Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

Advance Access publication November 5, 2014

*Political Analysis* (2015) 23:21–41
doi:10.1093/pan/mpu016

# Fuzzy Sets on Shaky Ground: Parameter Sensitivity and Confirmation Bias in fsQCA

**Chris Krogslund**
*Travers Department of Political Science, UC Berkeley*
*email: ckrogslund@berkeley.edu (corresponding author)*

**Donghyun Danny Choi**
*Travers Department of Political Science, UC Berkeley*
*email: dhchoi@berkeley.edu*

**Mathias Poertner**
*Travers Department of Political Science, UC Berkeley*
*email: mathias.poertner@berkeley.edu*

Edited by Jonathan Katz

Scholars have increasingly turned to fuzzy set Qualitative Comparative Analysis (fsQCA) to conduct small- and medium-*N* studies, arguing that it combines the most desired elements of variable-oriented and case-oriented research. This article demonstrates, however, that fsQCA is an extraordinarily sensitive method whose results are worryingly susceptible to minor parametric and model specification changes. We make two specific claims. First, the causal conditions identified by fsQCA as being sufficient for an outcome to occur are highly contingent upon the values of several key parameters selected by the user. Second, fsQCA results are subject to marked confirmation bias. Given its tendency toward finding complex connections between variables, the method is highly likely to identify as sufficient for an outcome causal combinations containing even randomly generated variables. To support these arguments, we replicate three articles utilizing fsQCA and conduct sensitivity analyses and Monte Carlo simulations to assess the impact of small changes in parameter values and the method's built-in confirmation bias on the overall conclusions about sufficient conditions.

## 1 Introduction

For as long as social science has been split between scholars employing large-*N*, quantitative methods and those employing small-*N*, qualitative methods, enterprising methodologists have sought to create new analytical techniques that might span this epistemological divide. Despite several highly regarded efforts to synthesize quantitative and qualitative methods in the social sciences (e.g., King, Keohane, and Verba 1994; Brady and Collier 2010; Freedman 2010; Gerring 2011), most work in this area simply acknowledges the contrasting analytical strengths and weaknesses of the two methods while imploring researchers to make active use of both. This so-called "multi-method" approach proscribes no specific technique, suggesting instead the use of any and all methodologies that can be productively applied to a given social scientific inquiry. Few efforts have succeeded, however, in charting a new and synergetic middle ground that bridges the quantitative–qualitative divide.

*Authors' note:* Supplementary materials for this article are available on the *Political Analysis* Web site. (Chris Krogslund, Donghyun Danny Choi, and Mathias Poertner, 2014, "Fuzzy Sets on Shaky Ground: Parameter Sensitivity and Confirmation Bias in fsQCA", http://dx.doi.org/10.7910/DVN/27100 Dataverse [Distributor] V1 [Version]). We thank the editors, the anonymous reviewers, Henry Brady, Ruth Berins Collier, Thad Dunning, Zachary Elkins, Marcus Kurtz, Katherine Michel, Robert Mickey, Gerardo Munck, Jack Paine, Slim Pickens, Roxanna Ramzipoor, Ingo Rohlfing, Jason Seawright, Laura Stoker, Sean Tanner, Alrik Thiem, Alison Varney, and Sherry Zaks for their very helpful comments. A special note of thanks goes out to David Collier, who was instrumental in shaping this project.

© The Author 2014. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

A potential exception is fuzzy set Qualitative Comparative Analysis (fsQCA), a variant of the comparative method originally developed by Charles Ragin and commonly referred to as Qualitative Comparative Analysis (QCA) (Ragin 1987, 2000). fsQCA is advocated by its supporters as a synthetic comparative research strategy that combines the most-desired elements of variable-oriented and case-oriented research into a single method, including the ability to (1) examine a large number of cases; (2) address complex causal conjunctions; (3) produce parsimonious explanations; (4) investigate cases both as wholes and as parts; and (5) evaluate competing explanations.[1] Like QCA, fsQCA utilizes Boolean minimization to identify essential prime implicants, or combinations of explanatory variables deemed sufficient to produce a given outcome. Unlike QCA, however, it departs from the binary version of set membership traditionally assumed in Boolean algebra (crisp sets) to introduce a method of configurational minimization that can handle partial or incomplete set membership (fuzzy sets). Proponents of this method argue that it is a superior approach to social inquiry, as it avoids the troublesome assumptions of most quantitative methods, explains outcome variation in both kind and degree, and allows scholars to draw upon their substantial case knowledge in a precise yet flexible way.[2]

Plaudits notwithstanding, fsQCA suffers from several troubling weaknesses. Adding to recent critiques of this method (Achen 2005; Seawright 2005a, 2005b) and proposed enhancements (Schneider and Wagemann 2012; Glaesser and Cooper 2014; Maggetti and Levi-Faur 2013), this article will demonstrate that fsQCA is an extraordinarily sensitive method whose results are worryingly susceptible to minor parametric and model specification changes. In particular, we build upon the efforts of Skaaning (2011), Hug (2013), and Lucas and Szatrowski (2014), who first raised concerns about the potential instability of QCA results. We depart from their work on several fronts, however—notably in our use of highly systematized, rigorous simulations that can detect result sensitivity to even the smallest of fsQCA parameter changes, in our introduction of tests for result sensitivity to model specification error and inherent confirmation bias, as well as in our overall finding that fsQCA results *are*, in fact, markedly sensitive to very small parameter changes.

We make two specific claims. First, the causal conditions—i.e., the essential prime implicants—identified by fsQCA as being sufficient for an outcome appear excessively sensitive to the values of several key parameters. These include the *minimum frequency threshold*, the *minimum sufficiency inclusion score*, and the *maximum sufficiency inclusion score* used during the Boolean minimization, as well as the raw data anchors for full set membership, full set nonmembership, and the crossover point. Small changes in these parameters produce results that are strikingly different and often contradictory.

Second, while it has been argued that fsQCA provides greater flexibility with regard to model specification than conventional methods (such as regression analysis), this flexibility comes at the cost of marked confirmation bias. Given its tendency toward finding complex connections between variables, the method is highly likely to identify causal combinations containing even randomly generated variables as sufficient for an outcome.

To substantiate these arguments, we replicate three carefully executed articles on important topics in political science that utilize fsQCA as one of their research methods—two published in prominent journals, the other in a regional policy journal. Specifically, we consider Mathias Koenig-Archibugi's "Explaining Government Preferences for Institutional Change in EU Foreign and Security Policy" (*International Organization*, 2004), Steven Samford's "Averting 'Disruption and Reversal': Reassessing the Logic of Rapid Trade Reform in Latin America" (*Politics & Society*, 2010), and Sang-Hoon Ahn and Sophia Lee, "Explaining Korean Welfare State Development with New Empirical Data and Methods" (*Asian Social Work and Policy Review*, 2012). For each paper, we conduct sensitivity analyses and Monte Carlo simulations to assess the impact of small changes in parameter values and model specifications on the essential prime implicants identified by fsQCA. Our findings suggest that fsQCA produces tenuous results.

[1]See Ragin (1987, 121).
[2]See Ragin (2000, 5–14).

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

The remainder of the article proceeds as follows. Section 2 briefly reviews the basic protocol for utilizing fsQCA. Section 3 explores several aspects of fsQCA that are potentially problematic for the stability of findings, while Sections 4 and 5 present the methodology and findings of our analysis, respectively. Section 6 concludes by suggesting that, contrary to current practice, users of fsQCA could strengthen their findings by first presenting their results graphically for all possible parameter values used during Boolean minimization, and only afterward marshal case knowledge to justify specific parameter choices. Furthermore, users would do well to include sensitivity analyses of their results with regard to both the calibration of fuzzy set membership and the specification of causal models.

## 2 An fsQCA Primer

### 2.1 *Measurement*

fsQCA is a variant of Qualitative Comparative Analysis (QCA) originally developed by Charles Ragin as a tool for the inference of necessary and sufficient conditions with small to medium sample sizes. In its original set-theoretic formulation, QCA applies a binary classification to establish membership in one or more sets of interest to the researcher (often referred to as "crisp sets"). Each observation in a given universe of cases is similarly classified in terms of its membership in these sets, after which a logical reduction is performed so as to identify combinations of set memberships deemed sufficient to produce a particular outcome.

By contrast, fsQCA goes beyond binary set membership classifications in favor of more graduated classifications drawn from a range of values. These so-called "fuzzy sets" attempt to capture set memberships that are neither fully complete nor fully incomplete, and that vary in degree as well as in kind. Crisp set membership scores are drawn from the set $C = \{0, 1\}$, while fuzzy set membership scores are drawn from $F = \{f | f \in [0, 1]\}$. Membership scores along the interval (0.5,1] are said to be relatively more "in" than "out" of a given set, while the opposite is true for scores contained in [0,0.5). Scores equal to 0.5 are thought to be neither more "in" nor "out" of the set.

Moving from crisp to fuzzy set membership scores has implications for the general analytical process outlined by QCA. In the first place, significant resources must be dedicated to properly coding the degree of set membership for a universe of cases and sets. Users of fsQCA have several strategies for assigning fuzzy set membership scores. All of these strategies require that the researcher first identifies cases that, for a given set, represent three qualitative anchor points—namely, the full set membership and full set non-membership points, as well as a crossover point, where a case is considered equally "in" and "out" of the set. Once the cases occupying these anchor points have been identified, the researcher can either continue to use case knowledge to directly assign the remaining fuzzy set membership scores, or use one of many predefined algorithms drawing on numerical data and the previously specified anchor points to assign scores.

### 2.2 *Causal Assessment*

After the fuzzy set scores are assigned, a crucial challenge arises in the second phase of the analysis. Here, the researcher carries out the logical reduction used to identify combinations of set memberships that are sufficient to produce a given outcome.

The challenge in this second phase is indeed substantial because, notwithstanding the gradations captured with fuzzy sets, the analytical procedures for causal assessment call for entering the findings into a truth table. This requires reducing the findings once again to dichotomies. The fine-grained choices in fuzzy set scoring can be enormously consequential for the cut-points used in returning to dichotomies.

The fsQCA user is thus required to specify three parameters that help establish what—even in the fuzzy set version—is ultimately a dichotomous separation between cases considered relatively more "in" an outcome set or a particular causal configuration and those that are "out." Only once these parameters are specified can one perform a logical reduction that identifies sufficient causal conditions.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

Users must first specify a *minimum frequency threshold*. Precisely defined, this is an integer greater than or equal to 1 indicating the number of cases that must have a membership score of at least 0.5 in a given set or combination of sets for that set or combination of sets to be included in the subsequent logical reduction. For membership in a single set, a case's membership score is simply its fuzzy set membership score. For membership in a causal combination (i.e., simultaneous membership in two or more sets), a case's membership score is the minimum of its fuzzy set membership scores across all sets in the given causal combination. The intention of the minimum frequency threshold is thus to identify and avoid causal combinations that are irrelevant, meaning those that are rarely seen to occur in the data.

