Advance Access publication July 25, 2015

Political Analysis (2015) 23:471–487
doi:10.1093/pan/mpv017

Guarding Against False Positives in Qualitative Comparative Analysis

Bear F. Braumoeller
Department of Political Science, The Ohio State University, Columbus, OH 43210
e-mail: braumoeller.1@osu.edu (corresponding author)

Edited by: Jonathan Katz

The various methodological techniques that fall under the umbrella description of qualitative comparative analysis (QCA) are increasingly popular for modeling causal complexity and necessary or sufficient conditions in medium-N settings. Because QCA methods are not designed as statistical techniques, however, there is no way to assess the probability that the patterns they uncover are the result of chance. Moreover, the implications of the multiple hypothesis tests inherent in these techniques for the false positive rate of the results are not widely understood. This article fills both gaps by tailoring a simple permutation test to the needs of QCA users and adjusting the Type I error rate of the test to take into account the multiple hypothesis tests inherent in QCA. An empirical application—a reexamination of a study of protest-movement success in the Arab Spring—highlights the need for such a test by showing that even very strong QCA results may plausibly be the result of chance.

1 Introduction

A large and growing number of social scientists are being drawn to Charles Ragin's qualitative comparative analysis (QCA) as a methodology for understanding complex causation (Figure 1). The COMPArative Methods for Systematic cross-caSe analySis (COMPASSS) web site, an online compendium of QCA-related resources, lists 112 citations on QCA-related methodology and 362 applications in fields as diverse as comparative politics, sociology, health, environmental studies, business, and law.[1] Ragin's book *The Comparative Method* (Ragin 1987), which introduces the technique now known as crisp-set QCA, or csQCA, has amassed more than 5000 Google Scholar citations. A follow-up volume, *Fuzzy-Set Social Science* (Ragin 2000), which introduces fuzzy-set QCA (fsQCA), has been cited more than 2000 times. QCA-based findings have been used across all empirical subfields in political science and have been applied to practical questions as diverse and important as foreign policy (Paul and Clarke 2014) and health care decisions (Thygeson, Peikes, and Zutshi 2013). Even detractors acknowledge the influence of set-theoretic methods: as Monroe and Gold (2004, 1) put it in an otherwise highly critical article, "If we measured such things, Ragin's work would be a certified social science double platinum smash hit."

QCA—a blanket term denoting a family of techniques, including fsQCA, csQCA, and multi-valued QCA (mvQCA)—is designed for use in medium-N situations, when there are too many cases for systematic case studies but too few to bring traditional statistical inference to bear. The Boolean minimization procedure at the heart of QCA techniques allows analysts to find combinations of conditions that are strongly associated with outcomes (or nonoutcomes) and to do so with relative ease.

This inferential power comes at a cost, however. Because QCA techniques eschew statistical inference, they are not designed to answer questions that many practitioners and methodologists

Author's note: Thanks to Christopher Achen, David Collier, Jirka Lewandowski, the scholars who attended the 2014 summer seminars on Boolean logit at WZB Berlin, and those who attended my sessions at the 2014 IQMR Summer Institute in Syracuse, New York, for valuable feedback, and to Andrew Rosenberg and Austin Knuppe for invaluable research assistance. Replication Data are available on the Dataverse site for this article, http://dx.doi.org/10.7910/DVN/GY6P9I.

[1] http://www.compasss.org/bibdata.htm.

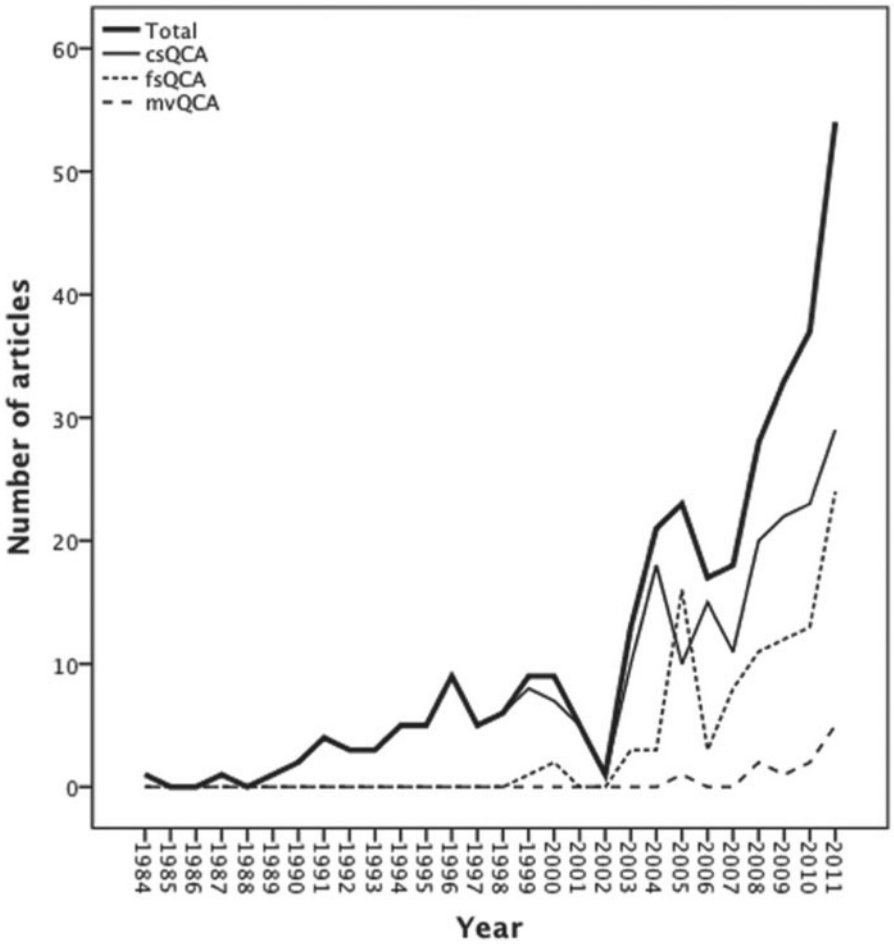

Fig. 1 Number of articles by QCA technique, 1984–2011 (Rihoux et al. 2013).

consider to be crucial. Foremost among these is the issue of how likely it is that the researcher could have found a given association by chance.

This question is made more complicated by the fact that, rather than starting out with a single hypothesized relationship among the data and implementing a test to ascertain how well or poorly the data conform to it, an analyst using set-theoretic methods starts out with a set of conditions and utilizes a truth table and a Boolean minimization procedure to find those combinations that are most consistently associated with the outcome. This practice amounts to testing a large number of possible hypotheses, and the results must be treated accordingly if we are to run an acceptable risk of false positives, or Type I error.

To address these issues, this article provides a new test that is designed to calculate the probability of encountering a given relationship between a Boolean combination of conditions and some outcome, or a more extreme one, by chance given that the relationship was arrived at via multiple hypothesis tests.[2] I describe both the test and the correction to the Type I error rate that must be made to account for the multiplicity of implied hypothesis tests. I then use the test to reevaluate a study of the determinants of protest-movement success in the Arab Spring (Hussain and Howard 2013). The results demonstrate that all four of the findings in the original article are plausibly the

[2]The replication file and data for the analyses described in this article have been archived at Braumoeller (2015). The procedures described herein are implemented in the `R` package `QCAfalsePositive`, available from the author or via the Comprehensive R Archive Network (CRAN).

result of random chance. The substantial margin by which these seemingly strong results fail the test underscores the magnitude of the threats to inference that this procedure addresses.

