regions and that, on the level of the

solution, density is logically redundant. But it is more relevant to look at individual configurations. The proportion of ∼density regions for configurations I, II, III, and IV is 0.11, 0.17, 0.27, and 0.21, respectively (calculated over the consistent cases only, Table 2). Multiplying the proportion of the confounder and ∼confounder cases gives a maximum of 0.25 (0.50 × 0.50) and a minimum of 0.00 (0.00 × 1.00). Dividing this product by 0.25 (the maximum score) returns a heterogeneity value between 1.00 and 0.00. The heterogeneity of the four configurations are .3916, .5644, .7884, and .6636, respectively. This means that density may be logically redundant only in configuration III and that density is part of configuration I—which is plausible given that configuration I is the diversity mechanism and that diversity "happens' in urban regions. Inspecting the ∼density regions in configurations II and IV learns that these regions are either part of larger urban areas or rural regions with their population concentrated in a central city. This means that configurations II and IV effectively reflect urban mechanisms.

The Robustness Table and Its Interpretation

The results of the various robustness tests may be summarized in a robustness table (Figure 5). The layout of this table and the symbols used are designed to match QCA's solution table and offer an easy-to-eyeball overview of the robustness test findings. Authors may want to include the robustness table in the results section, or in an Online Annex (which can be found at http://smr.sagepub.com/supplemental/). It may be interpreted as follows:

- Lenient and conservative alternate calibrations and thresholds for consistency and frequency present easy and difficult tests for the choice solution. Small adjustments do not calibrate semantically different sets and should produce similar solutions. "Lenient" lowers the bar for cases to have membership in a solution, while "conservative" makes such membership more difficult. Consequently, one would expect the choice solution to pass the easy tests but not necessarily to also pass the difficult tests. Borrowing from process tracing, "lenient" sets "hoop tests" (Beach and Pedersen 2013:102; Goertz and Mahoney 2012:93-94). Passing a hoop test is necessary but not sufficient for the choice solution to be valid; it must pass all hoop tests. Failing any one hoop test significantly reduces confidence in the validity of the choice solution, while passing it does not substantially increase confidence. If the choice solution fails a hoop test, calibrations and thresholds must be reevaluated on the basis of empirical,

	Configuration I	Configuration II	Configuration III	Configuration IV
Alternate calibrations				
Conservative calibration	⊗	◓	◓	⊗
Lenient calibration	◒	●	●	◒
Alternate consistency thresholds				
Conservative threshold	N/A	N/A	N/A	N/A
Lenient threshold	N/A	N/A	N/A	N/A
Alternate frequency thresholds				
Conservative threshold	N/A	N/A	N/A	N/A
Lenient threshold	●	◓	●	◒
Accuracy	0.90	0.70	1.00	0.80
Heterogeneity: urbanization	0,3916	0,5644	0,7884	0,6636

Configuration: ● replicated ◓ identified as superset ◒ identified as subset ⊗ not identified **[N/A]** not applicable

Figure 5. Robustness table.

substantive, and theoretical knowledge until the choice solution passes all hoop tests. Again borrowing from process tracing, this may be referred to as a process of updating (Beach and Pedersen 2013:96) confidence in the validity of the choice solution. Referring to the conservative, difficult tests as "smoking-gun tests" perhaps take the analogy too far. However, passing a difficult test substantially increases confidence in the validity of the choice solution while failing it does not much decrease confidence. No updating is required when the choice solution fails a difficult test.