For those configurations that score above the minimum frequency threshold, the fsQCA procedure then requires researchers to specify an additional pair of "consistency" scores that jointly establish whether a given causal configuration should be coded as a sufficient condition for the outcome of interest. Each causal configuration's consistency score measures the degree to which that configuration is associated with the outcome. Mathematically, the consistency score $S$ for a given causal configuration $c$ is bounded by [0,1] and defined as

$$S(c) = \sum_{i=1}^{n} \frac{\min(X_{i,c}, Y_i)}{X_{i,c}}, \tag{1}$$

where $X_i$ is the membership score for case $i$ of $n$ total cases in the causal configuration and $Y_i$ is the membership score for case $i$ in the outcome of interest.

To establish exactly which causal configurations are considered sufficient for an outcome, the user is asked to specify a *minimum sufficiency inclusion* threshold that provides a lower bound for the consistency scores of sufficient causal configurations, as well as a *maximum sufficiency inclusion* threshold that provides an upper bound for the consistency scores of causal conditions that are deemed insufficient for an outcome.

## 3 Threats to Validity in fsQCA

As with all methods, fsQCA requires some potentially consequential assumptions. While fsQCA "frees social scientists from many of the restrictive, homogenizing assumptions of conventional variable-oriented research," it largely trades one set of assumptions for another.[3] Setting the values for the basic parameters—which are used to calibrate fuzzy set membership scores and perform Boolean minimization—should ultimately be understood as taking place "by assumption." Given the high sensitivity of findings to minor changes in these parameters, the strong dependence on the assumptions becomes clear. A related yet distinct concern is with the method's built-in confirmation bias.

### 3.1 *Threat #1: Parameter Specification for Calibration and Reduction*

Perhaps the most troubling assumptions required by fsQCA concern the specification of key parameters for calibration and reduction. As noted by Skaaning (2011, 394), the process of selecting fsQCA parameter values, "despite attempts of theoretical and/or empirical justification, introduce[s] some degree of arbitrariness." Recall from above that users must draw upon their "extensive base of relevant substantive knowledge . . . [to] specify appropriate qualitative anchors defining full membership, full non-membership, and the crossover point."[4] Whereas crisp set QCA assumes that cases are either completely within a given set or entirely excluded from it, fsQCA admits a continuum of set membership scores along the interval [0,1].

As should be evident from the discussion above, classifying cases in this manner is quite understandably half science and half art. Researchers must utilize their substantive, inevitably somewhat intuitive, knowledge of the cases to identify what are often fundamentally hard-to-define cut-points. For calibration of fuzzy set membership scores via raw data anchor points, the fsQCA user must

[3]See Ragin (2000, 120).
[4]See Ragin (2000, 166).

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

identify the *exact* values at which cases jump from one meaningful set membership status to another. For direct assignation of fuzzy set membership scores, before identifying the key qualitative anchor points, the fsQCA user must also be able to first *order* the cases in terms of set membership. In any scenario, fsQCA demands somewhat ironically that what are ultimately converted to crisp values be assigned to points that can be very fuzzy. Herein lies a major limitation on the robustness of fsQCA results. Due to its highly subjective procedures for assigning discrete values to integral parameters used in calibrating fuzzy sets and carrying out logical reductions, we have to ask whether fsQCA results have a strong *a priori* expectation of invalidity due to methodological artifacts.[5]

Consider, for instance, the canonical fsQCA example of assigning fuzzy set membership scores for the set of rich countries. Figure 1 shows the distribution of GDP per capita for 199 countries in 2005. Where exactly a country moves from being considered poor to rich (corresponding to the crossover point) is very much open to interpretation when established on the basis of per capita GDP alone. Consider two candidate points in the raw data for the crossover point: the mean and median of the distribution. In 2005, St. Kitts & Nevis, Hungary, Barbados, the Seychelles, the Slovakia, Antigua & Barbuda, Oman, Trinidad & Tobago, the Czech Republic, and Saudi Arabia were all within 10% of the mean GDP per capita. Likewise, Ecuador, Algeria, Belarus, Tunisia, the Maldives, Serbia, Colombia, Namibia, Suriname, the Dominican Republic, Fiji, and Montenegro were all within 10% of the median GDP per capita. Despite the fact that the data help order the cases in terms of "richness," most analysts would be hard-pressed to specify a single point at which a country moves from being relatively poor to relatively rich.

The calibration procedures at the heart of fsQCA assume, however, that the professional expertise of its users can (and will) meaningfully draw the line separating relatively rich countries from relatively poor countries. Though they are free to utilize any and all sources of knowledge to assign cases to one of an infinite number of fuzzy set membership scores, there can be nothing fuzzy about the scores themselves. What makes this demand placed upon the researcher so troubling is that, as one tries to distinguish between a decreasing number of candidates for a given raw data anchor (such as the crossover point), given data that imperfectly approximate the concept of interest, the likelihood of correctly ordering the cases in just one attempt decreases markedly.

Figure 2 shows just such a result using simulated data and sample sizes that are traditionally regarded as small- or medium-$N$ in social science. For each simulation, a random sample of $n$ integers bounded by [−100,100] was taken. Each of these values $x_i$ was then transformed according to the equation

$$y_i = x_i + \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{2}$$

with $\sigma^2$ values of 1, 3, 5, 10, and 20. The percentage of the total $n(n-1)$ possible, nonidentical comparisons of the form $\{x_i, x_{-i}\}\forall i$ such that the ordering of $\{y_i, y_{-i}\}$ was equal to the ordering of $\{x_i, x_{-i}\}$ was then calculated, with each consistent ordering being weighted by the inverse square root of the absolute distance between $x_i$ and $x_{-i}$. The purpose of the weight is to count consistent orderings that are "hard" (where the $x$-values are relatively close to one another) more highly than consistent orderings that are "easy" (where the $x$-values are relatively far apart). This process was repeated fifty times for sample sizes ($n$) ranging from 2 to 50. The results show that, while lower standard deviations for the error term generally increase the weighted percentage of correctly identified orderings of the $x$-values on the basis of the $y$-values, the consistency of these percentages begins to decline rapidly once the sample size goes below 20–25.

The practical consequence of this finding is that researchers will have a diminished ability to correctly identify those cases occupying meaningful transition points in the concept of interest—such as which country truly lies at the border between rich and poor. This is troubling, because even slight changes in the assigned fuzzy set membership scores or the raw data anchor points used in the calibration process can significantly alter the results produced via fsQCA.

[5]We consider here only validity concerns derived from parameter specification, though similar concerns have been raised with respect to the fuzzy set membership score functions used for indirect score assignment. See Thiem (2014).

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

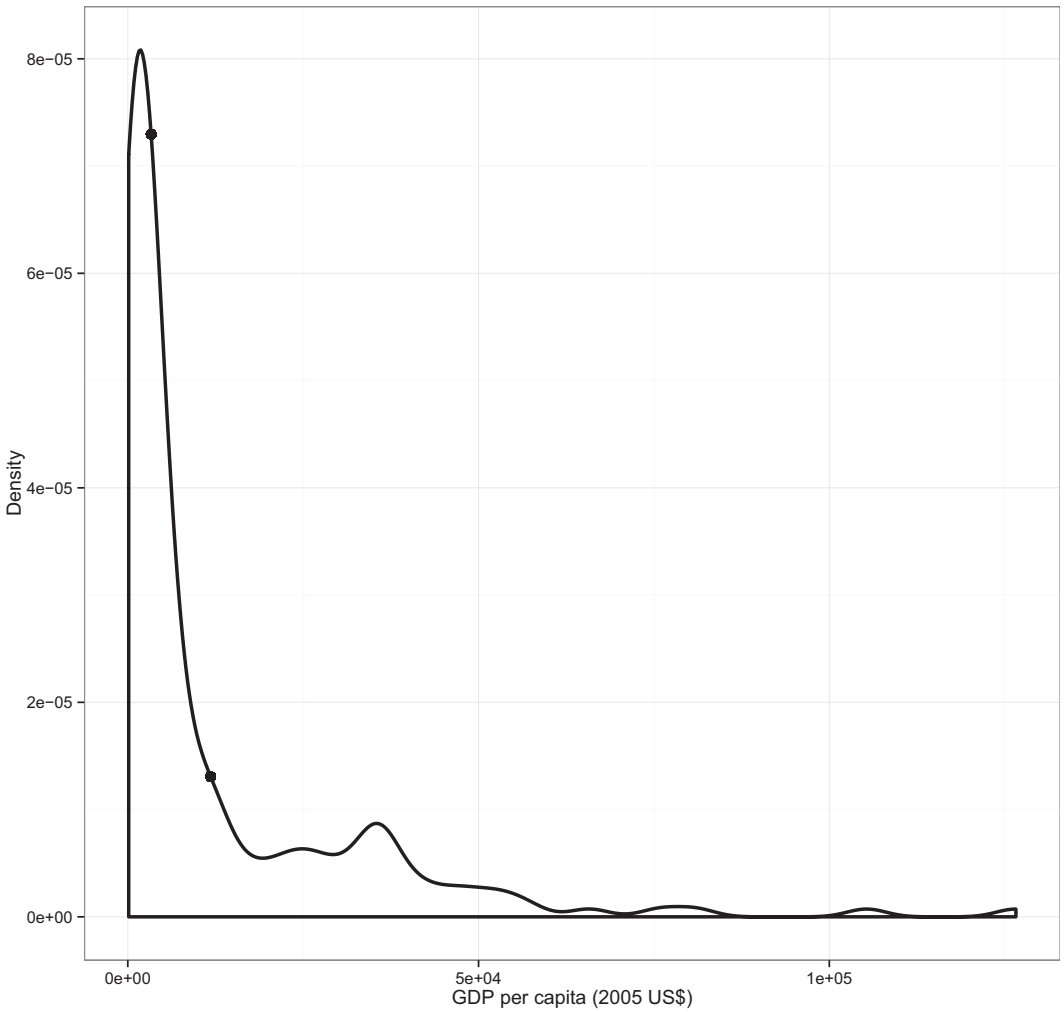


**Fig. 1** Density of GDP per capita (2005 US$), 199 countries. Left-most dot represents the median value, right-most dot the mean value. *Source:* World Bank, *World Development Indicators*, accessed July 23, 2013.