2 QCA and fsQCA

The QCA family of methodologies is designed to examine intersections among sets in order to uncover conditions under which outcomes are either nearly assured or nearly precluded. In language more familiar to conventional methodologists, that amounts to looking for empty cells in an *n*-dimensional crosstab, under the assumption that such cells tell us something interesting about the circumstances under which observations do not occur.[3]

A simple example will help to clarify this idea. Imagine that we categorize all states in the international system that have initiated wars since 1815 as being either democratic (D) or nondemocratic (d), using some reasonable threshold in the widely used Polity or Freedom House data as a cutoff. Imagine further that we categorize the targets of those war initiations as being either democratic or nondemocratic, using the same cutoff. In Ragin's terminology, these countries have been divided into *crisp sets*, to which they can either belong or not—there is no such thing as a degree of membership. The goal of csQCA is to tell us, for example, that the intersection of the set of democratic war initiators and the set of democratic defenders is empty. In so doing, we could either uncover or find evidence in favor of an intuition about the conflict behavior of democratic states.[4]

When more than two sets are involved, csQCA seeks to uncover Boolean combinations of sets that produce a given outcome with a high degree of regularity. This involves constructing a truth table that lists all possible configurations of sets and the outcomes with which those configurations are associated and then utilizing an algorithm to reduce that truth table to a series of prime implicants. The reduced-form equation that results from the Boolean minimization algorithm succinctly describes the set intersections that are most consistently associated with the outcome. For example, a study that explores the relationship among soil quality, rainfall, and crop growth might find that growth is dramatically curtailed in the absence of either good soil or rainfall—that is, that the set "crops that grow well" is a subset of the intersection of the sets "crops grown in good soil" and "crops grown in areas with dependable rainfall."[5]

While the Boolean combination of sets might suggest a parallel between QCA-based techniques and multiplicative interaction terms in regression equations, the similarity is superficial: whereas statistically significant interaction effects may be substantively trivial, the goal of methods in the QCA family is to find relationships with an extraordinarily high degree of substantive significance, that is, conditions under which outcomes are either nearly assured or nearly precluded. These are often described as combinations of conditions that are either necessary or sufficient to produce a given outcome, though many social scientists shy away from language that suggests a deterministic relationship among variables.

The democratic peace example also highlights two of the major shortcomings of crisp-set QCA: it often forces phenomena that are conceptually continuous, like wealth and political democracy, into discrete categories, and it is vulnerable to small numbers of counterexamples. In the case of the democratic peace, the existence of a single war between democracies would ensure that csQCA returns a null result, even if the relationship between democracy and war is otherwise very strong. Moreover, as Hug (2013) points out, csQCA's sensitivity to counterexamples makes it vulnerable to even a small amount of measurement error.

With concerns like these in mind, Cronqvist and Berg-Schlosser (2009) developed multi-value QCA, which is simply crisp-set QCA adapted to include categorical variables with more than two values, and Ragin developed fuzzy-set QCA (Ragin 2000, 2008), a refinement of csQCA that is

[3]The distinctions between configurations of variables and configurations of cases, and between subset relationships and correlations, are a constant theme in writings about QCA. Nevertheless, the isomorphisms that exist between the two conceptual frameworks make these distinctions more important in theory than they are in practice.

[4]In fact, Michael Doyle (1983a, 1983b) followed a procedure very much like this one in his groundbreaking study of the democratic peace.

[5]Nikita Khrushchev discovered precisely this relationship during the Virgin Lands Program of the 1950s–60s, in fact.

designed to take into account degrees of membership in sets and to allow for a limited number of counterexamples. To continue the above example, an analyst using fsQCA could code each country's degree of membership in the sets of democratic initiators and democratic defenders and then explore the extent to which a certain degree of membership in the set of democratic initiators implies a similar (or lesser) degree of membership in the set of democratic defenders. The visual result of such an investigation would be a scatterplot with many observations below the $X = Y$ diagonal and few if any observations above it. In set-theoretic terms, the set of nondemocratic defenders would be a subset of the set of democratic initiators.

Because relationships in social science are rarely perfect, Ragin introduces the concepts of *consistency* and *coverage* to quantify the degree of imperfection and to ensure that it remains within reasonable bounds. Consistency is defined as the percentage of consistent cases in Ragin (2000) but is refined to take into account near-misses, becoming $\frac{\sum \min(X, Y)}{\sum X}$ for results in which $X \leq Y$ (sufficient conditions) and $\frac{\sum \min(X, Y)}{\sum Y}$ for results in which $X \geq Y$ (necessary conditions) in Ragin (2008, 52–53).

3 Inferential Hazards

As useful as the QCA family of techniques is, it does not address one of the most fundamental questions in statistical inference: How often would we find data that are equally or more consistent with the hypothesis just by chance—that is, if the variables (or sets) were really unrelated? More succinctly, what is the probability of finding a false positive result? This problem is exacerbated by the fact that the Boolean minimization procedure at the heart of QCA reports only the best of a (sometimes substantial) number of results, a fact that renders ordinary p values problematic.

I address each issue in turn. Because fsQCA is the most nuanced of the three techniques and its goodness-of-fit metrics require special care, I spell out both the inferential challenges and their solutions in the context of fsQCA, after which I return to the simpler cases of csQCA and mvQCA.

3.1 *False Positives*

While the patterns of data returned by techniques in the QCA family can appear quite compelling, scholars rarely if ever pay attention to the question of how likely they are to have occurred by chance. Although some extant methods might seem applicable to this task, none is ideal. Tests of necessary conditions (e.g., Dion 1998; Braumoeller and Goertz 2000), while they could be adapted to test sufficient conditions as well, are designed for binary rather than continuous variables[6] and tend to base their tests on asymptotic properties that are unlikely to hold in precisely the middle-N situations envisioned by researchers using set-theoretic methods. Marx and Dusa (2011) establish lower bounds on combinations of cases and conditions to ensure that fsQCA results can even *in principle* be distinguished from chance variation, but they do so by generating random data and measuring the number of false positives rather than by constructing a test designed to assess the probability that any given outcome was the result of random variation.

The most relevant procedure to date is the goodness-of-fit test created by Eliason and Stryker (2009), which is primarily designed to test the hypotheses of perfect sufficiency and perfect necessity in fuzzy-set data. It does so by assessing the degree to which the data do, or do not, conform to the perfect upper-triangular and lower-triangular data patterns implied by such hypotheses. The hypotheses are rejected if the number and magnitude of any counterobservations in the data exceed what the researcher would expect due to measurement error.[7] In so doing, the authors

[6]The exception is Braumoeller and Goertz (2000), which describes a test for continuous necessary conditions—but the test is univariate and not readily applicable to this special case.

[7]One substantial drawback to the test is that the magnitude of that measurement error must be assumed (115–116), and the accuracy of that assumption, which cannot be evaluated, is critical for the result.

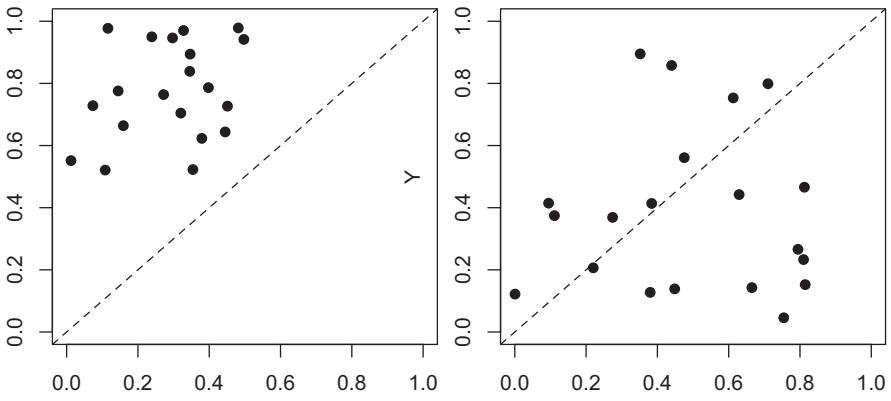

Fig. 2 Two hypothetical data sets in which Y and X are independent.

also describe a null hypothesis test (110–111), the goal of which is to test the argument that the outcome variable, Y, and the combined set membership, X, are statistically independent.