- The accuracy of the choice solution presents a Bayesian problem. If the truth-table analysis is redone 10 times minus randomly deleted cases, the choice solution could be "hit" or "missed" 10 times. That is, the set of potential "hits" is 10, while the set of actual hits varies between 0 and 10. The Bayesian problem then is What is the probability that the choice solution is true given the different alternate solutions that we found? (Bennett and Checkel 2015:279–85; see Figure 6). Actually calculating a Bayesian probability may be very difficult; because of the problems connected to randomly deleting cases, we are effectively comparing apples and oranges. But the Bayesian logic is clear. Since the overlap between the choice and alternate solutions is about the consistency of the set relationship, we may use QCA's commonly used consistency threshold of .8 for subset relationships to interpret the accuracy value as an alternative to actually calculating probabilities. However, we may exercise some leniency because of the above problems. For example, the accuracy of configuration II is 0.7, but since it passes all hoop tests and also a difficult test, we may accept 0.7 as accurate.
- Heterogeneity directly affects the set theoretic logic underlying QCA (Beach and Pedersen 2016:242; Goertz 2017:226). The question is what proportion of confounder/$\sim$confounder cases still suggests the absence of a constant association between confounder and outcome? Setting this proportion is arbitrary, but, given the importance of causal homogeneity, a high threshold is recommended. A proportion of 20/80 would result in a heterogeneity of [$(0.20 \times 0.80)/0.25 =$] 0.64, which seems low. A proportion of 25/75 or 30/70 results in a heterogeneity of 0.75 and 0.84, respectively, and that is perhaps where the heterogeneity threshold should be. (Approximately 0.8 conforms to other customary thresholds in QCA and thus facilitates interpretation.)

The robustness table thus shows that the set relationships in the openness values and regional innovation study are robust and that,

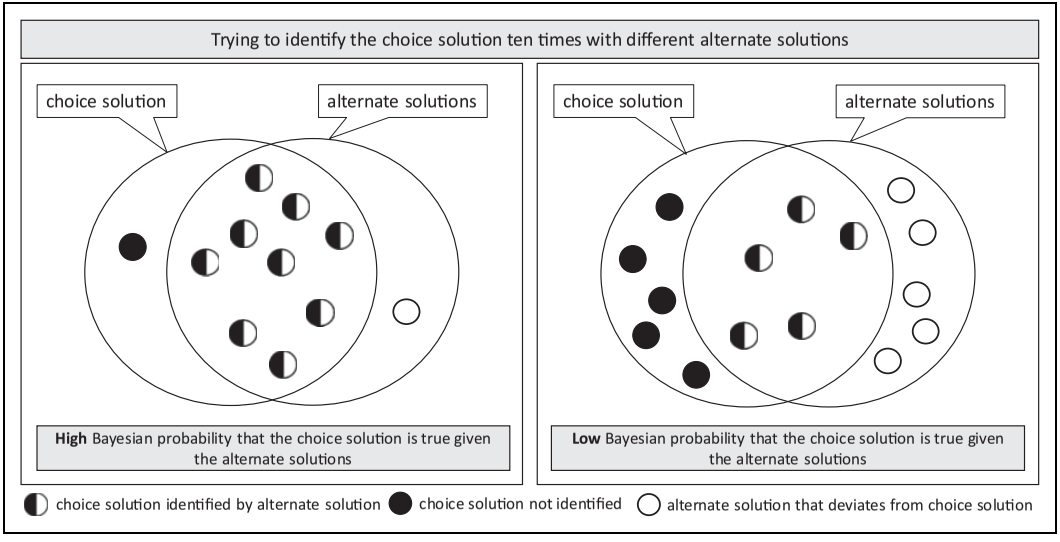

Figure 6. Overlapping choice and alternate solutions as a Bayesian problem.

consequently, the causal claims they support are (empirically) valid. As argued, conducting robustness tests is not a one-time exercise. Instead, large-*N* QCA researchers must use robustness tests findings to "update" confidence in their choice solution by reevaluating calibrations and thresholds. This does not mean adhering slavishly to the above robustness thresholds. There must always be a room for substantive interpretation around robustness test thresholds. Most importantly, the process of updating disciplines large-*N* QCA researchers to go back to their data, mirroring the going back to the cases of small-*N* QCA. It prevents large-*N* QCA from regressing into a decontextualized exercise in truth-table minimization.

