A Bootstrapped Robustness Assessment for Qualitative Comparative Analysis*

C. Ben Gibson[†] and Burrel Vann Jr[‡]

Forthcoming in the *Journal of Mathematical Sociology*

Abstract

Qualitative Comparative Analysis (QCA) has been increasingly used in recent years due to its purported construction of a middle path between case-oriented and variable-oriented methods. Despite its popularity, a key element of the method has been criticized for possibly not distinguishing random from real patterns in data, rendering its usefulness questionable. QCA methodologists have suggested certain quantitative thresholds to protect against spurious results. We test the effectiveness of these thresholds using repeated random sampling of data. We find evidence for the effectiveness of these thresholds, but this effectiveness is attenuated by basic properties of the underlying data. Using the intuition of the bootstrap, we develop techniques to determine which QCA thresholds are most appropriate for the input data. This assessment can be used as a hypothesis test for QCA, with an interpretation similar to a p-value.

*We are grateful to Charles Ragin and Carter Butts for their thoughtful comments on earlier versions of this work. We also thank Zack Almquist and the *many* anonymous reviewers who have provided feedback on this article. This research was made possible thanks to The Devil Wears Prada, and by generous funding from the Ford Foundation Pre-Doctoral Fellowship.

†C. Ben Gibson is an Associate Social/Behavioral Scientist at RAND. Direct correspondence to: C. Ben Gibson via Email: bgibson@rand.org

‡Burrel Vann Jr is an Associate Professor in the School of Public Affairs at San Diego State University.

1 Introduction

Qualitative Comparative Analysis (QCA), introduced in Charles Ragin's 1987 book *The Comparative Method*, is a set of techniques designed to be a "middle-path" between quantitative and qualitative analysis (Ragin 1987). At base, QCA provides a set-theoretic approach to social science grounded in Boolean algebra. Within this framework, QCA can identify necessary conditions and multiple combinations or configurations of conditions (known as "recipes", "pathways," or "solutions") for explaining the value of an outcome.

QCA can provide a useful alternative for analyzing complex causation, broadening the reach of current research strategies to integrate a combinatorial, logic-of-sets framework. For example, this multi-method technique enables the user to be "in dialogue" with the results – encouraged to bring in their own case-oriented knowledge to establish causal conditions, examine the results of QCA's "truth table" which displays identified configurations or pathways to an outcome, investigate the categorization of cases, and redefine conditions as needed. As such, QCA's interpretive qualities for have been heralded above the mean-based approach of regression analysis, which may fail to provide robust results for smaller samples (Katz et al. 2005; Seawright 2005; Vis 2012). Ragin and others have since published practical extensions of the QCA method, including the use of 'fuzzy-sets,' the development of software, and many substantive applications (Dusa and Thiem 2014; Caramani 2008; Ragin 2008; Rihoux and Ragin 2009; Schneider and Wagemann 2007; Thiem and Dusa 2013).

Despite recent popularity, QCA is not without its detractors.[1] Specifically, opponents of QCA are skeptical of its ability to successfully identify randomly-drawn data as patternless, which is an important benchmark of any method. Most notable is Lieberson's (2004) assertion that applying QCA to random data would result in spurious

[1]One such criticism is regarding the use of the truth table and subsequent algorithm to determine a minimum configuration for a given outcome. The truth table summarizes agreement on an outcome value for cases with the same configuration of causal conditions; the algorithm combines many configurations in a reduced configuration that results in an outcome.

configurations rather than null configurations. The chief criticism of this randomness problem is that "QCA is less prepared to allow for chance and probabilistic processes" than other methods, and that "procedures do not rule out the possibility that the observations are all a random matter" (Lieberson 2004:13). What is more, researchers question QCA's usefulness on the grounds that 1) it does not incorporate assessments of error commonplace in hypothesis tests (Hug 2013), and 2) the sensitivity of the method to variation to user-selected robustness thresholds remains unknown (Skaaning 2011). Though Braumoeller (2015) used permutation tests to identify randomness in *returned* configurations, a principled test of QCA, such as suggested by (Lieberson 2004) – against *totally* random data – while also varying the user-selected robustness specifications and the structure of data has not been undertaken.

Such a test is useful in two ways. First, it is useful to generally assess the ability of QCA to filter out random patterns, when applying reasonable thresholds for robustness parameters (i.e. configurational N and consistency score). Second, this test can be used for directly calculating the probability of returning random results according to unique features of any data set. This would be useful as an application-specific diagnostic assessment for QCA's "truth table" procedure, and would help protect users against wrongfully spotting patterns via stochastic properties of measurable phenomena. The interpretation of such an assessment would be the probability of returning a random configuration using the selected data and selected robustness specifications. Moreover, such a test could provide recommendations regarding the robustness thresholds necessary to reach a desired level of confidence against a random result.

Below, we assess whether QCA is robust to randomness, generally. We systematically apply QCA to thousands of random data sets, incrementally changing elements of the data structure – sample size and the distribution of variables in the data set, as well as elements under the control of the user – consistency score threshold, configurational N threshold, as well as 'complex' versus 'parsimonious' solutions. We then use logistic regression to determine which of these elements affects the probability of returning a random configuration – a configuration returned from random data.

The primary purpose of this article is to introduce the Bootstrapped Robustness Assessment for Qualitative Comparative Analysis – software (the `braQCA` package in the `R` language) designed to evaluate any QCA result. This is accomplished by generating many random data sets of the same data structure (i.e. sample size and variable distributions) used in an application of QCA, then applying QCA repeatedly at the thresholds under control of the user (i.e consistency score thresholds, configurational N, and parsimonious vs. complex solutions). The result is the probability that a given QCA model would return a random result, given that the result is based upon random data of similar size and distribution. We hope that this method will provide straightforward, easily-interpreted recommendations for researchers who desire unarbitrarily-drawn thresholds of choice.

2 'Probabilistic Processes' and QCA

In a recent *Sociological Methodology* symposium on the methodological merits of QCA, researchers have called attention to, and attempted to address, some of the fundamental problems of QCA (see Lucas and Szatrowski 2014; Vaisey 2014; Ragin 2014), with many evaluating the reliability of QCA under both statistical and epistemological conditions. Seawright (2005, 2014), for example, finds that while QCA shares some of the problems also inherent in regression analysis – with regard to the functional form of the relationship between variables and asserting causation from association – QCA is largely problematic with respect to dealing with missing variables as well as with unobserved configurations. One piece (Collier 2014) highlights several studies that address whether QCA's algorithm returns robust configurations, finding that QCA solutions are highly contingent upon specific values of thresholds chosen by the user (see Hug 2013; Schneider and Wagemann 2012; Krogslund et al. 2015). In this article, we attempt to remedy QCA's "randomness problem" – the extent to which QCA's 'truth table' analysis and algorithm are able to filter randomness in data given the robustness checks currently available.

Although QCA is a broad set of techniques to analyze small-to-medium-n data, the primary criticism of QCA has been its analysis of the 'truth table.' The QCA 'truth table' is a decomposition of data that assesses each combination of causal conditions found in the data, the number of cases within each combination, and the extent to which cases that share these causal conditions exhibit the same value on an outcome. An algorithm is applied to the truth table, combining the information into one or more "causal recipes" or solutions that result in an outcome. Lieberson (2004) imagines a test of his assertion that truth table analysis returns random results: apply QCA to a collected data set versus data where values of the independent and dependent variables are randomly reassigned, keeping the marginal distributions intact. If QCA returns a configuration in both cases, it has a serious problem with being able to distinguish real patterns from random ones.

In a rebuttal, Rihoux and Ragin (2009) argue that such a test would show that random patterns would be filtered out by probabilistic procedures used in any application of QCA. One such procedure is the use of a high *consistency score threshold*: the minimum proportion of cases that must exhibit the same value on an outcome for a given configuration or solution. This threshold is designed to prevent configurations that have high probability of being random from being included in the QCA algorithm. According to Ragin (2014, 1987), a recommended threshold is .85, meaning that 85 percent of the cases with the same specific configuration of causal conditions, must exhibit the same value on the outcome. However, a consistency score threshold may have a differential impact on filtering random configurations depending on the marginal distributions of the conditions used in the QCA model. For example, imagine an application of QCA had an outcome whose cases have a 90 percent probability of being 1. Any combination of categories now has a .9 probability of having an outcome of value 1, with some rate of error. However, if attempting to explain the negation of the outcome, each combination of variables has a .1

probability of having an outcome with value 0. In these situations, the consistency score needed to guard the user against observing a random pattern varies considerably – at base, a .9 is needed in the first case, while the lowest value is needed for predicting the negation. A more direct estimation of randomness – one that takes into account the marginal distribution of the outcome used in the analysis – would therefore be helpful for providing an application-specific recommendation for the consistency score.

A second probabilistic procedure designed to prevent spurious results is the *configurational N threshold*: the minimum number of cases required for a given configuration or solution to be considered in the final result. This prevents the user from making conclusions about a small number of cases, especially about being overconfident about just one case with a unique set of conditions. To prevent such a scenario, the user can set a certain configurational N threshold (usually 2 or 3) to throw out those combinations of conditions that do not have a sufficient number of cases to make conclusions.

Ostensibly, a high consistency score should be sufficient to account for error in causal conditions. As we have argued, these thresholds should have varying rates of success with respect to the marginal probability of conditions present in the analysis. The utility of these procedures to distinguish a random data set from a collected one is an empirical question. Therefore, our first goal is to determine if the current user-selected thresholds are enough to ensure that the results returned are not due to random chance.

This paper addresses two areas of inquiry. First: To what extent does a consistency score and configurational N threshold actually reduce the chance of returning a spurious result? Relatedly, how does this effectiveness differ according to the structure of the data? For example, is a consistency score of .9 effective at all sample sizes? How does the distribution of the outcome affect the usefulness of a high configurational N threshold? Our first set of results demonstrates the general effects of user choice on spuriousness in QCA's 'truth table' analysis, with a special attention given to its variation with data structure. Our second set of questions refers to a practical application of these results. If results are highly dependent upon the structure of data analyzed, an application-specific robustness check would be useful for providing specific recommendations for user choice. The second set of results is a practical application of our procedure that uses a given QCA model, simulates many random data sets from this model, and gives 1) specific recommendations for ensuring against random results and 2) a specific value for the probability that a given application of QCA would return a random result. The latter application has an analogous interpretation as a p-value for a QCA result.