A null hypothesis test of statistical independence is—arguably—appropriate in the context of ordinary least squares and related techniques, in which the test statistic reflects the degree of correlation. Critics have pointed out, with good reason, that the null hypothesis may be uninteresting or implausible. In fsQCA, however, the null hypothesis is both meaningful and plausible: with a small or medium-sized n and quite a few possible configurations of sets, it is entirely possible to produce a "finding" from entirely uncorrelated data.

As researchers who use set-theoretic methods have made clear, however, the logic of set-theoretic methodologies is very different from that of correlational ones (see, e.g., Ragin 2000, chap. 1). Consider, for example, the two hypothetical data sets plotted in Fig. 2. In both cases, Y is statistically independent of X, but fsQCA researchers would interpret them very differently: while the right-hand plot resembles a classic null result, the left-hand plot would be taken as strong evidence in favor of a sufficiency hypothesis.

While traditional quantitative researchers might argue that a strong finding based on uncorrelated data reflects a problem with fsQCA, to do so is to miss the point. The primary test statistics in fsQCA are based on the number and magnitude of *disconfirming* cases: the consistency measure takes into account the distance between counterexamples and the $Y = X$ diagonal, and the counterexamples measure takes into account their number. The distribution of cases that conform to the hypothesis—those above the $Y = X$ line for sufficient conditions and below it for necessary conditions—is immaterial to either measure. A condition is not "more necessary" if the conforming cases lie closer to the $Y = X$ line; from an fsQCA perspective, the only thing that matters is that $X > Y$.

What that implies is that *only the distribution of cases that fail to conform to the hypothesis is relevant to the test statistic*. The relevant null hypothesis for fsQCA is not that Y and X are independent but rather that an equal or greater number of counterexamples, or an equal or lower degree of consistency, could plausibly have been observed by chance.[8] Because the distribution of conforming cases is immaterial to the test statistic, it should not matter to a test of the null hypothesis. Yet the distribution of such cases, which constitute the overwhelming majority of observations in most studies, is the focus of the goodness-of-fit test.

What fsQCA still lacks, therefore, is a null hypothesis test that is tailored specifically to its unusual test statistics. Given the richness of the fsQCA research agenda and the number of

[8]To give credit where it is due, the necessary- and sufficient-condition goodness-of-fit tests in Eliason and Stryker (2009) are designed with this intuition in mind, unlike the null hypothesis test. They are unsuitable as null hypothesis tests themselves, however, for two reasons. First, while the tests assume measurement error, the magnitude of that measurement error must be assumed (115–116), and the accuracy of that assumption, which cannot be evaluated, is critical for the result. Second, the test assesses the fit of the data to the necessity-sufficiency hypothesis relative to their fit to the null hypothesis. I have already argued that the latter is problematic.

studies that it has spawned, the absence of such a test is cause for alarm. With a relatively small number of observations, the probability that all or nearly all of the observations in a data set would just happen to lie above or below the line $X = Y$ may not seem great, but it cannot be discounted either.

3.2 *Multiple Inference*

The danger of false positives is especially acute in the QCA family of procedures because they explicitly consider many possible relationships among variables before reporting only the strongest of them. For example, Krogslund, Choi, and Poertner (2015) reexamine three prominent studies that use fsQCA and find that completely random variables added to the model specifications are identified as part of at least one sufficient condition quite often—from 75.2% of the time to 100% of the time, depending on the study. As dire as these results may sound, the practice of examining multiple relationships is not at all problematic from the point of view of hypothesis testing, *if* it is done in a statistically principled manner. What that implies is accounting for the problem of multiple inference—performing more than one test on the same data—that is inherent in these techniques.

Multiple inference is a nontrivial issue for QCA-based techniques because the truth tables upon which QCA inferences are based explore many, and in some cases all, possible combinations of sets, each one of which must be counted as a hypothesis test. For example, if there are three sets, A, B, and C, and uppercase letters denote membership while lowercase letters denote nonmembership, possible solutions include ABC, ABc, AbC, Abc, aBC, aBc, abC, and abc. If each combination is examined as a possible solution, eight hypothesis tests have taken place, and each must be considered a separate test for the purposes of assessing the risk of a false positive.

Given that the number of possible combinations of n sets is 2^n, accounting for multiple inference could in principle make it very difficult for fsQCA results to pass a null hypothesis test. In practice, however, the problem of limited diversity implies that analysts often encounter *logical remainders*, or sets of conditions that are not found in any observed case. In the above example, if we observed a nonzero number of cases with characteristics ABc, AbC, Abc, aBC, aBc, and abC, but none with ABC or abc,[9] and if logical remainders were dropped from the truth table prior to minimization, the number of implied hypothesis tests would be 6 rather than 8.

That said, it is not uncommon for analysts using fsQCA to incorporate logical remainders by making plausible simplifying assumptions (Ragin 2004; Schneider and Wagemann 2013, 211). These amount to counterfactual arguments that, had such a combination of characteristics been observed, theory and past experience lead us to conclude that we would also have observed a certain outcome. Analysts may also simply assume that unobserved combinations of sets are associated with negative outcomes. Indeed, Ragin, Strand, and Rubinson (2008, 79–81) specifically instruct users of their fsQCA software to code all remainders as negative outcomes in order to obtain the most complex solution.

The upshot is that the number of implicit hypothesis tests (call it k) is equal to the number of configurations—hypothetical or not—used by the truth table algorithm in fsQCA. If all logical remainders are dropped and all observed configurations are retained, k is equal to the number of distinct observed configurations. If all logical remainders are retained, k is equal to 2^n. If some logical remainders are retained, k is somewhere between these two extremes.

The magnitude of k is consequential in fsQCA because the key question is not whether a single observed pattern of data could be the result of chance: the question is whether one or more such patterns might emerge by chance when the algorithm considers k configurations of cases. Whereas

[9]The continuous coding of set-membership values in fsQCA makes the discussion of limited diversity a bit awkward. To clarify, a case is considered to be a member of a set if its membership for that set is greater than or equal to 0.5. For that reason, we might never observe (say) aBC, aBc, or abc if all cases have values of A that are greater than 0.5.

most standard statistical tests must specify a Type I error rate of (say) 0.05 for a single test, analysts using fsQCA must specify a Type I error rate for *each* test that ensures that the Type I error rate for *all* of them will not exceed 0.05.

3.3 *A Brief Aside: Statistical Inference and QCA*

Before proceeding to discuss solutions to these problems, it would be useful to address one objection head-on: Is statistical inference an appropriate component of an fsQCA analysis? As one reviewer of an earlier version of this article argued, "there is a differentiation in the social science community which has made it possible to develop different approaches to social science questions. Those who use QCA do not share the quantitative template which underlines the importance of inferences; otherwise, they would in fact use statistical techniques."

The most reasonable response to this argument is to point out that users of QCA do, in fact, use statistical tests. Although csQCA and fsQCA are not statistical approaches, methodologists have designed some statistical tests to complement them. Ragin himself (2000, 226–229), for example, proposes a simple p-test in fsQCA to assess whether the observed proportion of cases that are consistent with the hypothesis of sufficiency, $\hat{p}$, is consistent with a hypothesized population (or "benchmark") proportion of p, for a given significance level α, and Goertz, Hak, and Dul (2012) propose the use of quantile regression in fsQCA to establish a boundary between regions in which observations are plentiful and those in which observations are sparse.