Against this backdrop, the researcher reports many fsQCA results using a variety of raw data anchor points or directly assigned fuzzy set membership scores, which would help establish whether a given result is anything more than an artifact of the method.

In addition to parameter values used to calibrate fuzzy set membership scores, fsQCA requires the user to make parametric assumptions concerning the degree to which certain causal configurations should account for the outcome of interest. These involve, as noted, the minimum frequency threshold, the minimum sufficiency inclusion score, and the maximum inclusion sufficiency score. As explained in the previous section, these parameters establish how strongly a causal configuration must be associated with the outcome for it to be considered a sufficient condition. While fsQCA users are given some general guidelines for selecting these parameter values (higher is better, mainly), the choice of specific parameters is clearly underdetermined by prior theory or case knowledge. Unfortunately, quite small differences in the choice of parameter values can have an outsized impact on one's results.

Suppose that with a given data set and hypothesized causal model, fsQCA identified two causal configurations—call them *A* and *B*—as sufficient for an outcome when the minimum frequency threshold was set to 3 and both the minimum and maximum sufficiency inclusion scores were set to 0.75. If one were to change any of these parameters slightly (say, moving the minimum frequency threshold down to 2, or the sufficiency inclusion scores up to just 0.76), the fsQCA procedure might now identify only configuration *B* as sufficient for the outcome. Equally, it might find that three

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

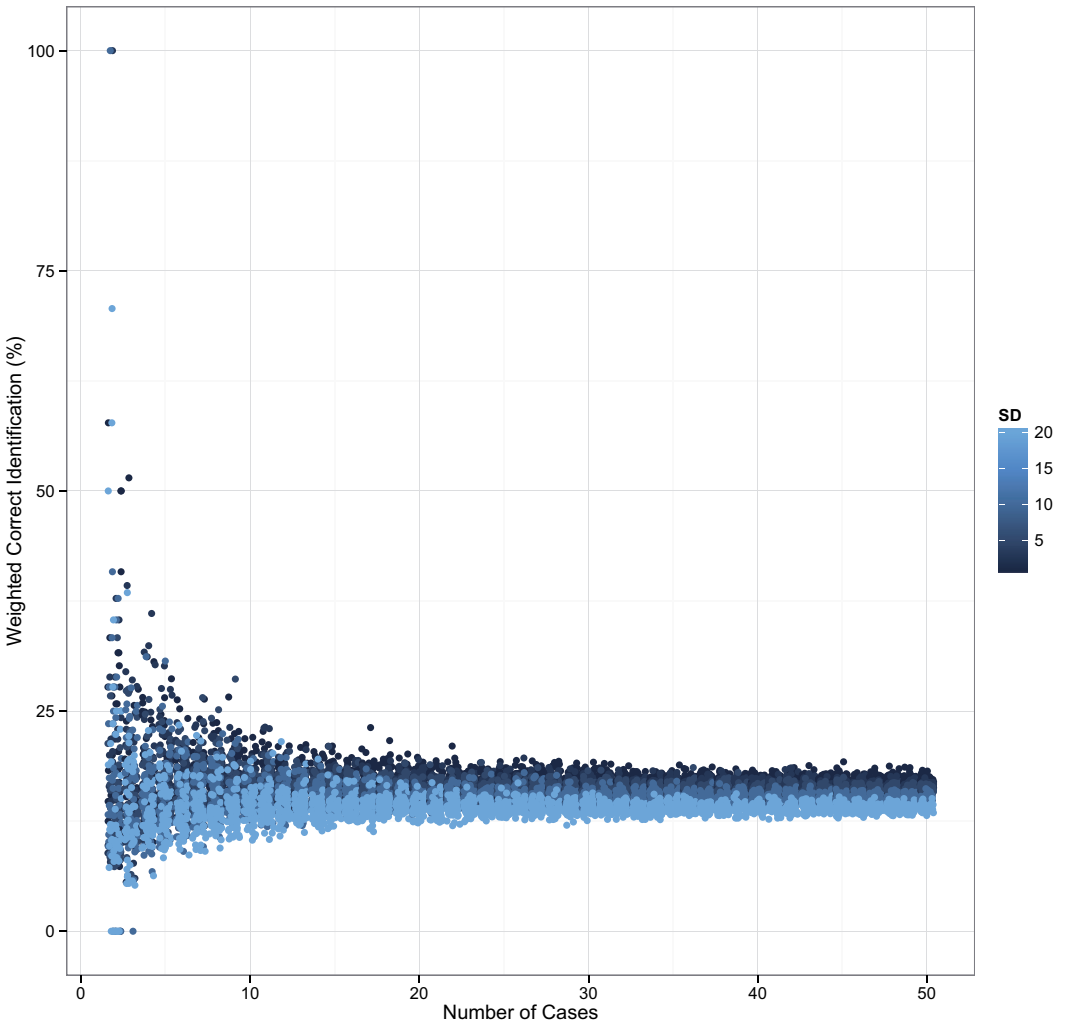


**Fig. 2** Simulated consistency for correct order classification with small- and medium-N sample sizes, weighted by the inverse square root of absolute case distances. Darker dots represent simulated results using normally distributed disturbances with higher standard deviations.

different conditions *C*, *D*, and *E* were now sufficient conditions. In fact, extremely small parameter changes of this sort could identify any number of the $2^k$ possible causal configurations (where $k$ is the number of independent variables included in the hypothesized causal model) as sufficient for an outcome. As with the selection of parameter values for calibrating fuzzy set membership scores, the assumption of specific parameters (from a literally *infinite* choice set for use during the Boolean minimization) presents an extraordinarily large threat to result validity from methodological artifacts.

### 3.2 *Threat #2: Model Specification and Confirmation Bias*

Aside from its many parametric assumptions, fsQCA—like many conventional methods—also demands the assumption of correct model specification. The inclusion or omission of different explanatory factors that are unrelated to the outcome can produce entirely different combinations of necessary and sufficient conditions. Of course, model specification errors are not so much a fault of the method as a fault of the user. But the robustness of fsQCA is potentially limited by the fact that it may be more likely than not to accept the sufficiency of causal configurations that include variables knowingly unrelated to the outcome. In simulations conducted on crisp set QCA,

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

previous research has hinted at QCA's inability to exclude random data from the essential prime implicants (Lucas and Szatrowski 2014). That is to say, fsQCA may have built-in confirmation bias, i.e., a proclivity to commit Type II errors.

Because fsQCA is a method intended primarily for small-*N* and medium-*N* applications, and given its focus on complex causal relationships between combinations of the independent variables and the outcome of interest, chance alone dictates that even variables consisting of random draws from a given distribution will often appear as showing some sufficiency relationship with the outcome. In its most common application, therefore, the method may have relatively little ability to discern between factors that are systematically related to the outcome (either alone or as part of a bigger causal configuration) and factors that are only randomly related to the outcome. By contrast, correlational methods such as regression will, by construction, discriminate against factors that have a weak correlation with the outcome. Any explanatory factor analyzed via fsQCA therefore has a relatively inflated likelihood of being classified as part of a sufficient condition for an outcome.

### 3.3 *Empirical Evidence for Theoretical Threats*

As noted above, a few previous studies have attempted to determine whether the theoretical sensitivity of QCA results translates into practical sensitivity. The verdict on this count is still unclear, with various authors claiming anywhere from minimal to overwhelming sensitivity.

For a variety of reasons, however, much more needs to be done. As with the voluminous use of sensitivity analysis to evaluate more conventional quantitative methods, this process of evaluation requires successive attempts, as scholars make what are often major strides forward in arriving at better tests.

Important gaps in previous simulations are quite evident. For example, Hug (2013) finds csQCA results to be highly sensitive to measurement error in the data. Yet, sensitivity to measurement error in csQCA may be different from the focus of the present article—fsQCA—which is now becoming far more prevalent vis-à-vis the crisp-set version. Great sensitivity to measurement error is certainly a serious deficiency, but other problems that are the focus here, which are even more integral to the basic QCA algorithms, raise even stronger questions about the method. Lucas and Szatrowski (2014) suggest that the introduction of variables known to be unrelated to the outcome routinely yields false positives. Yet again, however, they focus only on csQCA.

By contrast, Skaaning (2011) takes direct aim at the calibration and reduction parameters that are specific to fsQCA. He replicates several empirical works and finds that the theoretical sensitivity of fsQCA results does not lead to their practical sensitivity. But in simulating only a handful of parameter values that are relatively far apart, he fails to capture the full sensitivity of fsQCA results. Indeed, by not examining extremely fine-grained parameter changes across the whole spectrum of possible values, Skaaning effectively fails to test whether fsQCA results are sensitive to *very* small parameter changes. Sensitivity to the very small parameter changes we induce in our analysis would compromise the methodological viability of fsQCA because it is unlikely that any amount of theoretical or empirical expertise would enable the researcher to precisely distinguish *a priori* between these parameter values.

We seek to move beyond these shortcomings in the assessment of fsQCA result sensitivity. Specifically, in what follows, we present a variety of sensitivity tests that systematically assess the sensitivity of fsQCA results to even very small changes in its calibration and reduction parameters, as well as errors in the specification of potentially causal factors. Moreover, in an effort to ensure that fsQCA users will have easy access to sensitivity tests, we develop a user-friendly software package that automates these sensitivity tests and produces clearly interpretable graphical results. The specifics of our methods are detailed below.

## 4 Methodology

To establish how sensitive fsQCA results are to changes in its essential parameter values and model specification, we undertook a series of sensitivity analyses and Monte Carlo simulations on the

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

results of the three articles noted above. To preview the results, we confirm suspicions that small changes in these parameter values lead fsQCA to infer the sufficiency of different, often contradictory, causal combinations. Similarly, we find that the method is highly susceptible to the specification of incorrect causal models; randomly generated variables have a greater than 50% probability of being implicated as part of a sufficient causal combination by fsQCA.

The Boolean minimization at the heart of QCA is a computationally intensive process, requiring the user to identify the prime implicants from $2^k$ possible configurations, where $k$ is the number of explanatory factors. While minimization of this sort can be performed manually for analyses of up to roughly five explanatory factors (thirty-two configurations), anything much beyond that becomes cumbersome. Moreover, calculating the various statistics of fsQCA—including membership scores in the corner configurations of the vector space and scores for the consistency of each configuration as a sufficient condition for the outcome—makes the process even more time consuming.