3 Is QCA Robust to Randomness?

There are user-selected thresholds available prior to an analysis to reduce random configurations from being returned. The *consistency score threshold* restricts the analysis to only consider configurations that have a certain proportion of cases that all agree in the direction of the outcome variable. The *configurational N threshold* restricts the analysis to only consider those configurations that have a certain number of cases within them.

Though these are attempts to introduce probabilistic checks for QCA configurations, their use is often flexible, and general recommendations for which thresholds are hard to determine without a principled test of their usefulness. This section assesses the relative importance of each probabilistic check for filtering out random configurations from being returned by QCA.

3.1 Assessing the Robustness of QCA

We employ a straightforward assessment of QCA using simulations. First, we first simulate a random data set. Next, we apply QCA to this random data set, and record whether QCA returned a result at all from random data. If we discover that a result is returned, we know that QCA is returning a spurious result. We systematically vary several variables to determine which elements of data structure (marginal distribution of variables, number of causal conditions included in the model, and sample size) and features of user choice (consistency threshold, configurational N threshold, and complex versus parsimonious solutions) affect the probability of a spurious result.

Each observation in this data set is based on unique data sets. We create these several unique QCA data sets by doing the following:

1. We create unique data sets by varying the number of observations from 10 to 60, by tens ($n = 6$)

2. For each of these data sets, we create a unique set of data sets by varying the number of causal conditions in each, ranging from 2 to 7 causal conditions ($n = 6$)

3. For each, we create a unique set of data sets by varying the probability of the outcome variable being equal to one, ranging from 0.1 to 0.9, by 0.2 ($n = 5$)

4. For each, we create a unique set of data sets by varying the probability that the causal conditions being equal to one, ranging from 0.5 to 0.9, by 0.1 ($n = 5$)

5. For each, we next create a unique set of data sets by varying the value of the consistency score threshold required by QCA, ranging from 0.50 to 1.00, by .01 ($n = 51$)

6. For each, we create a unique set of data sets by varying the value of the configurational N threshold required by QCA, ranging from 1 to 6 ($n = 6$)

7. For each, we create a unique set of data sets by varying whether or not QCA calculates a complex or a parsimonious solution ($n = 2$)

8. For each, we create 5 simulations ($n = 5$)

Therefore, the resulting data set is greater than 2.5 million (2,754,000) observations, each case generated from an iteration of this procedure.

We employ logistic regression on these data, with the dependent variable being a 0-1 outcome of whether or not QCA returned a configuration from random data. The independent variables are the elements of data structure and user choice listed above. The primary question here is, which factors, when altered, independently decrease the probability of returning a random configuration in QCA?

Secondly, we assess whether user choice differentially affects the probability of a result given the structure of the data. For example, does a high consistency score threshold filter out spuriousness across all sample sizes? Does a high configurational N threshold filter spuriousness given all marginal probabilities of causal conditions? This is tested using interaction effects between variables measuring user choice and variables measuring data structure. If interaction effects are substantial, it suggests that additional assessments that take into account data structure need to applied to QCA to ensure robustness.

3.2 Results

Generally, choosing a higher consistency score and higher configurational N in QCA reduce the probability of returning a spurious result. Their effectiveness, however, is dependent upon the basic structure of the data: the distribution of the dependent variable, the number of variables used in the analysis, and the sample size affect how probable a result is spurious. First, we describe the main effects of each of these in turn. Then, we discuss how the structure of the data interacts strongly with user-selected thresholds .

3.2.1 The Effects of Data Structure and User Choice on Spuriousness

Model 1 in Table 1 shows the results of the logistic regression model predicting whether QCA returned a spurious result from simulated random data. Figure 1 displays the predicted effects of Model 1 graphically. When considering thresholds under the control of the user (in the left panel), we see that there is a general decrease in the likelihood of retrieving a spurious result from a QCA model. Specifically, higher configurational N

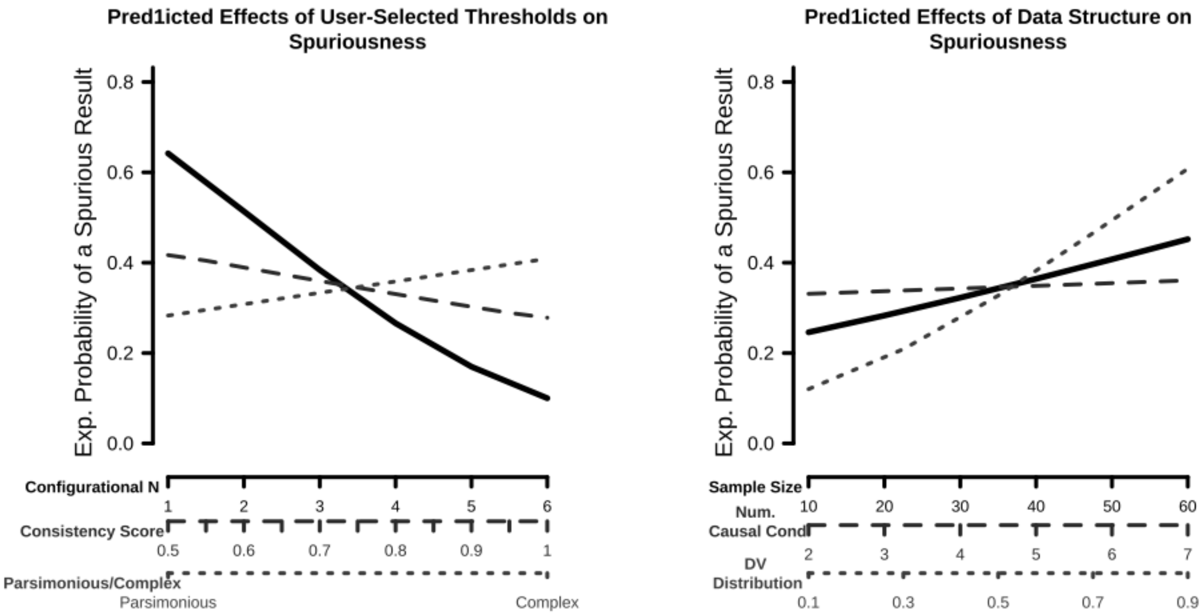

Figure 1: Predicted Effects on the Probability of a Spurious Result in QCA (Crisp Set)

thresholds, and higher consistency scores are associated with lower likelihoods of spuriousness, whereas choosing a complex (rather than a parsimonious) solution is associated with a higher likelihood of spuriousness. In particular, a one-unit increase consistency score is associated with a 2.020-unit decrease in the log odds of a spurious result (an 87 percent decrease). Similarly, a one unit increase in the configurational N is associated with a .733-unit decrease in the log odds of spuriousness (a 52 percent decrease). Importantly, choosing a complex solution is associated with a substantial increase in the likelihood of getting a spurious result. This latter finding is in line with the understanding that, relative to a parsimonious solution, a complex solution is more likely to result in a solution because it considers all possible combinations of conditions. Considering aspects of data structure (as seen in the right panel), we see that higher values for sample size, the number of causal conditions, and the marginal distribution of the dependent variable all increase the likelihood of getting a spurious result.

Table 1: *Effects of User-Selected Thresholds and Data Structure on Spurious QCA Result: Logistic Regression Estimates (N=2,754,000)*

	Model 1	Model 2
Consistency Score Threshold	-2.020^{***}	-2.450^{***}
	(0.011)	(0.053)
Configurational N Threshold	-0.733^{***}	0.545^{***}
	(0.001)	(0.008)
Complex Solution	0.874^{***}	0.731^{***}
	(0.003)	(0.023)
Sample Size	0.029^{***}	-0.054^{***}
	(0.0001)	(0.001)
Number of Causal Conditions	0.042^{***}	0.244^{***}
	(0.001)	(0.006)
Dependent Variable Distribution	4.127^{***}	-5.832^{***}
	(0.007)	(0.046)
Consistency Score Threshold × Configurational N Threshold		-1.308^{***}
		(0.009)
Consistency Score Threshold × Complex Solution		-2.903^{***}
		(0.025)
Consistency Score Threshold × Sample Size		-0.001
		(0.001)
Consistency Score Threshold × Number of Causal Conditions		0.171^{***}
		(0.007)
Consistency Score Threshold × Dependent Variable Distribution		8.446^{***}
		(0.050)
Configurational N Threshold × Complex Solution		0.130^{***}
		(0.002)
Configurational N Threshold × Sample Size		0.010^{***}
		(0.0001)
Configurational N Threshold × Number of Causal Conditions		-0.271^{***}
		(0.001)
Configurational N Threshold × Dependent Variable Distribution		0.490^{***}
		(0.005)
Complex Solution × Sample Size		-0.013^{***}
		(0.0002)
Complex Solution × Number of Causal Conditions		0.111^{***}
		(0.002)
Complex Solution × Dependent Variable Distribution		3.547^{***}
		(0.015)
Sample Size × Number of Causal Conditions		0.011^{***}
		(0.0001)
Sample Size × Dependent Variable Distribution		0.030^{***}
		(0.0004)
Number of Causal Conditions × Dependent Variable Distribution		0.022^{***}
		(0.004)
Intercept	-0.634^{***}	1.371^{***}
	(0.011)	(0.043)
Observations	2,754,000	2,754,000
Log Likelihood	−1,211,425.000	−1,042,433.000
AIC	2,422,863.000	2,084,909.000

Notes: Standard errors in parentheses.
$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$ (two-tailed test).