While advocates of QCA do sometimes resist the idea that set-theoretic methods are amenable to traditional hypothesis tests, their resistance seems to have more to do with the process by which conclusions are arrived at than the conceptual disparities between case-based and variable-based methods. Schneider and Wagemann (2012, 296), for example, point out that researchers using set-theoretic methods often "move back and forth between ideas and evidence" by refining their hypotheses, cases, and evidence based on preliminary results. The authors explicitly note (fn 20) that this is not an argument against the applicability of statistical inference to set-theoretic results; rather,

> when using set-theoretic methods, one either adheres to the standards of good practice of carefully crafting the data and thus cannot engage in straightforward hypothesis testing, or one performs proper hypothesis tests, which, however, can only be done by violating the standards of good practice of set-theoretic methods. (296)

While the authors' complaint about the mindless application of hypothesis tests is a reasonable one, the conclusion does not follow. In fact, there *are* proper hypothesis tests that accommodate researchers' practice of moving back and forth between ideas and evidence, which amounts to another source of multiple inference. The fact that researchers using set-theoretic methods are urged to do so by best practices underscores the need to address the issue systematically.

Moreover, researchers using QCA often argue that any pattern uncovered by QCA is only as good as the case knowledge that went into it and the within-case evidence on mechanism that backs it up. This raises an interesting question: what should we conclude if a solution formula backed up by convincing within-case evidence on causal mechanisms linking the solution to the outcome nevertheless fails to pass a statistical test? If this were to be the case, a null finding would indicate that *only* the within-case evidence warrants the conclusion. Similar situations often arise in, for example, medical case studies, where careful observations made on a single patient or a small number of patients offer valuable insights that are nevertheless not warranted by a controlled, double-blind study.

4 The Solutions

Solving the problem of potential false positives in fsQCA requires two steps. The first involves deriving a specific statistical test to assess the probability that the upper- or lower-triangular patterns of data typical of fsQCA findings arose as a result of chance. The second involves adjusting

the Type I error rate (α) of the individual tests to ensure a family-wise, or overall, error rate (FWER) that is in line with the standards of the field.[10]

4.1 *Permutation Tests*

The permutation test is one of the most flexible and valuable tools that could be adopted for use with fsQCA. Designed by Sir Ronald Fisher in the 1930s, permutation tests are designed to ascertain the probability that an observed difference or pattern across groups of observations arose by chance. They do so by permuting the group labels many times and counting the number of instances in which the permuted outcomes are as extreme as, or more extreme than, those observed in the real world.

Imagine, for example, that we have data on the heights of twelve men and twelve women, that the men are on average a bit taller than the women, and that we want to know whether their height difference is plausibly the result of chance. We would start by holding the vector of heights constant, reassigning the male–female group labels at random, and measuring the height difference in the permuted data. We would then repeat this procedure hundreds, thousands, or tens of thousands of times in order to find out how extreme our observed height difference would be if it had occurred by chance in a world in which sex is unrelated to height. The p value for the test is simply the fraction of differences in the permuted data that are as extreme as, or more extreme than, the observed difference.

Exact permutation tests, which examine all possible permutations, give exact p values but can be computationally expensive even in medium-N settings: twenty observations, for example, produce a staggering number of permutations—more than two quintillion. Fortunately, Monte Carlo permutation tests are asymptotically equivalent to exact permutation tests. Although there is some uncertainty surrounding the p value in a Monte Carlo permutation test, it is possible to quantify that uncertainty using the binomial distribution: $\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{k}}$, where k is the number of trails, gives the 95% confidence intervals around $\hat{p}$ (Good 2005).

The advantages of permutation tests are considerable. They do not in general rely on distributional assumptions. They are exceptionally versatile: the test statistic can be a difference in means, a difference in variances, or any more specialized quantity implied by a hypothesis. They are capable of uncovering statistically significant differences even in very small samples. And they do not rely on random sampling, because they are designed for within-sample inference—though as Ludbrook and Dudley (1998) note, if the sample is drawn at random from a larger population, a strong case can be made for generalizing back to that population based on within-sample differences. The main assumption upon which they rely is *exchangeability*—a fairly mild assumption that is closely related to the i.i.d. assumption in statistical studies.[11] For all of these reasons, permutation tests are especially useful in exactly the medium-N situations that are common in fsQCA analysis.

4.2 *Joint Hypothesis Tests*

Because exact FWERs depend on the independence of samples, observations, and variables (or sets), no single formula exists to correct for multiple inference across different fsQCA-based studies. We can, however, establish an upper bound for the FWER. The most straightforward way to do so

[10] Jirka Lewandowski (private communication) has argued in favor of using the false discovery rate (FDR; see Benjamini and Hochberg 1995) rather than the FWER. The FWER seeks to control the probability of even one false positive among the results, while the FDR seeks to control the expected *proportion* of false positives. In practical terms, the FDR provides a more powerful test than the FWER, but it allows a greater number of false positives. My own sense is that the FWER more closely reflects the concerns of social scientists, but reasonable scholars may differ. None of the test results presented in this article would differ if the FDR were to be used instead.

[11] Formally, exchangeability is the property that the order of a sequence of random variables does not affect its joint probability distribution. See Good (2005, 268) for details. The assumption is fairly innocuous because, if exchangeability does not hold, the researcher almost certainly shouldn't be using any kind of QCA technique in the first place.

would be to use the Bonferroni correction (Welkowitz, Cohen, and Lea 2012, chap. 13), which adjusts the error rate for each of k tests to $\frac{\alpha}{k}$ in order to ensure an FWER of no more than α. The Bonferroni correction is a rather conservative adjustment, however, that may well overcompensate by producing an FWER well below α.

The Holm–Bonferroni test (Holm 1979) is an iterative procedure that is both more powerful than the original Bonferroni correction and better suited to QCA.[12] For a desired Type I error rate α, the procedure is as follows:

- Run permutation tests on each of the k possible combinations of sets, saving the estimated p values.
- Assign each of the p values a rank, r, from largest/least significant ($r = 1$) to smallest/most significant ($r = k$).
- Calculate adjusted p values, $p_{\text{adj}} = pr$.
- Starting with the *lowest*-ranked test (i.e., the one with the smallest raw p value) and moving in order to the highest, reject the null hypothesis for every combination in which $p_{\text{adj}} \leq \alpha$ until you reach a combination for which $p_{\text{adj}} > \alpha$. Fail to reject the null hypothesis for that combination and every higher-ranked combination.

Because the test is sequential and the failure of one test to reject the null hypothesis implies the failure of all higher-ranked tests, the values of p_{adj} must logically be nondecreasing. For this reason, the adjusted p value for a test of the rth hypothesis is often calculated as the maximum of p_{adj} for hypotheses r through k.

Because of its mathematical simplicity, the Holm–Bonferroni adjustment provides an easy and statistically powerful way for practitioners to correct for multiple inference in fsQCA.

5 The Null Hypothesis Permutation Test

The goal of null hypothesis significance testing in QCA is to ascertain the probability that the observed pattern of data would have arisen by chance if the variables (or sets) are unrelated, taking into account the multiple inference inherent in the Boolean minimization algorithm.

One challenge facing such a test in the fsQCA context is the fact that the underlying distribution of data is unknown, so tests that rely on distributional assumptions will be problematic to an unknowable degree. Another challenge is that scholars using fsQCA are not interested in typical quantities of interest, such as the average increase in Y given a unit increase in X. Their focus tends to be on the number of observed counterobservations for a given grouping or the consistency of a given Boolean solution. Permutation tests are exceptionally well suited to both challenges.