Discussion and Conclusion

This article connects QCA's substantive interpretation of empirically observed cross-case regularities (i.e., configurations) into causal mechanisms to critical realism. Neither QCA nor critical realism considers empirical regularities as evidence of causality, as do regularity methods, but as empirical manifestations of underlying causal mechanisms. The key implication is that empirical robustness, while important, is not decisive for the validity of causal claims in QCA. Instead, when identifying causal mechanisms, QCA studies must answer the question: How and why does the presence of the cause make it possible for the outcome to occur? The plausibility (analytical robustness) of the answer is ultimately more important than the robustness of the empirical regularities that are the starting point for substantive interpretation. The article draws on critical realism to explain what construct, internal, and external validity mean in a QCA context and suggests empirical robustness tests for them. An important innovation is that this article visualizes all robustness test findings in a robustness table and suggests ways to interpret them. The article suggests the robustness table as a tool and invites researchers to "update" confidence in the validity of their causal claims in case robustness tests leave room for doubt. The robustness table as a tool thus serves two purposes. First, it answers calls in the literature for a systematic evaluation of the robustness of (large-*N*) QCA studies. Second, it forces large-*N* QCA researchers to go back and forth between empirical findings and contextual and theoretical knowledge in a way similar to the going back to the cases of small-*N* QCA. This is to prevent large-*N* QCA from degenerating into a decontextualized exercise in truth-table minimization.

Using critical realism to confirm QCA as a substantive-interpretation method has two further implications. First, QCA should not be applied to very large databases where cases are no longer analytically relevant but decomposed into conditions in much the same way that regression analysis decomposes cases into variables. Without good contextual knowledge and without cases to go back to, large-*N* QCA will have to infer causality from empirical regularities, defeating its interpretivist nature. Second, applying QCA to simulated truth tables in order to test the robustness of the truth-table analysis is meaningless. Being a case-based method, QCA is designed to be sensitive to dropping or adding cases and to changing consistency and frequency thresholds. Robustness tests on simulated truth tables merely reproduce how the method works (Emmenegger et al. 2014; Olsen 2014). Whether the different empirical findings are meaningful depends on context, but these studies have no context or cases to go back to, which makes them rely exclusively on regularities to validate causal claims. Ultimately, what defines (large-*N*) QCA is that cases are analytically relevant. Thus, setting the boundaries of "the comparative method" helps the discussion on the robustness of (large-*N*) QCA because comparative analysts no longer need to engage with robustness issues of exercises in decontextualized truth-table analysis. There is no actual distinction between QCA as an approach and QCA as a method (Ragin 2008:173; Wagemann and Schneider 2010:378), the approach is the method. If, substantiated by critical realism, QCA is indeed about detecting causes of effects (outcomes) by triangulating empirical, theoretical, and substantive knowledge, comparative analysts have very little to say on "stand-alone" truth-table analysis.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


Roel Rutten https://orcid.org/0000-0002-5933-2996

Supplemental Material

Supplemental material for this article is available online.

References

Abbott, A. 1998. "The Causal Devolution." *Sociological Methods & Research* 27: 148-81.

Álvarez-Coque, J., F. Mas-Verdú, and R. Roig-Tierno. 2017. "Technological Innovation versus Non-technological Innovation: Different Conditions in Different Regional Contexts?" *Quality and Quantity* 51:1955-67.

Baumgartner, M. 2015. "Parsimony and Causality." *Quality & Quantity* 49: 839-56.

Baumgartner, M. and A. Thiem. 2017. "Often Trusted but Never (Properly) Tested: Evaluating Qualitative Comparative Analysis." *Sociological Methods & Research* 46:345-57.

Beach, D. and R. Pedersen. 2013. *Process-tracing Methods: Foundations and Guidelines*. Ann Arbor: University of Michigan Press.

Beach, D. and R. Pedersen. 2016. *Causal Case Study Methods: Foundations and Guidelines for Comparing, Matching and Tracing*. Ann Arbor: University of Michigan Press.

Bennett, A. and J. Checkel. 2015. "Process Tracing: From Philosophical Roots to Best Practices." Pp. 41-73 in *Process Tracing: From Metaphor to Analytical Tool*, edited by A. Bennett and J. Checkel. Cambridge, MA: Cambridge University Press.