3.2.2 The Contingent Effects of User Choice on Spuriousness

How effective is user choice in reducing the spuriousness of QCA under different conditions of data structure? For example, does increasing the consistency score threshold effectively reduce the chance of a spurious result when the sample size is high? Model 2 of Table 1 demonstrates how various choices a user can make (e.g. the consistency score, configurational N, and solution type) interact with the various elements of data structure (e.g. the marginal distribution of the dependent variable, number of causal conditions, and sample size) to affect the likelihood of retrieving a spurious result. For ease of interpretation, these interactions are displayed in the figures below.

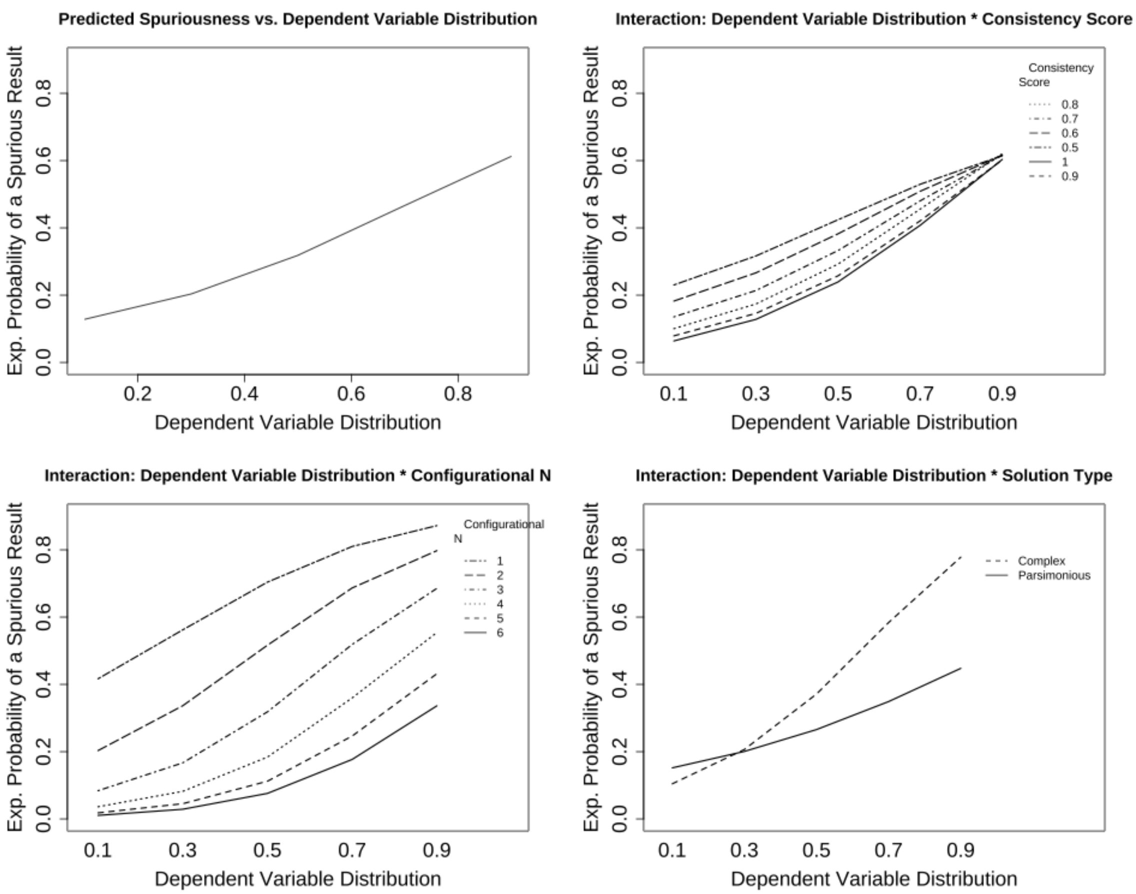

Figure 2: Predicted Interactions between Elements of User Choice and Outcome Distribution

Figure 2 shows the predicted interaction between user choice and the distribution of the dependent variable on the probability of spuriousness. On the y-axis is the expected probability predicted by the logistic regression model; the x-axis shows the dependent variable distribution in the data used to fit the QCA model. The top-left plot shows the predicted spuriousness across outcome distributions, with all other variables held at their means. The top-right shows that the consistency score threshold assists in reducing spuriousness by a great deal at lower distributions; when all cases have a .1 probability of being '1' on the outcome, increasing the consistency score from .5 to 1 decreases the probability of a spurious result. The effectiveness of consistency score threshold on spuriousness decreases as the outcome distribution changes, however. When all cases have a .9 probability of being '1,' the predicted effect of increasing the consistency score is slightly negative. In data sets where cases almost all agree on an outcome, the consistency score may not be the most effective tool to prevent spuriousness. Higher consistency scores in this circumstance will filter out configurations that randomly vary from the .9 baseline probability in the data set, returning a "random" configuration that made the cut while filtering out "random" configurations that predictably varied below the threshold in their incidence of the outcome.

The configurational N threshold substantially impacts the probability of robustness at all levels of the outcome distribution. Unlike the consistency score threshold, there is no change in the direction of the effect; increasing the configurational N threshold decreases the probability of spuriousness at all levels of the outcome distribution. The extent to the effect, however, decreases with an increase in proportion being '1' in the outcome. The probability of spuriousness at a configurational N threshold of 6 increases from .1 to .25 when the outcome distribution of '1s' increases from .1 to .9. This effect is small, however, relative to the effect of consistency score threshold times distribution of the outcome.

As shown in the bottom-right plot in Figure 2, the 'parsimonious' solution is almost always more robust than the complex solution, regardless of the outcome distribution.

Figure 3 shows the interactions between elements of user choice and the number of

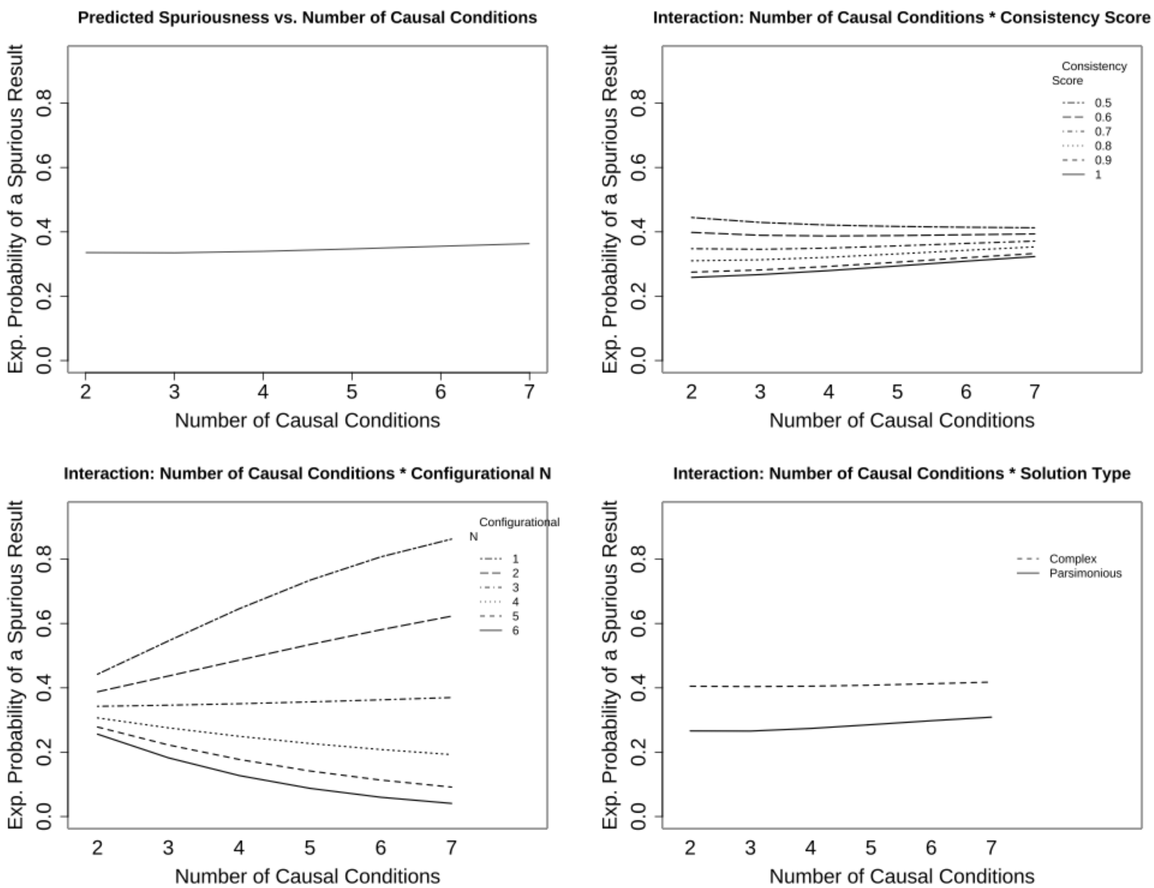

Figure 3: Predicted Interactions between Elements of User Choice and Number of Causal Conditions

causal conditions used in the analysis. Generally, a larger number of causal conditions is predicted to have a higher probability of spurious results. As seen in the top-right plot, as the number of causal conditions increases, the effect of the consistency score on spuriousness also (generally) increases. An interesting trend occurs when interacting configurational N threshold with the number of causal conditions – the effect of configurational N on spuriousness is similar at lower numbers of causal conditions, but different at higher numbers of conditions. Specifically, at higher numbers of causal conditions, a lower configurational N is related to a higher probability of spuriousness whereas a higher configurational N is related to a lower probability of spuriousness.

Ostensibly, this is due to fewer configurations being included in the truth table analysis when the number of possible configurations increases. Additionally, at all levels of causal conditions, the parsimonious solution is better for reducing the probability of spuriousness.