The first step is to specify the test statistic. If we are interested in the number of counterobservations, the test statistic is simply the number of observations for which $X > Y$ for upper-triangular, or sufficient condition, relationships or the number of observations for which $Y > X$ for lower-triangular, or necessary condition, relationships. If we are interested in consistency, the test statistic is $\frac{\sum \min(X, Y)}{\sum X}$ for sufficient condition relationships and $\frac{\sum \min(X, Y)}{\sum Y}$ for necessary condition relationships.

The second step is to permute the data many times by reassigning the values of the dependent variable at random, without replacement, and then calculate the test statistic for the permuted data. Following Good (2005, chap. 9), the values of the independent variables for a single observation are maintained as a single, indivisible vector. The number of iterations necessary depends on both the FWER and the desired level of specificity. If, for example, five tests had been carried out, so that the most statistically significant test would have to reach $p < 0.01$ in order to ensure $\alpha = 0.05$, we would need at least 100 iterations in order to achieve the granularity to carry out a test at the desired level of significance. The estimated $\hat{p}$ would not be very precise given that number of iterations, however:

[12] Still more powerful tests exist, such as the Šidák correction, but many have the disadvantage of requiring that the hypothesis tests be independent, which they are surely not in QCA.

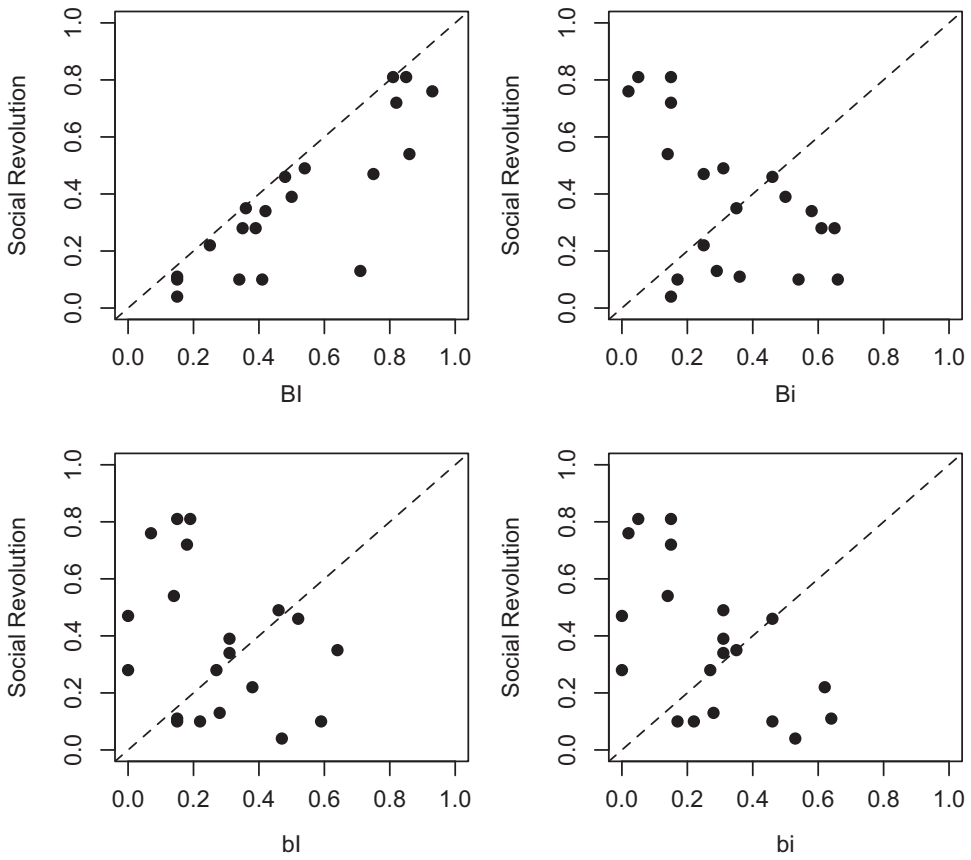

Fig. 3 Necessary conditions for social revolution (Ragin 2000).

if $\hat{p} = 0.01$, the standard error would be 0.01, rendering the test essentially useless for low values of $\hat{p}$. Increasing the number of iterations to 10,000 produces a much smaller standard error (0.001 if $\hat{p} = 0.01$, e.g.), which is useful for a much wider range of estimates of $\hat{p}$. Experimentation suggests that a good rule of thumb is to begin with $\frac{100k}{\alpha}$ permutations, calculate the standard error of $\hat{p}$, and if $\hat{p}$ is insufficiently precise to evaluate the null hypothesis, increase the number of permutations accordingly.

The final step is to calculate the adjusted p values, $\hat{p}_{\text{adj}}$, for the tests with the lowest p values and find the point at which $\hat{p}_{\text{adj}} > \alpha$, as described above, as a means of calibrating the probability of Type I error to take into account the multiple tests that are implied by the Boolean minimization procedure.

5.1 *Example: Ragin (2000)*

In the book that introduces fuzzy-set QCA, Ragin (2000, 218–221) offers an example that serves well to illustrate this process. He describes a hypothetical study of social revolutions with twenty outcomes. The two phenomena that are thought to be necessary but not sufficient for social revolution are popular insurrection and state breakdown. The cases are conceived of as having fuzzy membership in each of the three relevant sets (social revolution, popular insurrection, and state breakdown).

In Fig. 3, I illustrate the relationship between social revolution and each of the $2^2 = 4$ possible configurations of sets: state breakdown and insurrection (BI), state breakdown and no insurrection (Bi), no state breakdown and insurrection (bI), and neither state breakdown nor insurrection (bi). As the upper-left graph, which coincides exactly with Ragin's Figure 8.2, shows, when breakdown and insurrection are both present the observations form the lower-triangular pattern typical of

Table 1 $\hat{p}_{\text{adj}}$ for counterexamples and consistency from the Ragin (2000) social revolution example

	Counterexamples		*Consistency*	
Combination	$\hat{p}$	$\hat{p}_{adj}$	$\hat{p}$	$\hat{p}_{adj}$
BI	0.0000	0.0000	0.0000	0.0000
Bi	0.0624	0.1873	0.9296	1.0000
bI	0.9037	0.9090	0.7595	1.0000
bi	0.4545	0.9090	0.9529	1.0000

necessary-condition relationships. The same is true for breakdown and the absence of insurrection (Bi). For the other two combinations, bI and bi, the pattern of the data more closely resembles an upper-triangular or sufficient-condition relationship.

There are two possible test statistics that we might wish to examine: the number of counterexamples and the consistency of the data with the posited relationship $X \geq Y$, measured as $\frac{\sum \min(X, Y)}{\sum Y}$. Both were calculated for each of the four possible combinations of sets. Each combination was permuted 100,000 times because the raw p values for some combinations were very small. The value of $\hat{p}_{\text{adj}}$ was then calculated for each combination by multiplying the raw p value by the rank of that p value: 4 for the smallest p value, 3 for the next smallest, 2 for the next smallest, and 1 for the largest. Adjusted p values greater than 1 were rounded to 1, and p_{adj} values were adjusted to be nondecreasing.[13]

Of the four possible combinations, only one—the intersection of state breakdown (B) and popular insurrection (I)—produced values of $\hat{p}_{\text{adj}}$ below 0.05 for both the counterexamples and consistency test statistics. The raw p value for the third-ranked combination in terms of counterexamples, state breakdown in the absence of popular insurrection, came close to standard levels of significance, though the adjusted p value does not, so we are unable to reject the null hypothesis for this and all remaining tests. Had both BI and Bi passed their respective tests, the reduced-form formula B would have done so as well. Accordingly, for both consistency and counterexamples the sole combination to pass the null hypothesis test is BI. These results are shown in Table 1.