Bhaskar, R.[1975] 2008. *A Realist Theory of Science*. London, England: Verso.

Braumoeller, B. 2015. "Guarding against Falls Positives in Qualitative Comparative Analysis." *Political Analysis* 23:471-87.

Byrne, D. 2009. "Complex Realist and Configurational Approaches to Cases: A Radical Synthesis." Pp. 101-11 in *The Sage Handbook of Case-based Methods*, edited by D. Byrne and Ch. Ragin. London, England: Sage.

Collier, A. 1994. *Critical Realism: An Introduction to Roy Bhaskar's Philosophy*. London, England: Verso.

Cooper, B. and J. 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* 19:445-59.

Dubois, D. 2006. "Possibility Theory and Statistical Reasoning." *Contemporary Statistics & Data Analysis* 51:47-69.

Ebbinghaus, B. 2005. "When Less Is More: Selection Problems in Large-N and Small-N Cross-national Comparisons." *International Sociology* 20:133-52.

Emmenegger, P., D. Schraff, and A. Walter. 2014. "QCA, the Truth Table Analysis and Large-N Survey Data: The Benefits of Calibration and the Importance of Robustness Tests." Compasss Working Paper 2014-79. Retrieved February 01, 2018 (www.compasss.org).

Gerrits, L. and S. Verweij. 2013. "Critical Realism as a Meta-framework for Understanding the Relationship between Complexity and Qualitative Comparative Analysis." *Journal of Critical Realism* 12:166-82.

Goertz, G. 2017. *Multimethod Research, Causal Mechanisms, and Case Studies*. Princeton, NJ: Princeton University Press.

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

Greckhamer, T., V. Misangyi, and P. Fiss. 2013. "The Two QCAs: From a Small-N to a Large-N Set-theoretic Approach." Pp. 49-75 in *Configurational Theory and Methods in Organizational Research*, edited by P. Fiss, B. Cambré, and A. Marx. Binley, England: Emerald.

Hedström, P. and R. Swedberg. 1998. "Social Mechanisms: An Introductory Essay." Pp. 1-31 in *Social Mechanisms: An Analytical Approach*, edited by P. Hedström and R. Swedberg. Cambridge, MA: Cambridge University Press.

Hedström, P. and P. Ylikoski. 2010. "Causal Mechanisms in the Social Sciences." *Annual Review of Sociology* 36:49-67.

Hug, S. 2013. "Qualitative Comparative Analysis: How Inductive Use and Measurement Error Lead to Problematic Inference." *Political Analysis* 21:252-65.

Krogslund, C., D. C. Donghyun, and M. Poertner. 2015. "Fuzzy Sets on Shaky Grounds: Parameter Sensitivity and Confirmation Bias in QCA." *Political Analysis* 23:21-41.

Lucas, S. and A. Szatrowski. 2014. "Qualitative Comparative Analysis in Critical Perspective." *Sociological Methodology* 44:1-79.

Maggetti, M. and D. Levi-Faur. 2013. "Dealing with Errors in QCA." *Political Research Quarterly* 66:198-204.

Mahoney, J. 2001. "Beyond Correlational Analysis: Recent Innovations in Theory and Method." *Sociological Forum* 16:575-93.

Mahoney, J. 2008. "Toward a Unified Theory of Causality." *Comparative Political Studies* 41:412-36.

Marx, A., B. Cambré, and B. Rihoux. 2013. "Crisp-set Qualitative Comparative Analysis in Organizational Studies." Pp. 23-48 in *Configurational Theory and Methods in Organizational Research*, edited by P. Fiss, B. Cambré, and A. Marx. Binley, England: Emerald.

Olsen, W. 2014. "Comment: The Usefulness of QCA under Realist Assumptions." *Sociological Methodology* 44:101-07.

Ragin, C. C. 1992. "'Casing' and the Process of Social Inquiry." Pp. 217-26 in *What Is a Case? Exploring the Foundations of Social Inquiry*, edited by Ch. Ragin and H. Becker. Cambridge, MA: Cambridge University Press.