As Figure 4 shows, the effect of user choice varies the least when interacted with sample size. Table 1, again, shows effects for these interactions, the effect size is small compared to the interactions with outcome distribution and number of causal conditions. The top-left plot in this figure shows that the expected probability of spuriousness increases at higher sample sizes. Moreover, as seen in the top-right plot, this effect is unrelated to consistency score thresholds. Put another way, although higher consistency scores can reduce the probability of spuriousness, this effect does not differ at various sample sizes (the probability of spuriousness hangs around .35 for all numbers of causal conditions and all consistency scores examined). The bottom-left plot shows that the effect of sample size on spuriousness varies by configurational N. Higher configurational N thresholds can reduce the probability of spuriousness at all sample sizes, but this effect is stronger for larger samples. Finally, parsimonious solutions are more effective for reducing the probability of spuriousness for larger samples.

3.3 Discussion: QCA Robustness

The results here show that the probabilistic checks set by the user in a QCA truth table analysis are effective in reducing the probability of a spurious result. However, they also show that the effectiveness of user-set parameters to ensure robust results vary according to the structure of the data. In some cases, consistency score alone will not effectively filter out random patterns, especially when the outcome is distributed such that all configurations have a high probability of having the same outcome value. In some cases, however, a high consistency score is not needed to ensure a robust result. The interaction effects above show that elements of user choice differed in their effectiveness at differing levels of number of causal conditions used. A high configurational N threshold was more

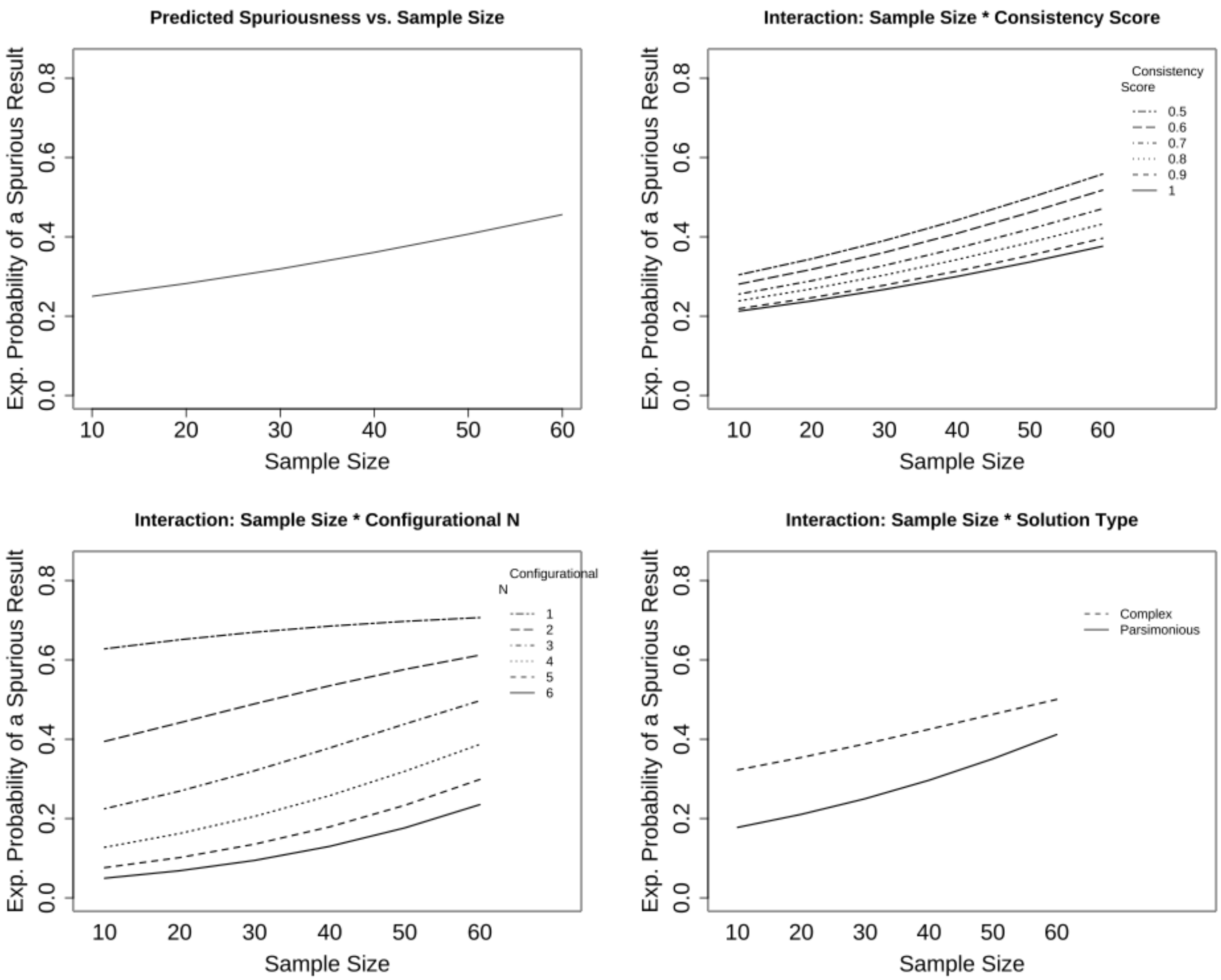

Figure 4: Predicted Interactions between Elements of User Choice and Sample Size

effective at higher numbers of causal conditions, while high consistency score was more effective when using fewer causal conditions, while . The difference between complex and parsimonious solutions was negligible at lower number of causal conditions, but represented a nearly four-fold increase in spuriousness – from 25 percent chance to 85 percent chance – when using seven causal conditions. Other interactions, while significant, were small in scale.

These results show that the choices users make when conducting a truth table analysis of QCA data have a differential effect upon spuriousness according to variations of basic features of the data. The large variation of these effects, as well their complexity, justifies the need for a more straightforward approach to probabilistic assessments for QCA's truth table analysis. While it would be helpful to advise researchers on general practices for choosing a consistency score, a configurational N, and whether to use complex or parsimonious solution, the large variations in their effects according to features of the data used prevent the authors from doing so.

To clarify interpretation, Table A1 displays the predicted probability of spuriousness for various combinations of data structure and user-set thresholds. For example, when a data set with 30 cases has an outcome that is 50 percent '1s' and 50 percent '0s', and the truth table analysis has a consistency score threshold of .9, a configurational N threshold is 4, and five causal conditions, the resulting configuration is predicted to have a 3 percent chance of being spurious. When a data set of 20 has a distribution of 90 percent '1s' and 10 percent '0s', a truth table analysis predicting an outcome of '1', a consistency score threshold of .9, four causal conditions, and a configurational N threshold of three, the probability of spuriousness is 40 percent. If we were to change the configurational N threshold to 4, the probability decreases to 14 percent. All three situations could easily be seen in a QCA truth table analysis, but offer vastly different probabilities that the resulting configuration is spurious.

In many cases, the robustness of a QCA truth table analysis can reach the .05 level conventionally required for other methods, while using relatively standard thresholds for QCA analysis. For example, a configurational N threshold of 4 is generally sufficient to ensure a 'p-value' of $< .05$ if the distribution of 1s and 0s is even. When the distribution is mostly 1s, a configurational N threshold of 5 is generally sufficient to achieve the same level of robustness. Generally, the consistency score threshold is only somewhat effective in practice, and more so when the distribution of 1s and 0s is even (or more 0s are present than 1s) when trying to predict the presence of the outcome.

In the section below, we present a method for providing a model-by-model estimate of spuriousness. This method has two functions: 1) to estimate the 'confidence level' of an existing QCA model; and 2) to provide a reasonable recommendation for setting the consistency score and configurational N thresholds to achieve a desired 'confidence level.' We first generally describe the method; we then give an example of its application

4 The Bootstrapped Robustness Assessment for QCA (`braQCA`)

The Bootstrapped Robustness Assessment for QCA (`braQCA`) is software developed in the R programming language, and is a procedural check of a QCA result which takes into account data structure (e.g. marginal distribution of variables) and user-selected thresholds to provide an estimate of the probability of spuriousness for a given result. Above, we demonstrated that user choices, which are designed to ensure robust analysis of a truth table, require vastly different thresholds vis-à-vis the structure of the data.

To build a robustness assessment while taking into account data structure, we draw a random data set using the same data structure as a QCA result. This includes using the same 1) number of causal conditions as the observed data set, 2) marginal distributions of the causal conditions and dependent variable as the observed data, and 3) the sample size. For this new random data set, we run a QCA model matching both 4) the consistency score threshold and 5) the configurational N threshold set by the user. This process is repeated thousands of times, and we take the simple proportion of times QCA returned *any* configuration from the random data sets, given the parameters of the observed data set. This proportion can be interpreted as the probability that the observed solution (e.g. combinations of configurations) returned in the QCA analysis is due to chance. The interpretation of this proportion is similar to a p-value, such that lower values indicate that the observed solution is unlikely due to chance.

By randomly sampling data over thousands of iterations, we are observing the probability of returning a random configuration. When a 95 percent confidence interval of the mean or proportion is calculated on a "random" variable, we can conclude with 95 percent confidence that the interval covers the true value of that mean or proportion. Our strategy, thus, identifies the probability that a given application of QCA, with the exact data structure and parameters of user choice, would, with some level of confidence, return a result given completely random data. If the value (and the surrounding confidence interval) is high, the user should be cautious about making claims from their observed data. If this value is low, however, the user can conclude with confidence that the result is unlikely due to random chance, and is robust to a direct comparison with random data.

Recent work has attempted to rectify the randomness problem in QCA. A notable example is Braumoeller's (2015) use of permutation tests to investigate the likelihood of false positive solutions in a QCA model. Importantly, the author argues that consistency score (the percent of cases with a configuration that also have the outcome) and the number of counterexamples (e.g. the number of cases with a given configuration but absent the outcome) are the most critical measures for understanding the efficacy of QCA in demonstrating causal relationships. For Braumoeller (2015), the relevant null hypothesis is that, by chance, the user would have observed either 1) an equal or greater number of counterexamples, or relatedly, 2) an equal or lower consistency score. Thus, such a test is principally concerned with sufficiency versus necessity – whether or not scores on (or, for crisp sets, presence of) a given configuration also result in presence of the outcome.