The example also highlights an interesting distinction between using counterexamples and using consistency as the test statistic. When state breakdown is present but insurrection is not (Bi; upper-right graph in Fig. 3), there are seven counterexamples out of twenty total cases. The scatterplot does not suggest a compelling relationship, but as it happens, the permutations rarely turn up a larger number of counterexamples. Accordingly, the permutation test for counterexamples very nearly reaches standard levels of statistical significance. The extremity of those misses dramatically reduces the consistency metric, however—so much so that the permutation test for consistency comes nowhere near statistical significance. The latter result comports more closely with intuition because consistency incorporates both the number of misses and their extremity.

To illustrate the workings of the permutation test in more detail, I have graphed the results of the test for the best-fitting combination, BI, in Fig. 4. The histogram on the right shows the distribution of counterexamples across all of the 100,000 permutations of the data. The small dot along the X axis indicates the number of counterobservations in the original data. As the figure suggests, zero counterexamples is an extraordinarily rare outcome when the data are permuted. The density on the left represents the distribution of consistency scores across all 100,000 permutations. Again, the observed consistency, 1.0, is extremely unusual. In both the histogram of counterexamples and the distribution of consistency scores, the rejection region for the null hypothesis, defined by the one-sided $1 - \frac{\alpha}{4}$ confidence interval,[14] is shaded a darker. Because the observed number of

[13]For example, the calculated value of $\hat{p}_{\text{adj}}$ for counterexamples in the case of the combination bi would be $0.59302 \times 1 = 0.59302$, but because the hypothesis test for bi and hypothesis test for the next-highest-ranked combination (bI) are sequential, the corrected value is the maximum of the two, or 0.71718.

[14]The rejection region for the second most statistically significant result would be the one-sided $1 - \frac{\alpha}{3}$ confidence interval, and so on down the line.

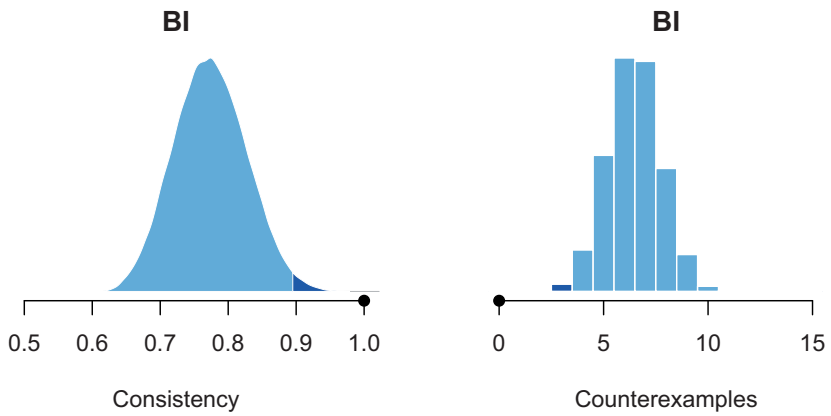

Fig. 4 State breakdown, popular insurrection, and social revolution. Dots along *X* axis indicate observed consistency and number of counterexamples in original study.

counterexamples (0) and observed consistency (1) are well within the rejection region, we can safely reject the null hypothesis that the observed association occurred by chance.

It is reasonable to wonder whether the multiple-inference penalty applies to researchers who eschew using the Quine–McCluskey algorithm in favor of simply testing individual necessary conditions. In this case, for example, the necessity of B and I follows logically from the necessity of BI, so a researcher who tests only B and I as necessary conditions might reasonably claim only to have conducted two tests. Strictly speaking, that would be correct, but only if other possible configurations are never considered, even informally: if a researcher generates plots of ten possible configurations and rejects nine of them without a formal test because they obviously wouldn't pass it, ten tests have still taken place.

As this example demonstrates, the null hypothesis permutation test is ideal for fsQCA applications. It is designed specifically for the unusual, asymmetric hypotheses that fsQCA seeks to assess. It provides a statistically principled means of gauging the probability of Type I error, adjusting for the fact that the Boolean minimization algorithm examines many possible combinations of sets and practitioners report only the best-fitting among them. The result is a stringent test that can nevertheless still be passed, at least in principle, even with a modest number of observations.

In the next section, I reexamine a study of protest cascades in the Arab Spring in order to highlight the dangers of multiple inference and to show that even very strong fsQCA results may not be as improbable as they first seem.

6 Application: Protest Cascades

While Ragin's hypothetical social-revolution example is a useful illustration of how the permutation test works under ideal conditions, it remains to be seen how well it works in practice. Hussain and Howard (2013) is an excellent candidate for reexamination for this purpose, both because its findings speak to an important substantive question (the sources of successful Arab Spring uprisings) and because its findings are impressively consistent (in the fsQCA sense of the word).

Hussain and Howard set out to uncover the sources of successful protest cascades, using the countries involved in the Arab Spring of 2011 as cases. They examine nine potential sources of protest-cascade success: degree of authoritarianism, per-capita GDP, income inequality, unemployment, urbanization, the size of the "youth bulge," the degree of mobile phone connectivity, the degree of internet connectivity, and the extent to which the economy depends on oil production.[15]

[15]The authors mention one additional measure, the sophistication of the regime's Internet censorship capacity, but it is unclear whether this measure was used in the empirical analysis. I assume that it was not, though little hinges on the question one way or the other.

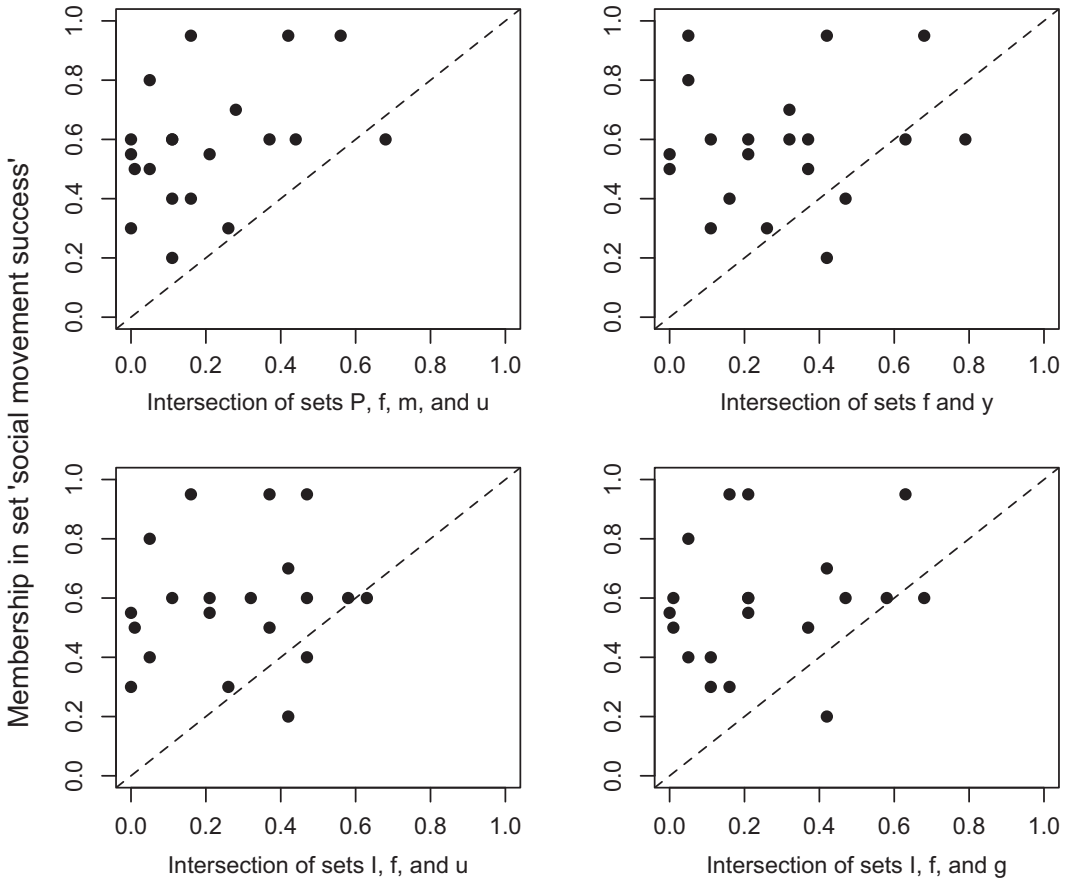

Fig. 5 Sufficient conditions for protest cascades (Hussain and Howard 2013).