To overcome these challenges, Charles Ragin, Sean Davey, and Kriss Drass developed software written for Microsoft Windows OS, entitled *fs/QCA*.[6] This program includes a Boolean minimization algorithm for the identification of prime implicants, automatically calculates fsQCA statistics, and is the workhorse program for fsQCA research in the social sciences. According to some estimates, it commands a market share upward of 80%.[7] Unfortunately, the fact that the program has an exclusively graphical user interface makes it a poor choice for conducting our sensitivity analyses and Monte Carlo simulations. However, Thiem and Duşa (2013) have written *QCA*, a package in the R programming language that includes fsQCA functionality.[8]

We developed three companion functions that conduct sensitivity analyses and Monte Carlo simulations of fsQCA results. As a reminder, our goals are to examine (1) how sensitive fsQCA results are to changes in parameter values used to calibrate fuzzy set membership scores and carry out logical reductions; and (2) how robust fsQCA results are to specifying causal models containing a single randomly generated variable.

We have called the core function utilized in our software *fsQCA.sim*. This function takes two arbitrarily large samples $X_{min}$ and $X_{max}$ of values from a uniform distribution bounded by the interval [0,1] and assigns these values in pairs to the minimum sufficiency inclusion score and the maximum sufficiency inclusion score, such that $X_{min} \geq X_{max}$.[9] For a given minimum frequency threshold, our software then utilizes the *eqmcc* function embedded in *QCA*, which identifies prime implicants on the basis of the enhanced Quine–McCluskey algorithm for Boolean minimization using the specified parameter values. This procedure is repeated for all $\{X_{min}, X_{max}\}$ pairs and across all viable values of the minimum frequency threshold. Thus, for a sufficiently large sample of parameter values, *fsQCA.sim* identifies all prime implicants in the data, across all possible combinations of parameter values for the minimum sufficiency inclusion score, maximum sufficiency inclusion score, and minimum frequency threshold.

A second function we developed, *fsQCA.anchor*, integrates with *fsQCAsim* to assess how sensitive fsQCA results are to changes in the procedures used to calibrate fuzzy set membership scores or, in the case that fuzzy set scores are directly assigned without mathematical mapping, how sensitive the results are to minor changes in the scores themselves. In the case that membership scores are calibrated on the basis of numerical criteria, the user is asked to specify the variance of a distribution with mean equal to 1 from which the overall degree of movement in raw data anchor points is established. For a given replication, one or more of the anchor points is multiplied by a value drawn from that distribution, thereby shifting the value of the raw data anchor point used during calibration. Each new set of anchor points is then calibrated via an algorithm fed into *fsQCA.anchor*. If fuzzy set membership scores were assigned directly, a range of scores is specified to receive adjustment (approximating, for instance, the scores of cases at the full membership, full

[6]See Charles Ragin and Sean Davey, fs/QCA, Version 2.5, Tucson: University of Arizona, 2009.
[7]According to http://www.compasss.org/software.htm (accessed October 13, 2014).
[8]See http://cran.r-project.org/web/packages/QCA/index.html (accessed October 13, 2014).
[9]The ordering limitation here helps avoid computational drag from combinations of parameter values for which fsQCA results are not computable.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

non-membership, and crossover points). Selecting a relatively higher distribution variance will lead to greater movement in the raw data anchor points and (potentially) fuzzy set membership scores, whereas relatively lower variances will lead to less overall movement. For a large enough number of simulated shifts in the raw data anchor points, *fsQCA.anchor* will identify all prime implicants in the data, across all possible combinations of parameter values for the minimum sufficiency inclusion score, maximum sufficiency inclusion score, and minimum frequency threshold for a given distribution of overall anchor point movement.

Third, we created the function *fsQCA.random*, which returns the frequency with which a random variable is included in the set of prime implicants identified by fsQCA. This was needed because, while *fsQCA.sim* and *fsQCA.anchor* allow the researcher to see the changes in his or her results for all possible parameter values required in the calibration and logical reduction processes, they do not give us any sense of how susceptible results are to Type II error. The variable may be drawn from any distribution and calibrated in any way—both require user specification. A random sample of "directly assigned" fuzzy set scores bounded by [0,1] can also be requested. The "calibrated" random variable is then combined with the other explanatory factors and fed into the *fsQCA.sim* function. This process is repeated for an arbitrarily large number of iterations. The function will then report the results of these simulations which, for a large number of repetitions with a sufficiently large sample of parameter values, will establish the probability that the random value will be included as a factor in one of the prime implicants, across all possible combinations of parameter values. Any probability greater than 50% can be taken as an indicator of confirmation bias.

## 5 Results

In this section, we detail the results of our sensitivity analyses and Monte Carlo simulations. Because the papers differ substantially with regard to their content, calibration procedures, the number of causal models tested, and the precision with which their methodologies are documented, we chose to tailor the specifics of our replications and simulations to meet the exigencies of each individual paper. To once again foreshadow our findings, the results reported in these articles are, in fact, *quite* sensitive to minor changes in the calibration and reduction parameters. Furthermore, in all three articles, fsQCA results are much more likely than not to identify a randomly drawn variable as being part of a causal configuration that is sufficient for an outcome.

### 5.1 *Ahn and Lee on Welfare State Development*

One of the deepest literatures in political science concerns the development of welfare states in the industrialized democracies, notably in the postwar period. Scholars working in this area have long sought to understand why some nations are so much more generous in their social expenditures than others. After more than fifty years of continued research, political science has identified a number of factors that lead to relatively greater entitlement provisions, including the progress of industrialization, the strength of left parties and trade unions, as well as certain kinds of political institutions.

Until recently, however, the vast majority of empirical work in this area has been focused on Europe and North America. Ahn and Lee (2012) attempt to bring some much-needed case diversity to the literature by investigating the causes of welfare state expansion in South Korea. The authors collect a host of new data from a variety of sources in order to add the Korean case to the well-known *Comparative Welfare States Data Set* (CWS), originally compiled by Huber et al. (2004). The authors estimate a number of different configurational and correlational models, attempting to test the various hypotheses for welfare state expansion on the South Korean data, first via several Prais–Winsten regressions and then via fsQCA.

#### 5.1.1 Original results and replication

Ahn and Lee use the fsQCA software developed by Ragin and Davey to search the configurations of five different sets of explanatory factors for combinations that are implicated as sufficient

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

conditions for having relatively high-welfare spending as a percentage of GDP. Their explanatory factors include (1) the percentage of the population aged greater than sixty-five (ELDERLY); (2) the unemployment rate (STUNEMR); (3) GDP per capita (CGDP); (4) total direct investment as a percentage of GDP (DINVOC); and (5) the presence of left government (LTPRD). The authors selected these factors for inclusion on the basis of simple bivariate Prais–Winsten regressions of the outcome on the various candidate explanatory factors. To avoid technical issues in the Boolean reduction derived from model overspecification, the authors apply the fsQCA procedure to different combinations of the explanatory factors. Though the precise sufficient conditions this technique yields do change with each different specification, the authors generally find that a relatively high GDP per capita and a relatively large elderly population are sufficient to produce a larger welfare state.

To replicate these findings, we used the Korean data from the CWS data set. Our task was immediately complicated, however, by the fact that the authors were imprecise in their methodological reporting. At each step, we reconstructed the data as faithfully as we could.[10]

Ahn and Lee report some of the parameters and other information needed for calibrating fuzzy set membership scores, but not all. Specifically, they give the raw data anchor values for the full membership score, the full nonmembership score, and the crossover point for each of their five explanatory factors. The authors state that their consistency cutoff is 0.85. Though it is not stated explicitly, we assume that the authors use this value for both the minimum sufficiency inclusion score and the maximum sufficiency inclusion score. The authors also note that their calibration procedure was carried out so that equal numbers of cases (years, in this instance) have fuzzy set membership scores above and below 0.5. What the authors do not provide is the frequency threshold.

We calibrated the data according to the authors' specifications. In particular, we used the *calibrate* function in *QCA* with the reported anchor points to assign fuzzy set membership scores to the cases, with half being above 0.5 and half below.

### 5.1.2 Sensitivity analysis for the frequency threshold and sufficiency inclusion scores

Since the absence of a stated frequency threshold prevents us from conducting a straight replication, we proceeded to feed the data directly into our *fsQCA.simulate* function. For each model specification tested by Ahn and Lee, we randomly sampled 3000 pairs of minimum sufficiency inclusion and maximum sufficiency inclusion scores. Only sensitivity analyses for minimum frequency thresholds ranging from one through six or seven were consistently estimable.

Figure 3 shows the sensitivity results for Ahn and Lee's second model; specifically,

$$\text{WELGDP} \Leftarrow \text{ELDERLY} + \text{STUNEMR} + \text{CGDP}. \tag{3}$$

Similar figures for their other models can be found in the online supplementary materials.[11] The figure for Model 2 is divided up into seven plots, each representing a different value of the minimum frequency threshold (noted at the top of each plot). For a given point within each plot, the horizontal coordinate represents the maximum sufficiency inclusion score used for a particular Boolean reduction, while the vertical coordinate represents the minimum sufficiency inclusion score.

[10] For instance, though they use social expenditures as a percentage of GDP for their dependent variable, they do not report which of the two such measures contained in the CWS they use. This is a problem, as the two measures have similar time spans but unequal starting points. Further, while the scores referred to by the authors for direct investment do match a variable in the CWS data set, this variable does not correspond to the description used by the authors. Where the authors use the variable to indicate direct investment as a percentage of GDP, this variable measures direct investment outflows only. Nor do the raw data anchor points they selected for this variable even appear within the variables range in the CWS data set. Moreover, initial replications of the simple bivariate regressions run using the measure of direct investment outflows yielded results that differed greatly from those of the authors. For this reason, we constructed a measure of foreign direct investment that is equal to the difference between foreign direct investment inflows and outflows, as reported in the CWS data set. This operationalization seemed to match the authors' raw data anchor points and regression findings much better than the simple measure of outflows.

[11] All replication materials and supplementary figures are available online. See Krogslund, Choi, and Poertner (2014).