The focus on the strength of equality between configuration and outcome is admirable. However, contrary to the claim that investigating QCA's problem of retrieving solutions from random data misses the point of QCA (see Braumoeller 2015), we argue that such an investigation is an appropriate *initial* test of QCA's robustness in terms of its ability to detect patterns in data, and not when given random data. While it is important to investigate the likelihood of an observed configuration relative to the distribution of possible configurations, our method makes no assumptions about specific configurations returned by the method, only that any configuration would be returned at all given random data random data of similar variable distributions and user-selected thresholds. As such, our assessment serves as a more general test of QCA, that can be used in conjunction with Braumoeller's (2015) test – that is, the Bootstrapped Robustness Assessment can, first, be used to determine if QCA solutions were, overall, the result of randomness, followed by Braumoeller's (2015) test of false positives to determine the likelihood of each individual configuration to appear by chance.

4.1 Package `braQCA`'s Process for Analyzing Robustness

This section outlines the step-by-step procedure of two methods for determining the probability that a given QCA application returns a spurious result, both available in the `braQCA` package. The first is the Bootstrapped Assessment for QCA (function `baQCA`), which can be applied using the following steps:

1. 'Fit' a QCA model with v causal conditions and n number of cases.

2. Simulate s random data sets, each with v causal conditions and size n. For each data set, sample (with replacement) both the outcome variable and the causal conditions,[2] ensuring that, for each simulated data set, each causal condition and outcome maintain the same distributional properties as those in the observed data. [3]

3. For each simulated data set, apply QCA using the thresholds selected by the user the observed model (configurational N and consistency score, etc.). Record whether QCA returned a spurious result – a configuration or solution from the simulated data.

[2]This follows the test described by Lieberson (2004), but users have the opportunity to sample both the outcome and the causal conditions or the outcome only.

[3]Based on convergence diagnostics from Gelman and Rubin (1992), we selected 2000 simulations. Researchers employing Monte Carlo simulations suggest that the findings from simulations tends to stabilize between 500 and 1500 simulations (Bukaçi et al. 2016). In practice, the number of simulations chosen depends on the structure and parameters of the data being simulated (Davidson and MacKinnon 2000), as well as the desired level of the test (e.g. α), but reaching the upper limit for simulations may be computationally-prohibitive for the user (Hall 1986). As such, we choose the start value of 2000 simulations, but the user has the ability to choose any number of simulations.

4. Take the simple proportion of times the QCA returned a configuration: $\frac{R}{s}$, where R = the number of times a QCA model returned a result from a randomly generated data set and s = the number of simulations used.

5. To measure uncertainty, we calculate bootstrapped standard errors (Efron and Tibshirani 1994) and take 95 percent quantiles of $\frac{R}{s}$ for our confidence interval.

The resulting value is the probability that QCA application would return a random result given the user's data, along with a 95 percent confidence interval.

A useful, related tool is the Bootstrapped Recommendation (`brQCA`) for a consistency score and configurational N threshold given the data, without having fit a QCA model at all. This recommendation can be applied using the following steps:

1. Simulate s random data sets, each with v causal conditions and size n. For each data set, sample (with replacement) both the outcome variable and the causal conditions, matching the distributions of the observed data

2. Apply QCA to all the generated data sets, but systematically vary parameters of user choice (configurational N, consistency score, etc.). Record whether each QCA returned a spurious result.

3. Apply a logistic regression model to the results, using the configurational N and consistency score thresholds as predictors.

4. Use this model to provide fitted values, calculating the minimum consistency score needed at every configurational N threshold to achieve a desired level of the test (e.g. α). The resulting table, therefore, provides a list of consistency scores and configurational N values and their related probabilities – the probability that, if using a given consistency score and with a given configurational N, the solutions returned in the QCA model is the result of chance. We also use the standard errors provided in the regression model to provide a confidence interval around each recommendation.

In practice, this method functionally eliminates the combinations of variables that do not have enough 'power' to be included in the final analysis. For example, it will eliminate from consideration the combinations of variables that contain fewer than n cases, which is standard procedure in QCA truth table analysis. This method simply provides a principled basis for which to choose those thresholds.[4]

Below, we provide a case study where reasonable thresholds were not quite enough to ensure good results, and where `baQCA` could be of service to suggest specific thresholds to ensure robustness.

4.2 Qualitative Comparative Analysis of Tea Party Rallies in Florida

In an application of this technique, we use a subset of data constructed by McVeigh et al. (2014) and Vann (2018) as part of a project on Tea Party organizations in U.S. counties. The data set includes several county-level measures, including demographic measures from the American Community Survey (ACS) 2005-2009 (U.S. Census Bureau 2015), measures of religious adherence from the Association of Religion Data Archives (ARDA) 2001, 2008 Presidential election measures from Congressional Quarterly's *America Votes*, and the number of Tea Party organizations between 2009 and 2010 from the Institute for Research & Education on Human Rights (Institute for Research & Education on Human Rights 2011). We extend the data set to include a new outcome variable: the number of rallies in each county between 2009 and 2010, also from IREHR.

We restrict the data set to counties in Florida for two reasons based on our own case knowledge (Ragin 2008). First, we choose Florida counties because, with the exception of California, all other states had fewer numbers of organizations. Second, and substantively important for the choice of causal conditions, we choose Florida due to the perceived impact the Tea Party movement had in the 2010 midterm election (Vann 2018; Miller and

[4] Like in any QCA model, users should take care to observe which cases are being eliminated, and consider how this would change the results if the thresholds were lowered.

Walling 2012). Restricting the data to counties in Florida leaves us with a middle-N data set of 67 cases for analysis (see Figure 5). Our analysis addresses the multiple causal pathways that lead to the occurrence of one or more Tea Party rallies in a Florida county.

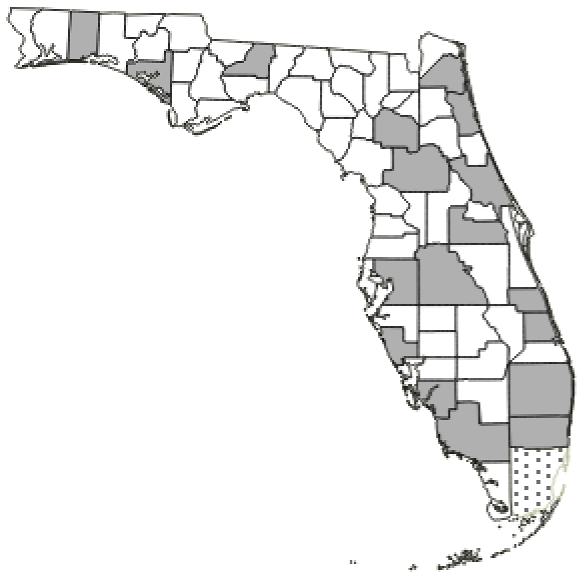

Figure 5: Tea Party Rallies in Florida Counties.
Note: Grey indicates at least one rally in county.

Because Qualitative Comparative Analysis arguments are combinational and often overlapping, a user's causal conditions require a deep knowledge of cases in the data set for adequate placement within particular sets of causal conditions (Ragin 2008). For example, to fully belong to a crisp outcome set, a Florida county must meet or exceed a minimum criterion. Therefore, it is necessary to both establish adequate causal conditions and inclusionary criteria for each condition. As discussed above, some have argued that because these criteria are based on user choice, criteria are biased (Lieberson 2004), which can lead some to believe that a user has cherry-picked their analyses. To combat this assumption, and for the sake of clarity, we employ a simple inclusionary-exclusionary criterion for membership in a causal condition (outlined below) for the analysis. It is important to note that for many causal conditions, we dichotomize on the mean. Because QCA is designed to allow the user to be in dialogue between results and the cases and to recalibrate, it is generally considered bad practice to create inclusionary-exclusionary criteria in this way. We do so as a pedagogical exercise for the bootstrapped robustness assessment for QCA and not as a manner with which to employ QCA.

We briefly discuss research on the Tea Party movement, which provides the basis for calibrating QCA causal conditions for the analysis of Tea Party rallies. Researchers find that while most Tea Party organizations were concentrated in conservative partisan environments (McVeigh et al. 2014; Skocpol and Williamson 2012), much of their on-the-ground rally activity took place in heavily populated, often left-leaning locales (Skocpol and Williamson 2012; Zernike 2010). Similar work has demonstrated the importance of educational background on support for the Tea Party (McVeigh et al. 2014; Skocpol and Williamson 2012), finding that supporters of the Tea Party movement are highly educated, and that Tea Party organizations were more likely to be established in U.S. counties characterized by a predominance of college graduates. Although supporters of the Tea Party tended to be relatively impervious to the economic recession of 2008 (Skocpol and Williamson 2012; Parker and Barreto 2013), many of the movement's grievances related to dismay about unemployment and the expanding reach of federal government (Skocpol and Williamson 2012; McVeigh et al. 2014; Parker and Barreto 2013). Scholars show that while many Tea Party supporters were Protestants, these were not of the Evangelical bent as depicted in the media (Zernike 2010; Skocpol and Williamson 2012; McVeigh et al. 2014). Finally, Parker and Barreto (2013) argue that support for the movement derived from racial backlash against the nation's first Black President, and the perception that he would create policies that favor Black Americans. For this reason, where Black individuals pose a perceived economic threat by virtue of their presence, the size of the Black population could encourage Tea Party activity.

This brief summary of research on the Tea Party provides insight into creating causal arguments about the presence of Tea Party rallies as a test of QCA and the bootstrapped robustness assessment for QCA. Importantly, because the analysis here employs crisp-set

QCA, each causal condition is coded as either one or zero. As previously mentioned, the outcome variable is Tea Party RALLIES. Full placement in the outcome set (1) requires that a county has at least one rally. In sum, there are 19 cases that have the outcome. Given the differential relationship between Republican partisan contexts and the presence organizations and the occurrence of rallies, we include the measure REPUBLICAN. This is coded as one if, during the 2008 Presidential Election, the Republican candidate received a majority of the votes in the Florida county. Based on the above discussion, we expect the absence (negation) of Republican context as an important component of a causal pathway to Tea Party rallies. We also include four causal conditions in which full inclusion is defined in a straightforward manner: full membership in the causal set (1) is determined by whether a value for a particular case falls at or above the mean for that variable.