The authors dropped all logical remainders from the analysis; their twenty cases left them with a truth table comprising fifteen rows, each of which corresponds to an observed configuration.[16]

The study finds that successful cascades are most likely in four different settings:

- authoritarian, non-oil-producing states with few mobile phone users and low unemployment (i.e., the intersection of sets P, f, m, and u, where capital letters denote membership and lowercase letters denote nonmembership);
- the absence of both oil dependence and a youth bulge (the intersection of sets f and y);
- high Internet connectivity in the absence of either high oil dependence or unemployment (the intersection of sets I, f, and u); or
- high Internet connectivity in the absence of either high oil dependence or inequality unemployment (the intersection of sets I, f, and g).

Figure 5 shows the relationships among these intersections and membership in the set of social movement success. Simple visual inspection suggests that the relationships uncovered are quite strong: in almost every case, a given level of membership in the intersection of the relevant sets is associated with at least the same level of membership in the set of social movement success. Put differently, the presence of a given quantity of the combined independent variables is sufficient, or very nearly so, to ensure the presence of the same quantity of the dependent variable—hence the near-absence of cases below the line $X = Y$. Accordingly, each of the four relationships is

[16]Personal communication with the authors, September 16, 2014. Some unpublished analyses dropped some of the less consistent cases and were carried out on an eight-row truth table, but the results reported in the article used all fifteen rows.

characterized by a high degree of consistency, ranging from 0.87 for the intersection of f and y to 0.95 for the intersection of P, f, m, and u.

Substantively, these conclusions reaffirm some parts of the conventional wisdom regarding the Arab Spring but run strongly contrary to others—in particular, to findings that suggest that demographics and unemployment both contributed to the spread of Arab Spring uprisings (see, e.g., Ansani and Daniele 2012; LaGraffe 2012). Moreover, where the conclusions do comport with conventional wisdom, the connections are sometimes tenuous. For example, the authors' conclusion that Internet connectivity, which appears in two of the four patterns associated with success, was a crucial determinant of success is odd given that nonmembership in the set of petrostates (f) appears in all four.

The fact that the findings are derived from a nontrivial number of hypothesis tests heightens concerns that the results may be spurious. In order to find any significant relationships at all given that fifteen separate configurations were tested, the raw p score of the most significant relationship (Pfmu) would have to be $\frac{0.05}{15} = 0.003$ or lower. Regardless of whether we use counterexamples ($\hat{p} = 0.30$) or consistency ($\hat{p} = 0.16$) as our test statistic, the estimated p value is far from significant.

This calculation highlights a potential pitfall when QCA is used in the medium-N situations for which it is explicitly designed: as the number of configurations increases, so does the number of implied hypothesis tests, and the analyst is forced to adopt an increasingly Draconian adjustment to the test's p values in order to mitigate the risk of false positives. Even data that conform perfectly to his or her hypothesis, in a medium-N world, might never be so improbable that they could rule out chance association. The distributions of 400,000 sets of permuted data (100,000 per row) in Fig. 6 bear out this concern. In the distribution of consistency scores for the fy configuration, the rejection region for H_0 is barely visible on the far right, indicating that only nearly perfectly consistent data would pass the test. In all of the other figures, the rejection region is too slim even to be plotted.

As it turns out in this case, the exact width of the rejection region is something of a moot point. None of the proposed groupings of sets would pass even if we ignored the problem of multiple inference entirely. As the small dots along the X axis in Fig. 6 show, most of the observed numbers of counterexamples and observed consistencies fall squarely within the range of outcomes that we would expect to see by chance. In fact, two of the four set intersections produce the *modal* number of counterobservations found in the permuted data. In short, the patterns that fsQCA found in the data are anything but improbable. The fact that they were deemed strong enough to merit publication underscores the danger of using fsQCA in the absence of null hypothesis testing.

7 Extension to csQCA and mvQCA

Because crisp-set and multi-value QCA are designed for binary or categorical data, and because they do not allow for counterexamples, the tests described above become much simpler when applied to those procedures. For a Boolean combination of sets to constitute a solution in csQCA and mvQCA, it must be associated with the outcome of interest with certainty: if the combination AbC is a solution, there may be one case that exhibits the pattern AbC or there could be many, but there can be no cases of AbC in which the outcome of interest does not occur.

To find the probability of a false positive result for a given solution, we need to calculate the probability that, if the outcomes were permuted many times, we would find that every case that conforms to the solution also corresponds to a case in which the outcome occurs. We need not run a permutation test to do so, however: if q represents the proportion of cases in which the outcome of interest occurs and s represents the number of cases in which the solution in question occurred, we can simply use the probability mass function of a binomial distribution to calculate the raw p value: $\binom{s}{0}q^0(1-q)^s$, or more simply, $(1-q)^s$. This p value can then be adjusted via the Holm–Bonferroni procedure, as above, to calculate the probability of Type I error given k tests (where k again refers to the number of configurations used by the truth table algorithm).

It is not hard to see that this will be a very difficult test for many csQCA and mvQCA findings to pass, especially those in which the solution in question is associated with a very small number of sets or the outcome in question is very common. This simply means that the test is doing its job. If a