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

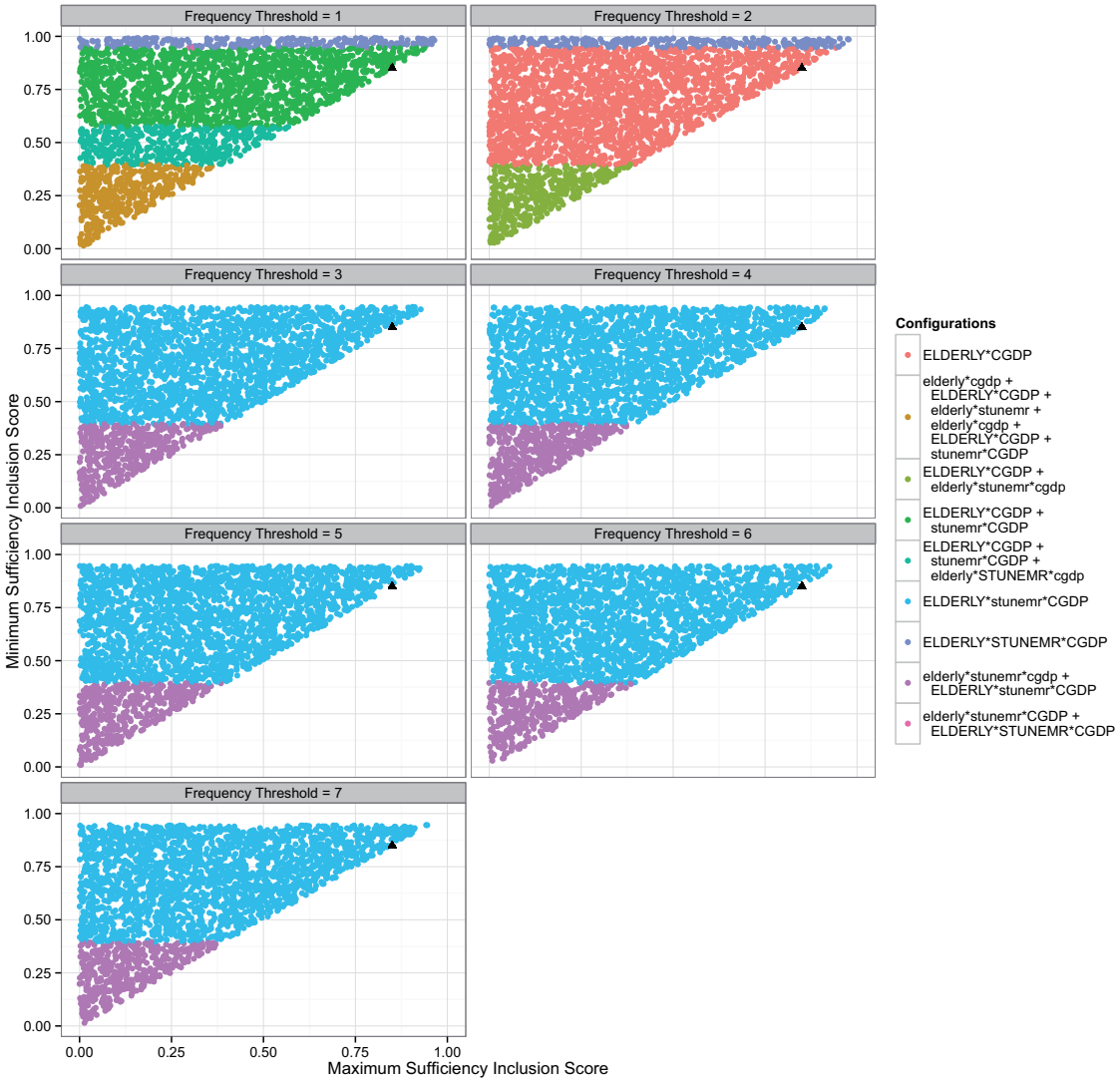


**Fig. 3** Sensitivity analysis for the frequency threshold and sufficiency inclusion scores, Ahn and Lee (2012), Model 2. Note that the triangular point represents the inclusion scores specified by the authors. A more interpretable color version of this figure is available online.

The color coding of each point corresponds to the set of prime implicants that are yielded by the fsQCA procedure for that combination of maximum sufficiency inclusion score, minimum sufficiency inclusion score, and minimum frequency threshold. These configurations are shown in the legend.

Like the authors, we find that the combination of ELDERLY and CGDP form a sufficient condition for greater welfare spending, though our results also require the absence of STUNEMR (the black triangle in the figure represents the result yield for the authors' parameter specifications). Note, however, how greatly the findings can change with just minor parameter changes. Moving the minimum frequency threshold but keeping the inclusion scores at those used by Ahn and Lee yields three different sufficient causal configurations: the presence of ELDERLY and CGDP; the absence of STUNEMR and the presence of CGDP; as well as the presence of ELDERLY and CGDP and the absence of STUNEMR.

Holding the minimum frequency threshold constant but varying the inclusion scores can also produce a number of different results. Consider the minimum frequency thresholds between three and seven. At a minimum sufficiency inclusion score of around 0.40, fsQCA results change from

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

yielding as sufficient conditions the absence of ELDERLY, STUNEMR, and CGDP, or the presence of ELDERLY and CGDP with the absence of STUNEMR, to just the former condition. For a less stable result, consider the minimum frequency threshold equal to 2. As one increases the inclusion scores from the authors' values, at around 0.95 the sufficient conditions switch from the presence of ELDERLY and CGDP to the presence of ELDERLY, CGDP, *and* STUNEMR. For a minimum frequency threshold of 1, the same inclusion score movement ends up with the same causal configuration above 0.95, but starts off with the presence of ELDERLY and CGDP or the presence of CGDP and absence of STUNEMR.

In total, across all possible parameter specifications, fsQCA takes a model with three explanatory factors and yields nine different causal configurations with ten unique sufficient conditions. While our results successfully replicate the findings of Ahn and Lee, they also show that they can be radically altered depending upon several parameter specifications.

### 5.1.3 Sensitivity analysis for the crossover point in calibration

To assess sensitivity of the results to the selection of raw data anchor points for the calibration of fuzzy set membership scores, we repeated the analysis just reported for a hundred randomly drawn inclusion score pairs and all computable minimum frequency thresholds. For each of these hundred pairs, we randomly moved the raw data anchor for the crossover point a hundred times by a factor drawn from a uniform distribution bounded by $[1 - d, 1 + d]$, for a total of 10,000 simulations. We repeated this process for $d$ with values of 0.01, 0.025, and 0.05. In other words, we simulated the effect of introducing up to 1.0, 2.5, and 5.0% identification error in the crossover point.

These simulations are shown in Fig. 4. Again, we focus on the model shown in equation (3), and results from the same procedure applied to the other models are available in the online supplementary materials. Unlike in the previous figures, Fig. 4 shows fsQCA results for all possible values of the inclusion scores, but only for when the minimum frequency threshold is equal to 2. This corresponds to the upper-right-hand plot in Fig. 3. In this case, each plot contains results for when a different level of identification error was applied to the raw data anchor for the crossover point. The amount of the error is indicated at the top of each plot.

The top left-hand plot is a simple replication of the original result found in the top-right-hand corner of Fig. 3, showing three different causal configurations produced by fsQCA. As the level of identification error in the crossover point increases, the original fsQCA results become corrupted. If one switches the raw data anchor for crossover point used in the calibration of fuzzy set scores by just 2.5% in either direction, the number of potential causal configurations produced by fsQCA more than doubles. Parameter combinations that once pointed only to the presence of ELDERLY and CGDP as a sufficient condition for welfare state expansion now also point to the presence of CGDP and the absence of STUNEMR. Likewise, parameter combinations that once yielded the presence of ELDERLY and CGDP—or the absence of ELDERLY, STUNEMR, and CGDP—as sufficient conditions for welfare state growth now find that only the absence of ELDERLY and STUNEMR is sufficient. And whereas the boundaries between the different causal conditions in terms of the inclusion score values were once quite well defined, just a little bit of identification error makes the borders rather fuzzy. Our results, then, point to the conspicuous sensitivity of fsQCA results to changes in calibration parameters.

### 5.1.4 Monte Carlo simulations for random variables

Finally, to gauge the extent to which fsQCA results are subject to Type II errors, we began once again by selecting a hundred randomly drawn inclusion score pairs. For each of these hundred pairs, we drew a hundred random variables and included them in the authors' explanatory model. For instance, the model represented by equation (3) above would now be represented as

$$\text{WELGDP} \Leftarrow \text{ELDERLY} + \text{STUNEMR} + \text{CGDP} + \text{RANDOM}. \quad (4)$$

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

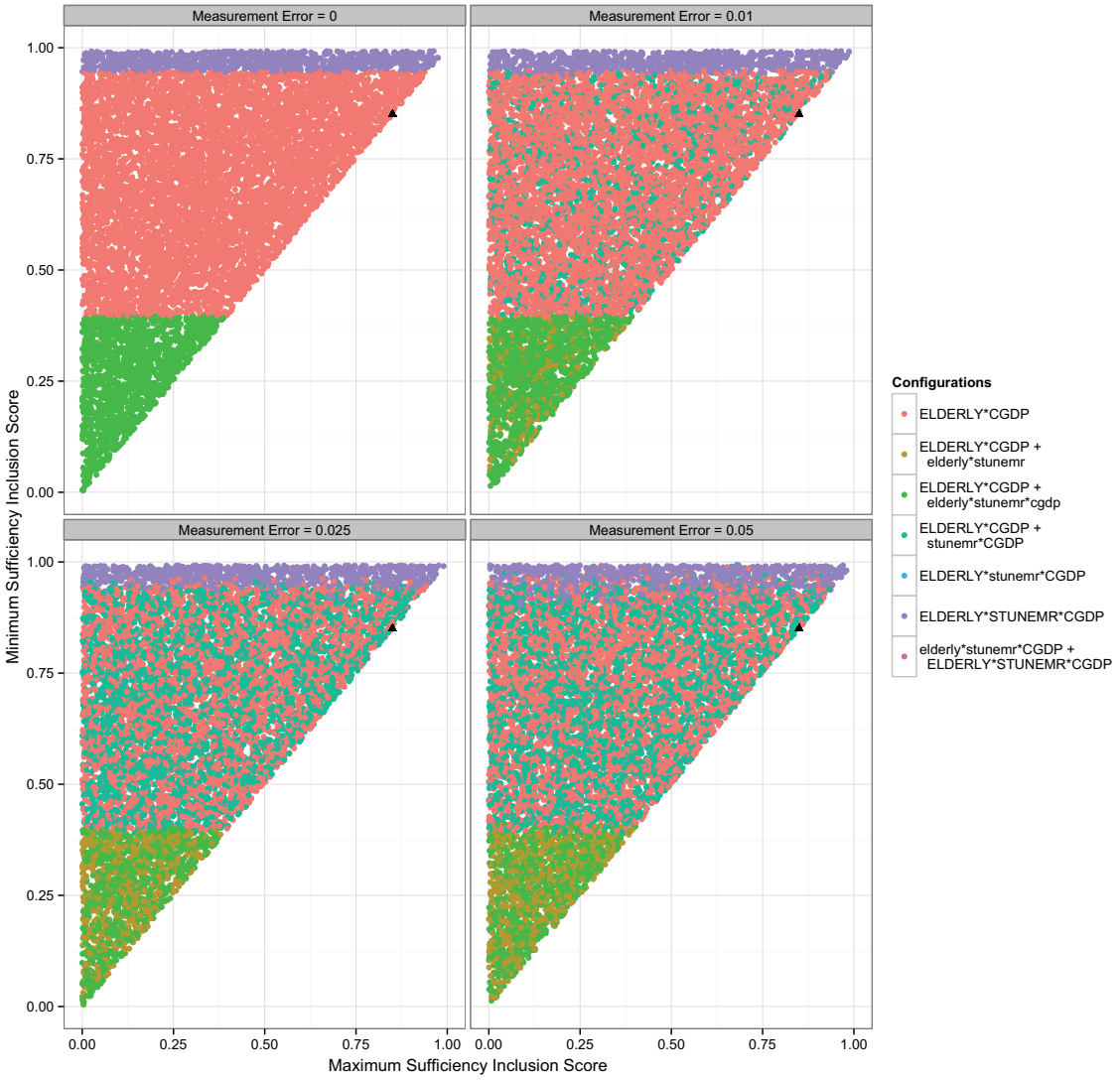


**Fig. 4** Sensitivity analysis for the crossover point in calibration, Ahn and Lee (2012), Model 2. Note that the triangular point represents the inclusion scores specified by the authors. A more interpretable color version of this figure is available online.