First, we include a measure of COLLEGE educated, the percentage of people in the county (aged 25 or above) who hold a Bachelor's degree. Second, we include a measure of UNEMPLOYMENT, measured as the percent of the county population which is unemployed. With regard to membership in the college educated set, we expect that the presence of a college educated population is an important component of the pathway to rallies. However, given that many Tea Party supporters were not actually unemployed (although the movement's rhetoric says otherwise), we expect that the absence of a high unemployed population is an important part of explaining Tea Party rallies. Third, we include a measure for the size of the BLACK population, measured as the percentage of African-Americans in the county. Fourth, we include a measure for the size of the EVANGELICAL population, measure as the percentage of Evangelical adherents in the Florida county.

In QCA, the logical representation of the presence of a causal condition is indicated by the variable name in all upper-case letters whereas negation is represented by all lower-case letters. Combinations of conditions (e.g. complex combinations of variables) in a pathway or recipe to an outcome are expressed as a string of variable names delineated by an asterisk, representing the logical operator "AND." If multiple pathways exist, each pathway is delineated by a plus symbol, the logical operator for "OR." Therefore, our main expectation, expressed in QCA notation, is:

republican * COLLEGE * unemployment * BLACK * evangelical

In this analysis, there are a total of 32 possible pathways to the outcome, based on the five (K) causal conditions ($2^K = 2^5 = 32$).

We begin by calculating the causal solutions for our Tea Party rally data (which are available in the braQCA package). The data are shown below.

```
> qca.data <- rallies[,8:13]

> head(qca.data)

                   P R C U E B

Bay County         1 1 1 0 1 0

Bradford County    0 1 0 0 1 1

Taylor County      0 1 0 1 1 1

Indian River County 1 1 1 1 0 0

Hillsborough County 1 0 1 0 0 1

Brevard County     0 1 1 0 0 0
```

For these data, `P` represents the presence of at least one Tea Party Rally (or Protest) in a county, `R` represents that the Republican candidate for President in 2008 (John McCain) secured the majority of the vote, `C` represents that the percent of residents in the county with a Bachelor's degree was above the statewide average, `U` represents higher than average percent unemployment in the county, `E` represents higher than average percent of county residents who identified as Evangelical, and `B` represents higher than average percent of Black residents in the county.

Next, we calculate a truth table (`truthTable` from the `QCA` package) with `P` as the outcome, with a sufficiency inclusion/minimum consistency score of .85 (the minimum proportion of cases explained by a causal configuration, `incl.cut1=0.85`), a configurational N threshold of 1.0 (the minimum number of cases required for a particular configuration, (`n.cut=1`), and return the results without the list of cases shown (`show.cases=FALSE`).

```
> truth <- truthTable(qca.data,outcome="P",
+   sort.by="incl", incl.cut1=0.85, n.cut=1, show.cases=FALSE)
> truth

  OUT: output value
    n: number of cases in configuration
 incl: sufficiency inclusion score
  PRI: proportional reduction in inconsistency


     R  C  U  E  B    OUT    n   incl  PRI
 6   0  0  1  0  1     1     1   1.000 1.000
10   0  1  0  0  1     1     6   1.000 1.000
12   0  1  0  1  1     1     1   1.000 1.000
28   1  1  0  1  1     1     1   1.000 1.000
27   1  1  0  1  0     0     4   0.500 0.500
29   1  1  1  0  0     0     2   0.500 0.500
25   1  1  0  0  0     0     10  0.400 0.400
 9   0  1  0  0  0     0     4   0.250 0.250
21   1  0  1  0  0     0     4   0.250 0.250
19   1  0  0  1  0     0     9   0.111 0.111
 5   0  0  1  0  0     0     1   0.000 0.000
 8   0  0  1  1  1     0     2   0.000 0.000
17   1  0  0  0  0     0     1   0.000 0.000
20   1  0  0  1  1     0     2   0.000 0.000
23   1  0  1  1  0     0     8   0.000 0.000
```

```
24   1  0  1  1  1     0     8    0.000 0.000
31   1  1  1  1  0     0     2    0.000 0.000
32   1  1  1  1  1     0     1    0.000 0.000
```

Next, we minimize the truth table into causal solutions (`minimize` from the `QCA` package) and return the results without the list of cases shown (`show.cases=FALSE`).

```
> model <- minimize(truth, details=TRUE, show.cases=FALSE)
> model

M1: ~R*C*~U*B + C*~U*E*B + ~R*~C*U*~E*B -> P

                 inclS   PRI   covS   covU
-------------------------------------------
1      ~R*C*~U*B  1.000  1.000  0.368  0.316
2       C*~U*E*B  1.000  1.000  0.105  0.053
3   ~R*~C*U*~E*B  1.000  1.000  0.053  0.053
-------------------------------------------
              M1  1.000  1.000  0.474
```

These results, reproduced in Table 2, indicate that only three configurations or combinations explain the presence of Tea Party rallies in Florida counties. Overall coverage for the three configurations is 47.4 percent, with only BLACK appearing in each. The presence or absence of REPUBLICAN, COLLEGE, UNEMPLOYMENT, and EVANGELICAL appear in various combinations, known as "insufficient but necessary components of causal combinations that are unnecessary but sufficient for the outcome" (Mackie 1980; Ragin 1987).

Table 2: QCA Results for Florida Tea Party Rallies

Solutions	Consistency	Coverage
COLLEGE * unemployment * BLACK * EVANGELICAL	100.0%	10.5%
republican * COLLEGE * unemployment * BLACK	100.0%	36.8%
republican * college * UNEMPLOYMENT * BLACK * evangelical	100.0%	5.3%

4.3 Applying `braQCA`

As previously mentioned, using a high consistency score or altering the minimum number of cases for a solution is not enough to ensure the final configurations are non-spurious. To assess spuriousness in the current set of solutions, we apply the `baQCA` method, which yields a probability of randomness score.

We apply the `baQCA` method to the minimized model, which applies the bootstrapped assessment using 2000 simulations (`sim=2000`).

```
> set.seed(1738)
> baQCA(model, sim=2000)

$Probability
[1] 0.9430


$‘95% Confidence Interval’
    2.5%    97.5%
0.9345 0.9520
```

Application of the `baQCA` method shows that causal configurations have a high probability of being random. In fact, of the simulated data sets, just under a 95 percent yielded a random result (p = .943, CI_{95} = .9345, .9520).

Given that our solutions exhibited a high probability of being random, we rely on a second function in the `braQCA` package – one which returns alternative configurational N and consistency threshold recommendations that would yield a non-spurious set of solutions, with significance values.

We apply the `brQCA` method to the data frame, which calculates significance levels for a potential model (from the data frame `qca.data`) with `P` as the outcome, and with a selected minimum configurational N (`ncut=4`), using 100 simulations (`sim=100`).

```
> set.seed(1738)
> brQCA(qca.data, outcome="P", ncut=4, sim=100)

$‘ncut=4’
       sig     solution consistency.score.rec lower.CI upper.CI
1  p < .10 parsimonious                  0.77     0.76     0.78
2               complex                  0.77     0.76     0.78
3  p < .05 parsimonious                  0.84     0.83     0.85
4               complex                  0.84     0.83     0.85
5  p < .01 parsimonious                    NA     0.99       NA
6               complex                    NA     0.99       NA
7 p < .001 parsimonious                    NA       NA       NA
8               complex                    NA       NA       NA
```

The output is also depicted in graphical form below. As shown and described above, to combat spuriousness in QCA, the user *should* select a high consistency threshold. In addition, the researcher should use their knowledge of the solutions and causal conditions to select an appropriate configurational N threshold. As shown in Figure 6, which depicts "typical" levels of significance (e.g. lower probabilities that the causal solutions result from randomness), in all graphs except the bottom-right ($p < .001$), all suggested consistency thresholds around .90 and configurational N thresholds that range from three cases (for higher probabilities of randomness/lower levels of significance) to five (for lower probabilities of randomness/higher levels of significance). Here, we select the recommendation for a configurational N threshold of 4, which is the smallest configurational N threshold for significance levels less than or equal to .05. While the recommended consistency score threshold for a significance level of .05 is .84, based on the figures below, we selected a more stringent threshold of .90. Importantly, the results are similar when using configurational N thresholds of three, four, or five. While it is possible to decrease the consistency score threshold and the configurational N threshold to yield additional configurations, these solutions would only cover one or two cases. Building solutions around one or two cases, instead of a medium-n set of cases, introduces randomness into the set of causal solutions and does not aid a user in developing substantive theory about the conditions that produce an outcome.

The selection of a minimum of four cases per solution and 90 percent consistency results in one solution. This result is shown in Table 3 below. We can see that the result somewhat confirms expectations from the as literature outlined above. Tea Party rallies were present in counties that were non-Republican, and had larger populations of college graduates and black residents, in combination with low levels of both unemployment and Evangelical adherents. This result is non-spurious at the .05 level.