Ragin, C. C. 2000. *Fuzzy-set Social Science*. Chicago: University of Chicago Press.

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

Ragin, C. C. and S. Davey. 2017. "fs/QCA [Computer Programme]." *Version 3.0*. Irvine: University of California.

Ragin, C. C. and P. Fiss. 2008. "Net Effects versus Configurations: An Empirical Demonstration." Pp. 190-212 in *Redesigning Social Inquiry: Fuzzy Sets and Beyond*, edited by C. C. Ragin. Chicago: University of Chicago Press.

Ragin, C. C. and J. Sonnett. 2008. "Limited Diversity and Counterfactual Cases." Pp. 147-59 in *Redesigning Social Inquiry: Fuzzy Sets and Beyond*, edited by C. C. Ragin. Chicago: University of Chicago Press.

Rohlfing, I. 2016. "Why Simulations Are Appropriate for Evaluating Qualitative Comparative Analysis." *Quality & Quantity* 50:2073-84.

Rohlfing, I. 2018. "Power and False Negatives in Qualitative Comparative Analysis: Foundation, Structure and Estimation for Empirical Studies." *Political Analysis* 26:72-89.

Rohlfing, I. and C. Schneider. 2014. "Clarifying Misunderstandings, Moving Forward: Towards Standards and Tools for Set-theoretic Methods." *Qualitative & Multi-Method Research* 12:27-34.

Rohlfing, I. and C. Schneider. 2018. "A Unifying Framework for Causal Analysis in Set-theoretic Multimethod Research." *Sociological Methods & Research* 47: 37-63.

Rubinson, C., L. Gerrits, R. Rutten, and Th. Greckhamer. 2019. *Avoiding Common Errors in QCA: A Short Guide for New Practitioners*. Retrieved July 02, 2019 (Compasss.org), pp. 1-6.

Rutten, R. 2019. "Openness Values and Regional Innovation: A Set-analysis." *Journal of Economic Geography* 19:1211-32.

Schneider, C. 2018. "Realists and Idealists in QCA." *Political Analysis* 26:246-54.

Schneider, C. and C. Wagemann. 2012. *Set-theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. Cambridge, MA: Cambridge University Press.

Seawright, J. 2014. "Comment: Limited Diversity and the Unreliability of QCA." *Sociological Methodology* 44:118-21.

Skaaning, S. 2011. "Assessing the Robustness of Crisp-set and Fuzzy-set QCA Results." *Sociological Methods & Research* 40:391-408.

Smithson, M. and J. Verkuilen. 2006. *Fuzzy Set Theory: Applications in the Social Sciences*. Thousand Oaks: Sage.

Taylor, S. 1978. "The Regularity of Randomness." *The Mathematical Gazette* 62:1-8.

Thiem, A., M. Baumgartner, and D. Bol. 2016. "Still Lost in Translation! A Correction of Three Misunderstandings between Configurational Comparativists and Regressional Analysts." *Comparative Political Studies* 49:742-74.

Thiem, A., R. Spöhel, and A. Duşa. 2016. "Enhancing Sensitivity Diagnostics for Qualitative Comparative Analysis: A Combinatory Approach." *Political Analysis* 24:104-20.

Thomann, E. and M. Maggetti. 2017. "Designing Research with Qualitative Comparative Analysis (QCA): Approaches, Challenges, and Tools." *Sociological Methods & Research*:1-31.

Wagemann, C. and C. Schneider. 2010. "Qualitative Comparative Analysis (QCA) and Fuzzy Sets: Agenda for a Research Approach and Data Analysis Technique." *Comparative Sociology* 9:376-96.

Author Biography

Roel Rutten is an assistant professor in Tilburg School of Social and Behavioral Sciences at Tilburg University. He is also a senior consultant at European Regional Affairs Consultants (ERAC) and a visiting professor at Northumbria University in Newcastle, United Kingdom. His research interests include the geography of knowledge creation, qualitative comparative analysis (QCA), and the organization of knowledge creation.