To avoid conceptual problems in trying to meaningfully calibrate fuzzy set scores for randomly generated data, we elected to draw "directly assigned" random fuzzy set scores from the interval [0,1].

The results of this Monte Carlo simulation are quite troubling. Across a thousand fsQCA simulations of Ahn and Lee's second model, 75.2% returned causal configurations containing at least one sufficient condition that included the randomly drawn variable. That is to say, three times out of four, fsQCA results find that a random variable is part of a sufficient condition for welfare state expansion. This suggests severe confirmation bias on the part of fsQCA, as it is more likely than not to identify even random variables as somehow sufficient for an outcome.

### 5.2 *Samford on Trade Liberalization in Latin America*

Samford (2010) aims to explain rapid trade liberalization in Latin America between 1970 and 2000. Drawing on arguments in the existing literature, he tests for the sufficiency of multiple conditions approximating "opportunities" for liberalization and "willingness" to liberalize, using fsQCA in combination with three brief case studies. The analysis considers policy outcomes under national

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

executives that govern for at least one year in eleven Latin American countries. Over the thirty-year period, this yields sixty-one cases.

### 5.2.1 Original results and replication

Fuzzy set membership scores for each of the cases are calibrated via the indirect method outlined in Ragin (2000). Using Ragin's original *fs/QCA* software, Samford tests seven factors as being sufficient alone or in combination for observing extraordinarily quick trade liberalization. Specifically, his causal model includes (1) the presence of an unconstrained executive (EXECUNCO); (2) a strong currency devaluation (DEVALU); (3) a previous history of hyperinflation (HYPERINF); (4) a strong manufacturing sector (MANUFAC); (5) strong economic growth (GROSTRON); (6) negative economic growth (GROWEAK); and (7) policy switching or dissimulation (SWITCHER). Samford's documentation is admirably clear and transparent in explaining the calibration steps taken, the raw data anchor points used, as well as many of the parameters used for carrying out the Boolean minimization.

In terms of results, Samford finds that a handful of conditions are ordinarily sufficient to push rapid trade reform—specifically, the presence of DEVALU and EXECUNCO; the presence of HYPERINF and EXECUNCO; and the presence of MANUFAC and EXECUNCO.

### 5.2.2 Sensitivity analysis for the frequency threshold and sufficiency inclusion scores

As with the work of Ahn and Lee, we are one parameter short of being able to exactly replicate Samford's findings. This leads us to once again proceed directly to the sensitivity analysis for the minimum frequency threshold and inclusion scores. As can be seen in Fig. 5, we replicate Samford's main result with inclusion scores equal to 0.75 and minimum frequency threshold equal to 2. As the figure indicates, however, the exact results produced by fsQCA are quite unstable. Note that the results are variable to such an extent that we were forced to put a floor of 0.35 on the minimum sufficiency inclusion scores considered. Had we not done so, it would have been impossible to fit all of the resulting fifty-two causal configurations on a single page.[12]

To illustrate the sensitivity of fsQCA results to minor parameter changes, consider Samford's main finding in the panel with minimum frequency threshold equal to 2 (noted by the black triangle). Were the author to have moved the inclusion scores from 0.75 to just 0.77, the previous finding would have been supplemented with another factor that was jointly sufficient for rapid liberalization: weak growth. Had the minimum frequency threshold been reduced to 1, fsQCA would have found both conditions as sufficient for the outcome. And had the minimum frequency threshold been set to 1 along with the inclusion scores at 0.77, three additional configurations involving the presence of SWITCHER, GROWEAK, and GROSTRON would have emerged. Once again, minor parametric changes lead to big differences in fsQCA results.

### 5.2.3 Sensitivity analysis for the crossover point in calibration

Figure 6 shows the results from applying the same test of crossover point identification error sensitivity used on the Ahn and Lee data to the Samford data (refer to Section 5.1 for the methodological details). For values of the minimum sufficiency inclusion score above 0.50, the results are robust to error in the crossover point (at least below the 5% level). There is some boundary blurring on the margins, but its overall effect appears minimal. Below the 0.50 inclusion score threshold, however, there is notable corruption in the results—even at just 1% identification error for the crossover point.

[12]The full graph can be found in the online supplementary materials.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

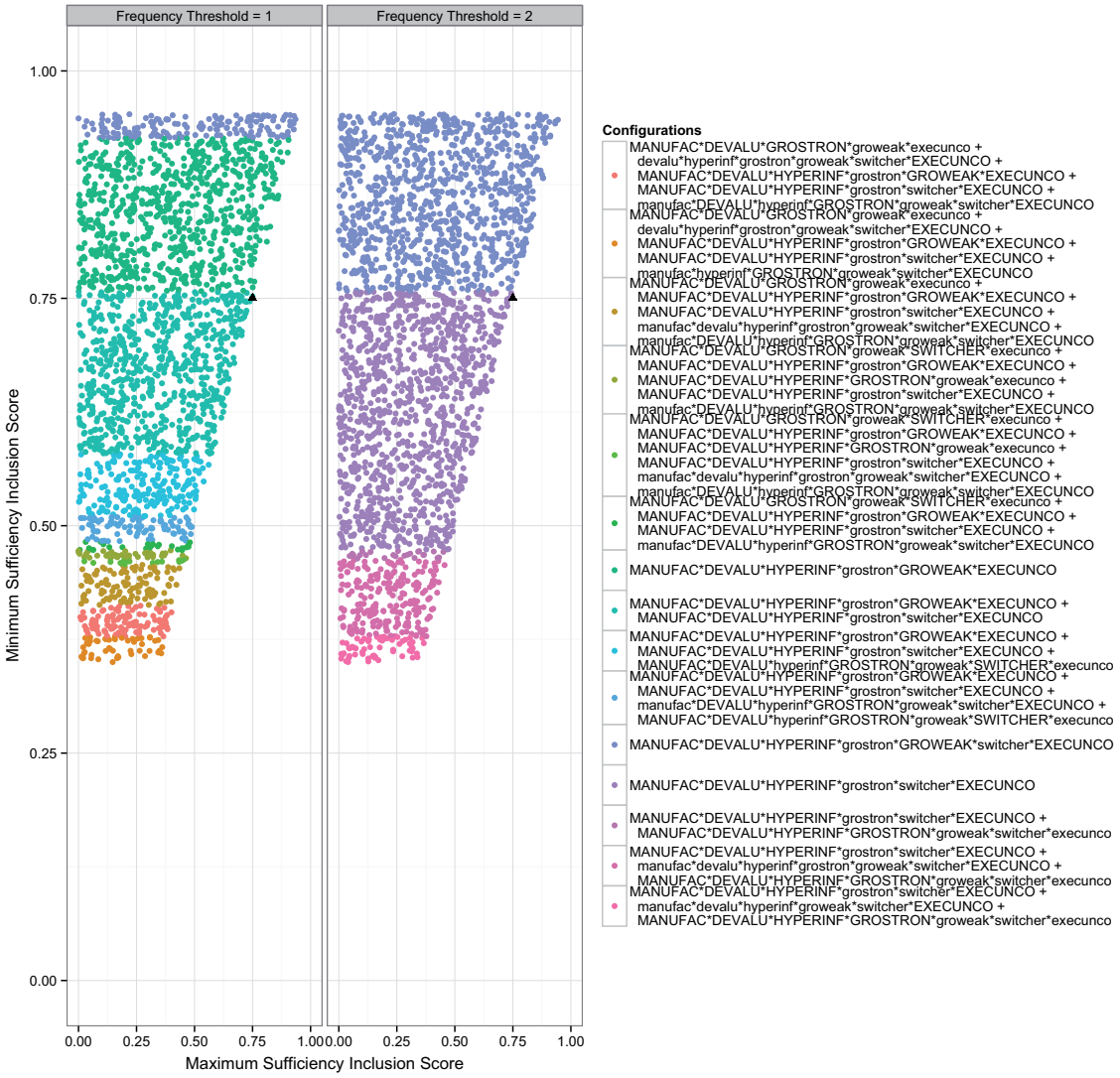


**Fig. 5** Sensitivity analysis for the frequency threshold and sufficiency inclusion scores, Samford (2010). Note that the triangular point represents the inclusion scores specified by the author. A more interpretable color version of this figure is available online.

### 5.2.4 Monte Carlo simulations for random variables

Applying the same Monte Carlo methodology used in Ahn and Lee (Section 5.1), 99% of the fsQCA results included the random variable in some capacity as part of a sufficient condition for an outcome. This is additional evidence of serious confirmation bias.