Table 3: QCA Results for Florida Tea Party Rallies, After Applying `brQCA`

Solutions	Consistency	Coverage
republican * COLLEGE * unemployment * BLACK * evangelical	100.0%	31.6%

While the overall solution coverage decreased after relying on threshold recommendations from `brQCA`, the subsequent set of solutions were less prone to randomness (checked by reapplying the `baQCA` method). The more robust configuration is a

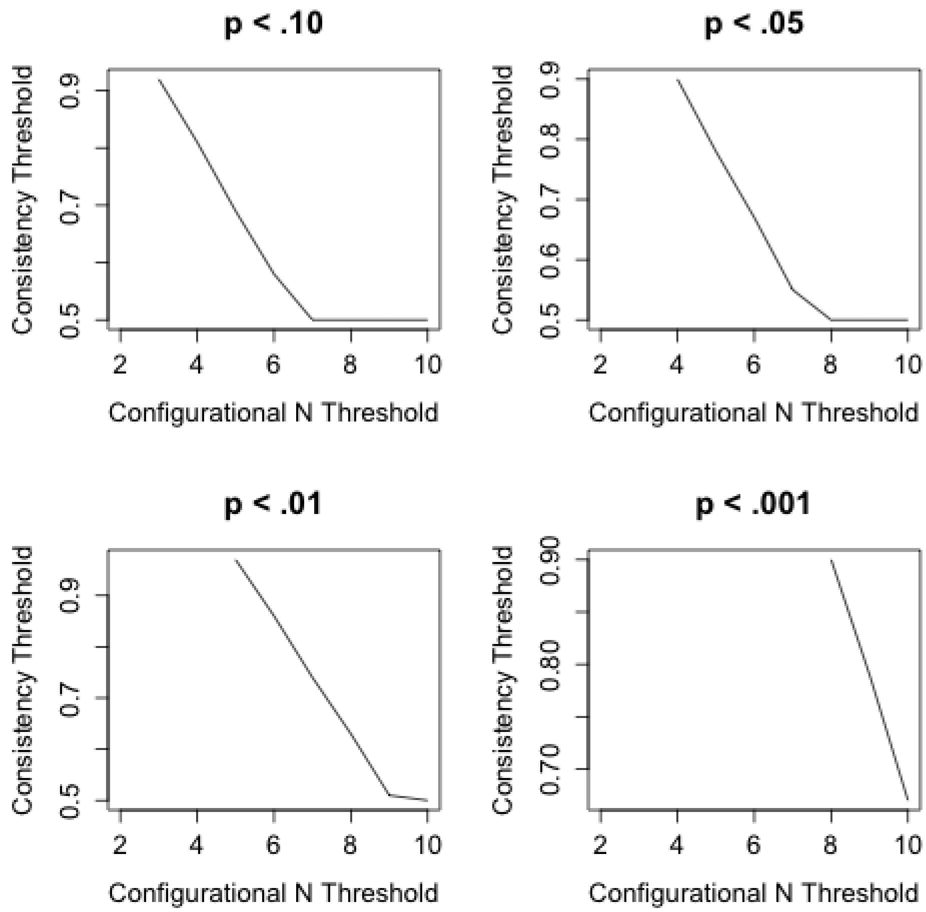

Figure 6: Suggested Consistency and Configurational N Thresholds by Desired Significance Level, per `brQCA`

subset of the less robust configuration, suggesting a simpler solution that excludes from analysis the rows of the truth table that have high probability of spuriousness. Though it has been suggested to remove rows of the truth table with low coverage in general, this test provides a systematic way of doing so. As exhibited in Table 4, applying the thresholds recommended by the `brQCA` method improves our solutions. We are thus highly confident that the result returned by the QCA analysis is not due to randomness.

Table 4: `baQCA` Results for Florida Tea Party Rallies, Difference by `brQCA` Application

Solutions	Probability of Randomness	95% Confidence Interval	
Solution Set, before `brQCA`	0.9430	0.9345	0.9520
Solution Set, after `brQCA`	0.0185	0.0135	0.0235

4.4 Discussion: `baQCA` and `brQCA` Results

The case above shows that a high consistency score threshold alone was not sufficient to prevent spuriousness from taking place. By increasing the configurational N threshold, we are able to greatly reduce the probability of spuriousness of the QCA result. In this case, the original consistency score was too conservative; the same level of significant is achieved by using a .9 consistency score threshold rather than 1. By estimating the probability of randomness via a direct comparison with random data, we are able to 1) estimate what thresholds are reasonable to achieve certain significance levels and 2) determine the probability of randomness of a QCA result.

5 Conclusion and General Recommendations

Though the Bootstrapped Assessment is a quantitative assessment of spuriousness, should be used to better inform the user about the conclusions drawn from the data rather than used as a black-and-white threshold of significance. This is to notify the user of whether the final result has a high probability of including spurious configurations. If the method concludes that a QCA result is not robust to randomness, the researcher may need draw on strong, case-oriented knowledge to clearly articulate how the QCA results hold up and are not due to random chance, based on their own case-oriented knowledge and analysis. That said, a configurational N threshold of 4-5 is generally sufficient for classical threshold of ‘significance.’ Increasing thresholds to improve robustness will inevitably result in fewer cases being considered, which may be of some use to final configurations.[5]

Some practical recommendations for a principled approach to QCA include:

1. Ensuring variability among the cases – so as to not select only those cases that confirm the user's suspicion, theory, or argument

2. Proceed with a typical application of a QCA model, using common thresholds (e.g. configurational N of 4, consistency score of 0.85)

3. Investigate the results from the truth table, asking whether the solutions make sense for the cases at hand

4. Provided the results make logical sense, proceed by assessing the robustness of the findings using the `baQCA` method

5. Once the level of spuriousness in the QCA model is determined, investigate whether or not the solutions need to be improved. If so, employ the `brQCA` method to improve the solutions and limit spuriousness

6. Read the results of the recommendations of the `brQCA` method, and decide a desired level of "significance" given the provided pseudo-p-value

7. Re-run the initial QCA model, applying the configurational N and consistency score recommendations provided in `brQCA` for a given or desired pseudo-p-value

8. Once the new QCA model is complete, re-assess randomness in the model by applying the `baQCA` method

9. Confirm that the new QCA model, with updated configurational N and consistency score thresholds is more robust to randomness than the initial QCA model.

[5]In situations where a lack of cases substantially reduces our ability to solve research questions, we have included in the software package a function that pulls out configurations excluded from analysis by nature of trying to improve robustness. When two models, say Model 1 with configurational N threshold of 2 and Model 2 with configurational threshold of 4, one can run QCAdiff(model1,model2) to determine which configurations were excluded. If the user has a compelling reason to reduce the robustness threshold in favor of a researcher-provided explanation, we welcome it.

Though QCA has been heavily criticized for not being robust to randomness, this simple assessment could help provide a systematic test of QCA against randomness. As we have shown, QCA can be robust to randomness using existing checks, but the ability of these checks to sufficiently protect against spuriousness differs according to data structure. We present a method that measures the confidence that a given QCA result is due to random chance, which provides information above and beyond the current robustness checks. Ideally, the Bootstrapped Robustness Assessment could be used as a standard assessment for QCA that provides precise recommendations for user choice.

References

Braumoeller, Bear F. 2015. "Guarding Against False Positives in Qualitative Comparative Analysis." *Political Analysis* Pp. 1–26.

Bukaçi, E, Th Korini, E Periku, S Allkja, and P Sheperi. 2016. "Number of Iterations Needed in Monte Carlo Simulation using Reliability Analysis for Tunnel Supports." *6*.

Caramani, Daniele. 2008. *Introduction to the Comparative Method With Boolean Algebra*. Thousand Oaks, CA: Sage.

Collier, David. 2014. "Comment: QCA Should Set Aside the Algorithms." *Sociological Methodology 44*:122–126.

Davidson, Russell and James G. MacKinnon. 2000. "Bootstrap Tests: How Many Bootstraps?" *Econometric Reviews 19*:55–68.

Dusa, Adrian and Alrik Thiem. 2014. "QCA: Qualitative Comparative Analysis." R Package Version 1.1-4.

Efron, Bradley and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press.

Gelman, Andrew and Donald B. Rubin. 1992. "Inference from Iterative Simulation using Multiple Sequences." *Statistical Science 7*:457–511.

Hall, Peter. 1986. "On the Number of Bootstrap Simulations Required to Construct a Confidence Interval." *14*:1453–1462.

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

Institute for Research & Education on Human Rights. 2011. *Tea Party Nationalism*. Data retrieved from (http://www.teapartynationalism.com).

Katz, Aaron, Matthias vom Hau, and James Mahoney. 2005. "Explaining the Great Reversal in Spanish America: Fuzzy-Set Analysis Versus Regression Analysis." *Sociological Methods & Research 33*:539–573.

Krogslund, Chris, Donghyun Danny Choi, and Mathias Poertner. 2015. "Fuzzy Sets on Shaky Ground: Parametric and Specification Sensitivity in fsQCA." *Political Analysis 23*:21–41.

Lieberson, Stanley. 2004. "Comments on the Use and Utility of QCA." *Qualitative Methods: Newsletter of the American Political Science Association Organized Section on Qualitative Methods 2*:13–14.

Lucas, Samuel R. and Alisa Szatrowski. 2014. "Qualitative Comparative Analysis in Critical Perspective." *Sociological Methodology 44*:1–79.

Mackie, J. L. 1980. *The Cement of the Universe: A Study of Causation*. Oxford, UK: Clarendon Press.

McVeigh, Rory, Kraig Beyerlein, Burrel Vann Jr, and Priyamvada Trivedi. 2014. "Educational Segregation, Tea Party Organizations, and Battles over Distributive Justice." *American Sociological Review 79*:630–652.

Miller, William J. and Jeremy D. Walling. 2012. *Tea Party Effects on 2010 U.S. Senate Elections: Stuck in Middle to Lose*. Lanham, MD: Lexington Books.

Parker, Christopher S. and Matt A. Barreto. 2013. *Change They Can't Believe In: The Tea Party and Reactionary Politics in America*. Princeton, NJ: Princeton University Press.

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

Ragin, Charles. 2014. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley, CA: University of California Press.

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

Rihoux, Benoît and Charles Ragin. 2009. *Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques*. Thousand Oaks, CA: Sage.

Schneider, Carsten Q. and Claudius Wagemann. 2007. *Qualitative Comparative Analysis (QCA) und Fuzzy Sets: Ein Lehrbuch für Anwender und jene, die es werden wollen*. Opladen, Germany: Barbara Budrich.

Schneider, Carsten Q. and Claudius Wagemann. 2012. *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. Boston, MA: Cambridge University Press.

Seawright, Jason. 2005. "Qualitative Comparative Analysis vis-à-vis Regression." *Studies in Comparative International Development 40*:3–26.