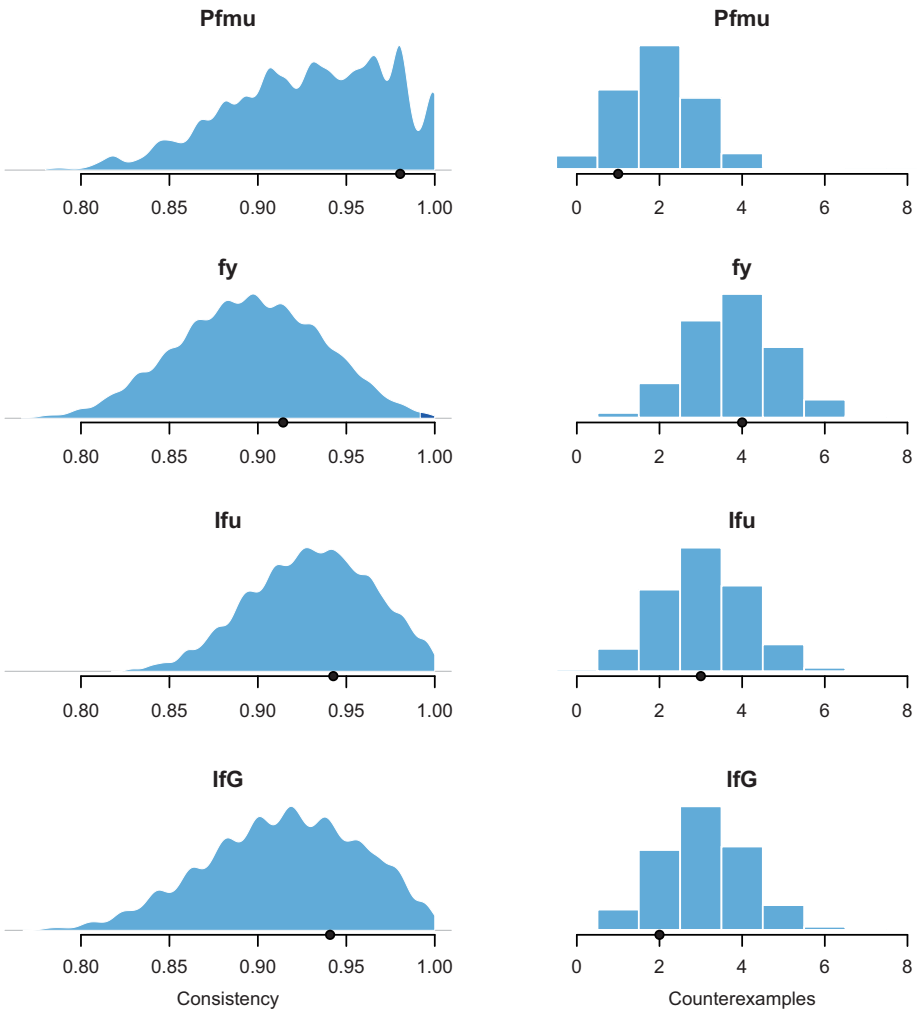

Fig. 6 Permutations of protest cascade data for four Boolean combinations. Dots along X axis indicate observed consistency and number of counterexamples in original study.

solution set is only found in a single case and the outcome of interest occurs in 90% of the observations, the probability of a false positive result is dangerously high, and the test result reflects that fact.

8 Conclusion

Despite the fact that QCA procedures are not designed as statistical techniques, it is quite possible to devise a statistical test that will help us reject, or fail to reject, the null hypothesis that observed QCA relationships are spurious even with a relatively small number of observations. In fsQCA, the most nuanced procedure in the QCA family, a simple permutation test is ideal for this purpose: it is very powerful, it requires no distributional assumptions, and it can be used with non-standard test statistics such as the number of observations above or below the line $X = Y$ or $\frac{\sum \min(X, Y)}{\sum X}$. In csQCA and mvQCA, the probability of a false positive is even simpler to calculate.

Because QCA involves multiple hypothesis tests, researchers must also make an adjustment for the number of combinations considered by the Boolean minimization algorithm in the course of producing the best QCA results. The Holm–Bonferroni adjustment to the tests' p values is a powerful and statistically principled way in which to do so, and it brings to light a heretofore-underappreciated fact: as the number of sets included in a QCA analysis grows, the number of

implied hypothesis tests increases, sometimes dramatically, and the probability of Type I error increases as well. Under such circumstances, the necessary adjustment for multiple inference may create challenges for even the most compelling findings. This is as it should be, because those findings often represent the best to come out of dozens or even hundreds of comparisons.

Conflict of interest statement. None declared.

References

Ansani, Andrea, and Vittorio Daniele. 2012. About a revolution: The economic motivations of the Arab Spring. *International Journal of Development and Conflict* 02(03):1–24.

Benjamini, Yoav, and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 57(1):289–300.

Braumoeller, Bear. 2015. Replication data for: Guarding against false positives in qualitative comparative analysis. Harvard Dataverse, V2 [Version], May 18, 2015. http://dx.doi.org/10.7910/DVN/GY6P9I.

Braumoeller, Bear, and Gary Goertz. 2000. The methodology of necessary conditions. *American Journal of Political Science* 44(4):844–58.

Cronqvist, Lasse, and Dirk Berg-Schlosser. 2009. Multi-value QCA (mvQCA). In *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques*, eds. Benoît Rihoux and Charles C. Ragin, 69–86. Thousand Oaks, CA: Sage Publications, Inc.

Dion, Douglas. 1998. Evidence and inference in the comparative case study. *Comparative Politics* 30(2):127–45.

Doyle, Michael W. 1983a. Kant, liberal legacies, and foreign affairs. *Philosophy and Public Affairs* 12(3):205–35.

——— 1983b. Kant, liberal legacies, and foreign affairs, part 2. *Philosophy and Public Affairs* 12(4):323–53.

Eliason, S. R., and R. Stryker. 2009. Goodness-of-fit tests and descriptive measures in fuzzy-set analysis. *Sociological Methods & Research* 38(1):102–46.

Goertz, Gary, Tony Hak, and Jan Dul. 2012. Ceilings and floors: Where are there no observations? *Sociological Methods & Research* 42(1):3–40.

Good, Phillip I. 2005. *Permutation, parametric and bootstrap tests of hypotheses, Springer Series in Statistics*. 3rd ed. New York: Springer.

Holm, Sture. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2):65–70.

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

Hussain, Muzammil M., and Philip N. Howard. 2013. What best explains successful protest cascades? ICTs and the fuzzy causes of the Arab Spring. *International Studies Review* 15(1):48–66.

Krogslund, C., D. D. Choi, and M. Poertner. 2015. Fuzzy sets on shaky ground: Parameter sensitivity and confirmation bias in fsQCA. *Political Analysis* 23(1):21–41.

LaGraffe, Dan. 2012. The youth bulge in Egypt: An intersection of demographics, security, and the Arab Spring. *Journal of Strategic Security* 5(2):65–80.

Ludbrook, John, and Hugh Dudley. 1998. Why permutation tests are superior to t and F tests in biomedical research. *American Statistician* 52(2):127–32.

Marx, Axel, and Adrian Dusa. 2011. Crisp-set qualitative comparative analysis (csQCA): Contradictions and consistency benchmarks for model specification. *Methodological Innovations Online* 6(2):103–48.

Monroe, Burt, and Suzanne Gold. 2004. A close look at the qualitative comparative analysis (QCA) family of methodologies. Paper presented at the Annual Meeting of the Midwest Political Science Association, Palmer House Hilton, Chicago, IL.

Paul, Christopher, and Colin P. Clarke. 2014. A broad approach to countering the Islamic state. *Washington Post*.

Ragin, Charles 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.

——— 2004. *Between complexity and parsimony: Limited diversity, counterfactual cases, and comparative analysis*. University of California. http://www.sscnet.ucla.edu/soc/soc237/papers/ragin.pdf (accessed on July 10, 2015).

——— 2008. *Redesigning social inquiry: Fuzzy sets and beyond*. Chicago: University of Chicago Press.

Ragin, Charles C., Sarah Ilene Strand, and Claude Rubinson. 2008. User's guide to fuzzy-set/qualitative comparative analysis. Manuscript, University of Arizona. http://www.u.arizona.edu/~cragin/fsQCA/download/fsQCAManual.pdf (accessed March 10, 2015).

Rihoux, Benoît, Priscilla Alamos-Concha, Damien Bol, Axel Marx, and Ilona Rezsöhazy. 2013. From niche to mainstream method? A comprehensive mapping of QCA applications in journal articles from 1984 to 2011. *Political Research Quarterly* 66(1):175–84.

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

———. 2013. Doing justice to logical remainders in QCA: Moving beyond the standard analysis. *Political Research Quarterly* 66(1):211–20.

Thygeson, N. Marcus, Deborah Peikes, and Aparajita Zutshi. 2013. Fuzzy-set qualitative comparative analysis: A configurational comparative method to identify multiple pathways to improve patient-centered medical home models. Technical Report AHRQ, Publication No. 13-0026-EF, Agency for Healthcare Research and Quality, Rockville, MD.

Welkowitz, Joan, Barry H. Cohen, and R. Brooke Lea. 2012. *Introductory statistics for the behavioral sciences*. 7th ed. Hoboken, NJ: Wiley.