## 5.3 *Koenig-Archibugi on Government Preferences for EU Foreign and Security Policy*

Koenig-Archibugi (2004) seeks to explain why some member states of the European Union push for common supranational foreign and security policy while others object to any such limitation on their sovereignty. Guided by international relations theory, he examines the impact of relative power capabilities, foreign policy interests, Europeanized identities (measured on the opinion leader level and mass level), and domestic multilevel governance on foreign and security policy cooperation among fifteen EU member states. He does so using both a more conventional regression-based approach as well as fsQCA. The analysis draws data from a variety of sources such as

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

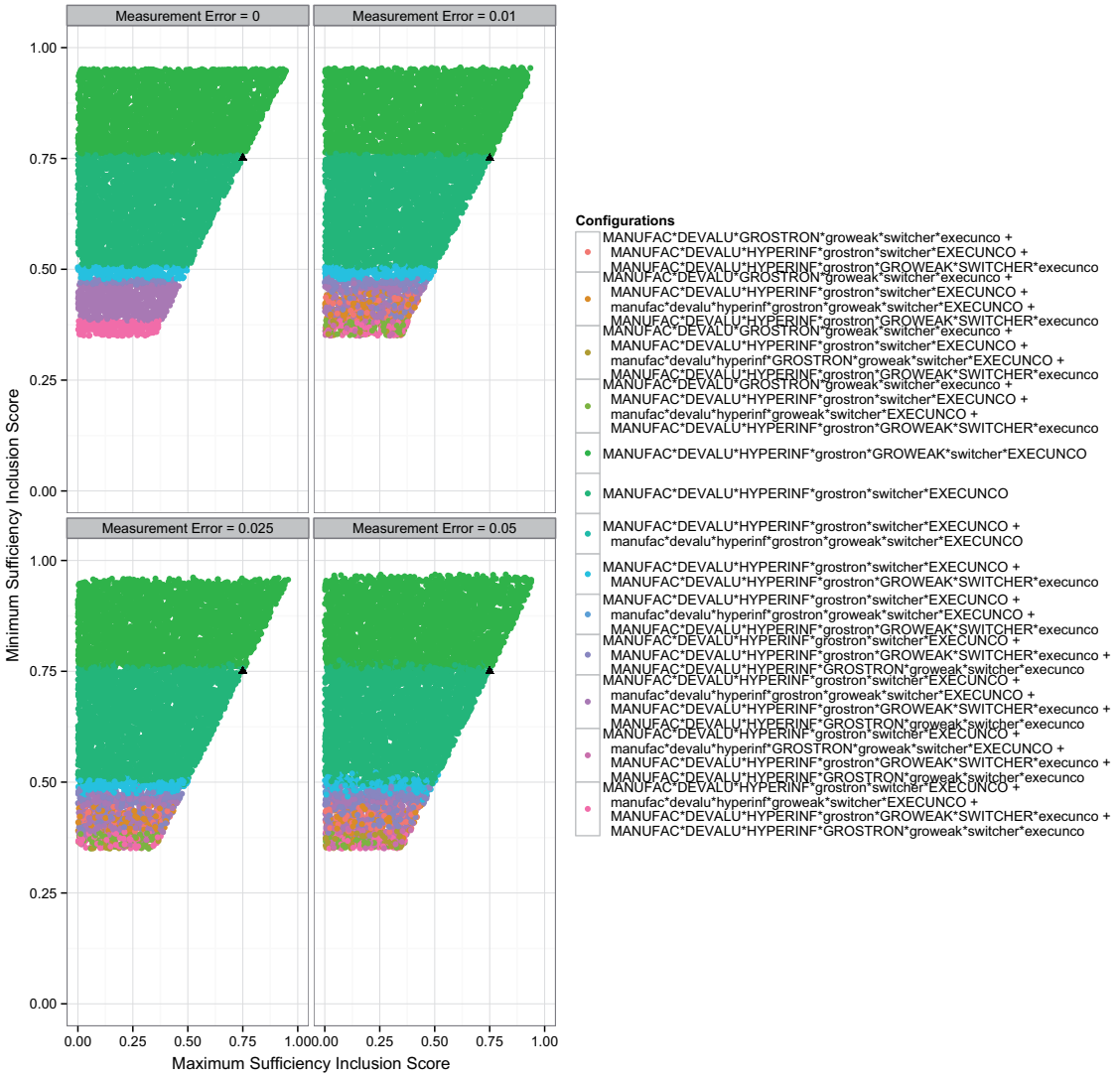


**Fig. 6** Sensitivity analysis for the crossover point in calibration, Samford (2010). Note that the triangular point represents the inclusion scores specified by the author. A more interpretable color version of this figure is available online.

governmental statements on the CFSP, voting behavior in the UN General Assembly, and Eurobarometer surveys to create a data set measuring the outcome of interest and five causal conditions across the EU member states.

### 5.3.1 Original results and replication

Using the *fs/QCA* software, Koenig-Archibugi tests for sufficient combinations of causal factors from among (1) the degree to which a state's policy preferences conform with the preferences of other states (CONF); (2) the strength of regional governments (REG); (3) the level of European identity in the general public (PUBID); (4) the level of European identity in opinion leaders (OPID); and (5) the level of "material capabilities" (MAT). He analyzes two causal models—one included the European identity measure for the general public, and the other included the opinion leaders measure. In general, the author finds that sufficient causal configurations for supranational foreign and security policy include the presence of REG and CONF, or the presence of REG, PUBID, or OPID, and the absence of MAT.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

### 5.3.2 Sensitivity analysis for the frequency threshold and sufficiency inclusion scores

Koenig-Archibugi's article does not report the sufficiency inclusion scores or the minimum frequency threshold employed in the analysis. Here again, we therefore focused mainly on carrying out a sensitivity analysis for all possible minimum frequency thresholds and sufficiency inclusion score pairs. These results for the first model with PUBID instead of OPID are shown in Fig. 7. Equivalent figures for the model containing OPID in place of PUBID are in the online supplementary materials.

With regard to replication, without knowing the author's parameter specifications for the reduction process, it is difficult to confirm that his results were replicated. Figure 7 does in fact include the causal configurations reported in the article. However, what is troubling for fsQCA is that these were by no means the *only* causal configurations found. Especially when the minimum frequency threshold is set to 1, the variability of fsQCA results is high. In fact, even at the highest levels for the minimum sufficiency inclusion score, the causal configurations yielded by the method change roughly every 0.06 threshold units.

### 5.3.3 Sensitivity analysis for all fuzzy set membership scores

Unlike for the previous replications, the Koenig-Archibugi article uses a normalized linear membership function to assign fuzzy set membership scores, which prevents the assignment of a crossover threshold. This complicates the task of testing for the sensitivity of fsQCA results to identification error in the crossover point. As a first-pass remedy, we applied the same methodology used for randomly moving the raw data anchor for the crossover point—i.e., the numerical value of some variable at which a case is considered equally in and out of a set—to the entire data set. That is, for each of the error simulations, every fuzzy set membership score for all cases was multiplied by a randomly chosen factor in $[1 - d, 1 + d]$. The technique is likely to yield results that are not directly equivalent to those yielded if only the crossover point were manipulated, but it still provides some general assessment of how sensitive fsQCA results are to minor changes in the fuzzy scores themselves.

The sensitivity of Koenig-Archibui's results (using PUBID instead of OPID) with the minimum frequency threshold set to 1 are shown in Fig. 8. While the results are stable up through 1% identification error, the findings begin to be corrupted at an error margin of 2.5%. By the time identification error reaches 5%, full contamination has set in—especially for results using minimum sufficiency inclusion scores of roughly 0.80 and below.

### 5.3.4 Monte Carlo simulations for random variables

Finally—and in a bit of an unfortunate milestone for fsQCA—over 10,000 simulations, every single causal configuration identified by the method as being sufficient for the outcome included at least one randomly drawn variable unrelated to the outcome. Thus, surprisingly, with 100% certainty either the presence or absence of a random variable is found to be part of a sufficient condition.

## 6 Conclusion

To be sure, fsQCA is one of the most innovative methodologies to emerge from social science that attempts to bridge the qualitative–quantitative divide. Its growing popularity and increasingly widespread use is a testament to the favor fsQCA has gained with researchers across several disciplines. However, this article has argued that fsQCA is a method whose results are questionably robust to even small changes in its calibration and reduction parameters, or to the incorrect specification of causal models. Our sensitivity analyses of the parameters required for utilizing fsQCA showed that the prime implicants identified by the method as being sufficient for a particular outcome are highly contingent upon specific values of the minimum sufficiency inclusion score, the maximum sufficiency inclusion score, the minimum frequency threshold, and the raw data anchor points. Furthermore, our Monte Carlo simulations of incorrect model specifications

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

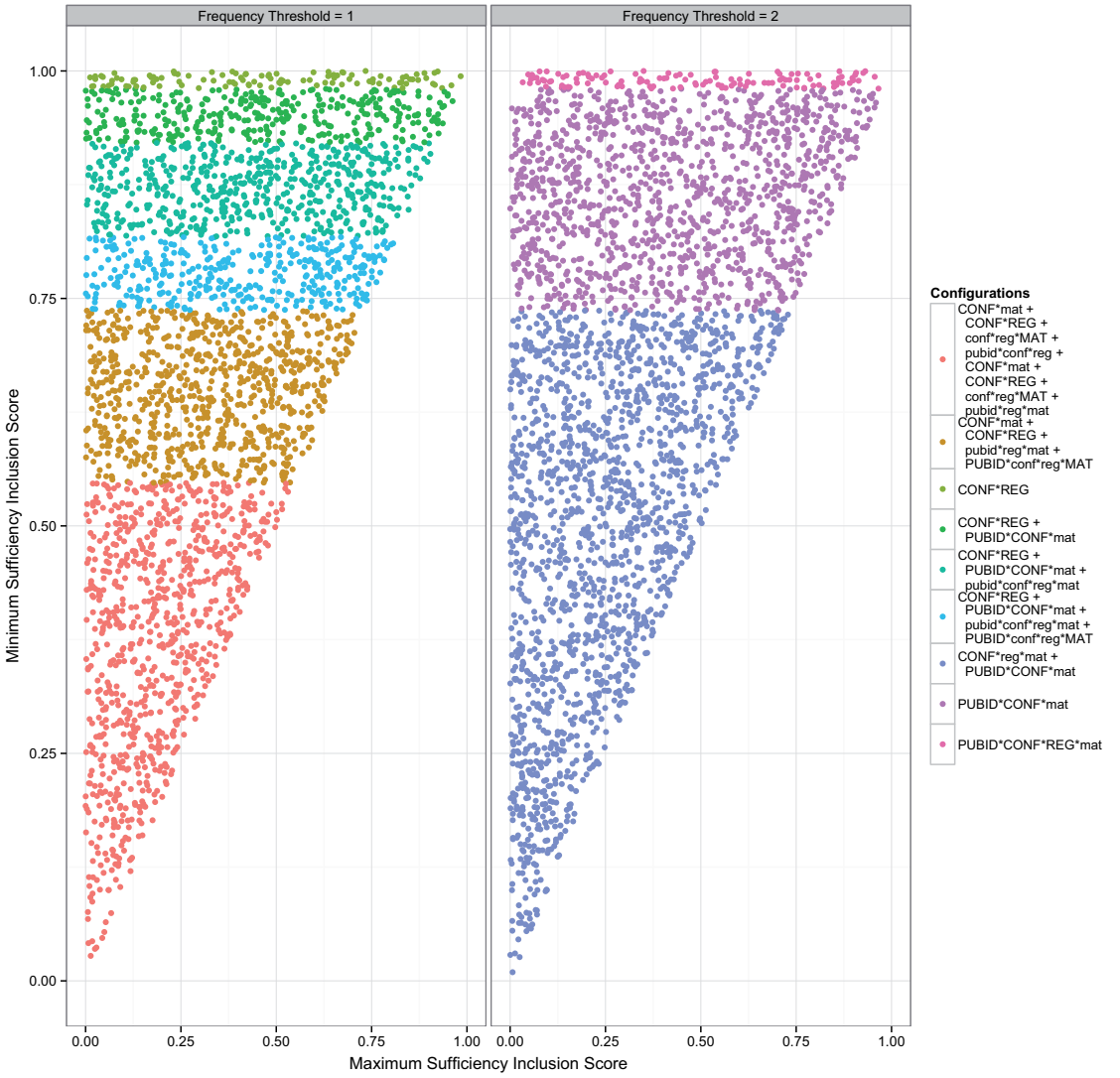


**Fig. 7** Sensitivity analysis for the frequency threshold and sufficiency inclusion scores, Koenig-Archibugi (2004), Model 1. A more interpretable color version of this figure is available online.

demonstrated that fsQCA is highly subject to confirmation bias, consistently failing to exclude random variables from the configurations found to be sufficient for a particular outcome.

Based on these results alone, however, we are not yet ready to suggest that researchers abandon fsQCA as a method for social scientific inquiry. It is entirely possible that an appropriately expert user could, in fact, assign theoretically meaningful and accurate values to the calibration parameters required by the fsQCA procedure and include only those explanatory factors that are causally connected to the outcome. And even short of these goals, some researchers may find fsQCA to be a helpful exploratory device for early-stage research. But we should be deliberate in trading the nuance and uncertainty of expert opinion for the overwhelming confidence and finality of a single number. For this reason, we suggest a departure from current practice when utilizing fsQCA. While there is little that can be done regarding the problem of false positives, a more adequate treatment of calibration and reduction parameters is possible, and could be valuable.

As it stands, users are asked to marshal their substantive case knowledge to pick calibration parameters in advance of attempting to identify prime implicants. This puts the consumer of fsQCA research at a distinct disadvantage relative to the researcher in being able to gauge how sensitive certain results are to parameter value choices, even when the researcher carries out the battery of robustness checks currently suggested by the literature (Skaaning 2011; Thiem 2013).

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

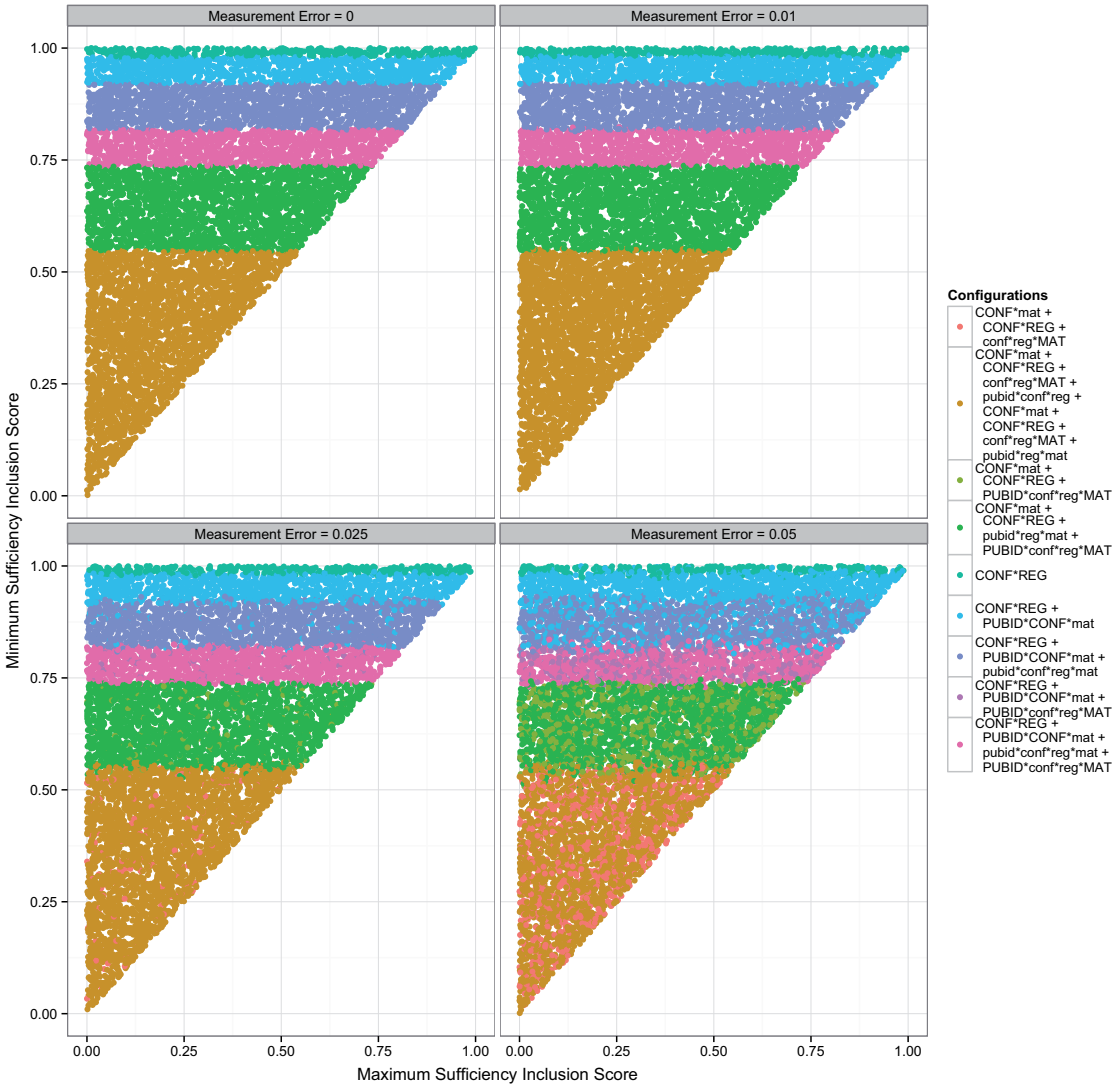


**Fig. 8** Sensitivity analysis for all fuzzy set membership scores, Koenig-Archibugi (2004), Model 1. A more interpretable color version of this figure is available online.

Instead of picking final parameter values in advance of initiating the search for prime implicants, fsQCA users could first report results for a large number of different values of each calibration parameter. Only after the reader has been presented with all possible fsQCA results derived from a large number of different parameter values should the researcher's substantive case knowledge be presented as justification for selecting specific parameter values. This step would increase the burden of proof placed upon the researcher, but would help convey the overall robustness of the results to the reader. Following such a procedure, though not a complete remedy for the method's limitations, would allow greater confidence in the excessively sensitive results derived from fsQCA.

## Funding

Choi acknowledges generous support from the Korea Foundation for Advanced Studies.

## References

Achen, C. 2005. Two cheers for Charles Ragin. *Studies in Comparative International Development* 40(1):27–32.

Ahn, S.-H., and S. S.-Y. Lee. 2012. Explaining Korean welfare state development with new empirical data and methods. *Asian Social Work and Policy Review* 6(2):67–85.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016

Brady, H. E., and D. Collier. 2010. *Rethinking social inquiry: Diverse tools, shared standards*. Lanham, MD: Rowman and Littlefield.

Freedman, D. 2010. *Statistical models and causal inference: A dialogue with the social sciences*. New York: Cambridge University Press.

Gerring, J. 2011. *Social science methodology: A criterial framework*. New York: Cambridge University Press.

Glaesser, J., and B. Cooper. 2014. Exploring the consequences of a recalibration of causal conditions when assessing sufficiency with fuzzy set QCA. *International Journal of Social Research Methodology* 17(4):1–15.

Huber, E., C. Ragin, J. D. Stephens, D. Brady, and J. Beckfield. 2004. *Comparative welfare states data set*. Northwestern University, University of North Carolina, Duke University, and Indiana University.

Hug, S. 2013. Qualitative comparative analysis: How inductive use and measurement error lead to problematic inference. *Political Analysis* 21(2):252–65.

King, G., R. O. Keohane, and S. Verba. 1994. *Designing Social Inquiry*. Princeton, NJ: Princeton University Press.

Koenig-Archibugi, M. 2004. Explaining government preferences for institutional change in EU foreign policy and security policy. *International Organization* 58(1):137–74.

Krogslund, C., D. D. Choi, and M. Poertner. 2014. Replication data for: Fuzzy sets on shaky ground: Parameter sensitivity and confirmation bias in fsQCA. Available at http://dx.doi.org/10.7910/DVN/27100 (accessed October 13, 2014).

Lucas, S. R., 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(1):199–204.

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

———. 2000. *Fuzzy set social science*. Chicago: University of Chicago Press.

Samford, S. 2010. Averting disruption and reversal: Reassessing the logic of rapid trade reform in Latin America. *Politics and Society* 38(3):373–407.

Schneider, C. Q., and C. Wagemann. 2012. *Set-theoretic methods for the social sciences: A guide to qualitative comparative analysis (QCA)*. Cambridge, UK: Cambridge University Press.

Seawright, J. 2005a. Assumptions, causal inferences, and the goals of QCA. *Studies in Comparative International Development* 40(1):39–42.

———. 2005b. Qualitative comparative analysis vis-à-vis regression. *Studies in Comparative International Development* 40(1):3–26.

Skaaning, S.-E. 2011. Assessing the robustness of crisp-set and fuzzy-set QCA results. *Social Methods Research* 40(2):391–408.

Thiem, A. 2014. Membership function sensitivity of descriptive statistics in fuzzy-set relations. *International Journal of Social Research Methodology* 17(6):625–42.

Thiem, A., and A. Duşa. 2013. QCA: A package for qualitative comparative analysis. *R Journal* 5(1):87–97.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 18 Aug 2017 at 20:40:41, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1093/pan/mpu016