Seawright, Jason. 2014. "Comment: Limited Diversity and the Unreliability of QCA." *Sociological Methodology 44*:118–121.

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

Skocpol, Theda and Vanessa Williamson. 2012. *The Tea Party and the Remaking of Republican Conservatism*. New York, NY: Oxford University Press.

Thiem, Alrik and Adrian Dusa. 2013. "QCA: A Package for Qualitative Comparative Analysis." *The R Journal 5*:87–97.

U.S. Census Bureau. 2015. *American Community Survey Five-Year Estimates*. Data retrieved from (http://factfinder2.census.gov).

Vaisey, Stephen. 2014. "Comment: QCA Works – When Used with Care." *Sociological Methodology 44*:108–112.

Vann, Burrel. 2018. "Movement-Countermovement Dynamics and Mobilizing the Electorate." *Mobilization 23*:285–305.

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

Zernike, Kate. 2010. *Boiling Mad: Inside Tea Party America*. New York, NY: Times Books.

Table A1: Predicted Probabilities for User-Selected QCA Thresholds

Consistency Score	Configurational N	N	# of Variables	Outcome Distribution	Predicted p
0.80	2	10.00	3	0.10	0.19
0.90	2	10.00	3	0.10	0.13
1.00	2	10.00	3	0.10	0.08
0.80	3	10.00	3	0.10	0.07
0.90	3	10.00	3	0.10	0.04
1.00	3	10.00	3	0.10	0.02
0.80	4	10.00	3	0.10	0.02
0.90	4	10.00	3	0.10	0.01
1.00	4	10.00	3	0.10	0.01
0.80	5	10.00	3	0.10	0.01
0.90	5	10.00	3	0.10	0.00
1.00	5	10.00	3	0.10	0.00
0.80	2	20.00	3	0.10	0.17
0.90	2	20.00	3	0.10	0.11
1.00	2	20.00	3	0.10	0.07
0.80	3	20.00	3	0.10	0.07
0.90	3	20.00	3	0.10	0.04
1.00	3	20.00	3	0.10	0.02
0.80	4	20.00	3	0.10	0.03
0.90	4	20.00	3	0.10	0.01
1.00	4	20.00	3	0.10	0.01
0.80	5	20.00	3	0.10	0.01
0.90	5	20.00	3	0.10	0.01
1.00	5	20.00	3	0.10	0.00
0.80	2	30.00	3	0.10	0.16
0.90	2	30.00	3	0.10	0.10
1.00	2	30.00	3	0.10	0.06

continued on next page

Table A1 – continued from previous page

Consistency Score	Configurational N	N	# of Variables	Outcome Distribution	Predicted p
0.80	3	30.00	3	0.10	0.07
0.90	3	30.00	3	0.10	0.04
1.00	3	30.00	3	0.10	0.02
0.80	4	30.00	3	0.10	0.03
0.90	4	30.00	3	0.10	0.02
1.00	4	30.00	3	0.10	0.01
0.80	5	30.00	3	0.10	0.01
0.90	5	30.00	3	0.10	0.01
1.00	5	30.00	3	0.10	0.00
0.80	2	10.00	4	0.10	0.18
0.90	2	10.00	4	0.10	0.12
1.00	2	10.00	4	0.10	0.08
0.80	3	10.00	4	0.10	0.05
0.90	3	10.00	4	0.10	0.03
1.00	3	10.00	4	0.10	0.02
0.80	4	10.00	4	0.10	0.01
0.90	4	10.00	4	0.10	0.01
1.00	4	10.00	4	0.10	0.00
0.80	5	10.00	4	0.10	0.00
0.90	5	10.00	4	0.10	0.00
1.00	5	10.00	4	0.10	0.00
0.80	2	20.00	4	0.10	0.18
0.90	2	20.00	4	0.10	0.12
1.00	2	20.00	4	0.10	0.08
0.80	3	20.00	4	0.10	0.05
0.90	3	20.00	4	0.10	0.03
1.00	3	20.00	4	0.10	0.02
0.80	4	20.00	4	0.10	0.01
0.90	4	20.00	4	0.10	0.01

continued on next page

Table A1 – continued from previous page

Consistency Score	Configurational N	N	# of Variables	Outcome Distribution	Predicted p
1.00	4	20.00	4	0.10	0.00
0.80	5	20.00	4	0.10	0.00
0.90	5	20.00	4	0.10	0.00
1.00	5	20.00	4	0.10	0.00
0.80	2	30.00	4	0.10	0.17
0.90	2	30.00	4	0.10	0.11
1.00	2	30.00	4	0.10	0.07
0.80	3	30.00	4	0.10	0.06
0.90	3	30.00	4	0.10	0.04
1.00	3	30.00	4	0.10	0.02
0.80	4	30.00	4	0.10	0.02
0.90	4	30.00	4	0.10	0.01
1.00	4	30.00	4	0.10	0.01
0.80	5	30.00	4	0.10	0.01
0.90	5	30.00	4	0.10	0.00
1.00	5	30.00	4	0.10	0.00
0.80	2	10.00	5	0.10	0.17
0.90	2	10.00	5	0.10	0.11
1.00	2	10.00	5	0.10	0.08
0.80	3	10.00	5	0.10	0.03
0.90	3	10.00	5	0.10	0.02
1.00	3	10.00	5	0.10	0.01
0.80	4	10.00	5	0.10	0.01
0.90	4	10.00	5	0.10	0.00
1.00	4	10.00	5	0.10	0.00
0.80	5	10.00	5	0.10	0.00
0.90	5	10.00	5	0.10	0.00
1.00	5	10.00	5	0.10	0.00
0.80	2	20.00	5	0.10	0.18

continued on next page

Table A1 – continued from previous page

Consistency Score	Configurational N	N	# of Variables	Outcome Distribution	Predicted p
0.90	2	20.00	5	0.10	0.12
1.00	2	20.00	5	0.10	0.08
0.80	3	20.00	5	0.10	0.04
0.90	3	20.00	5	0.10	0.02
1.00	3	20.00	5	0.10	0.01
0.80	4	20.00	5	0.10	0.01
0.90	4	20.00	5	0.10	0.00
1.00	4	20.00	5	0.10	0.00
0.80	5	20.00	5	0.10	0.00
0.90	5	20.00	5	0.10	0.00
1.00	5	20.00	5	0.10	0.00
0.80	2	30.00	5	0.10	0.19
0.90	2	30.00	5	0.10	0.13
1.00	2	30.00	5	0.10	0.09
0.80	3	30.00	5	0.10	0.05
0.90	3	30.00	5	0.10	0.03
1.00	3	30.00	5	0.10	0.02
0.80	4	30.00	5	0.10	0.01
0.90	4	30.00	5	0.10	0.01
1.00	4	30.00	5	0.10	0.00
0.80	5	30.00	5	0.10	0.00
0.90	5	30.00	5	0.10	0.00
1.00	5	30.00	5	0.10	0.00
0.80	2	10.00	3	0.50	0.35
0.90	2	10.00	3	0.50	0.33
1.00	2	10.00	3	0.50	0.31
0.80	3	10.00	3	0.50	0.16
0.90	3	10.00	3	0.50	0.14
1.00	3	10.00	3	0.50	0.12

continued on next page

Table A1 – continued from previous page

Consistency Score	Configurational N	N	# of Variables	Outcome Distribution	Predicted p
0.80	4	10.00	3	0.50	0.06
0.90	4	10.00	3	0.50	0.05
1.00	4	10.00	3	0.50	0.04
0.80	5	10.00	3	0.50	0.02
0.90	5	10.00	3	0.50	0.02
1.00	5	10.00	3	0.50	0.01
0.80	2	20.00	3	0.50	0.37
0.90	2	20.00	3	0.50	0.34
1.00	2	20.00	3	0.50	0.32
0.80	3	20.00	3	0.50	0.19
0.90	3	20.00	3	0.50	0.17
1.00	3	20.00	3	0.50	0.15
0.80	4	20.00	3	0.50	0.09
0.90	4	20.00	3	0.50	0.07
1.00	4	20.00	3	0.50	0.06
0.80	5	20.00	3	0.50	0.04
0.90	5	20.00	3	0.50	0.03
1.00	5	20.00	3	0.50	0.02
0.80	2	30.00	3	0.50	0.38
0.90	2	30.00	3	0.50	0.36
1.00	2	30.00	3	0.50	0.33
0.80	3	30.00	3	0.50	0.23
0.90	3	30.00	3	0.50	0.20
1.00	3	30.00	3	0.50	0.17
0.80	4	30.00	3	0.50	0.12
0.90	4	30.00	3	0.50	0.10
1.00	4	30.00	3	0.50	0.08
0.80	5	30.00	3	0.50	0.06
0.90	5	30.00	3	0.50	0.05

continued on next page

Table A1 – continued from previous page

Consistency Score	Configurational N	N	# of Variables	Outcome Distribution	Predicted p
1.00	5	30.00	3	0.50	0.04
0.80	2	10.00	4	0.50	0.30
0.90	2	10.00	4	0.50	0.29
1.00	2	10.00	4	0.50	0.28
0.80	3	10.00	4	0.50	0.10
0.90	3	10.00	4	0.50	0.09
1.00	3	10.00	4	0.50	0.08
0.80	4	10.00	4	0.50	0.03
0.90	4	10.00	4	0.50	0.02
1.00	4	10.00	4	0.50	0.02
0.80	5	10.00	4	0.50	0.01
0.90	5	10.00	4	0.50	0.01
1.00	5	10.00	4	0.50	0.00
0.80	2	20.00	4	0.50	0.34
0.90	2	20.00	4	0.50	0.32
1.00	2	20.00	4	0.50	0.31
0.80	3	20.00	4	0.50	0.13
0.90	3	20.00	4	0.50	0.12
1.00	3	20.00	4	0.50	0.10
0.80	4	20.00	4	0.50	0.04
0.90	4	20.00	4	0.50	0.03
1.00	4	20.00	4	0.50	0.03
0.80	5	20.00	4	0.50	0.01
0.90	5	20.00	4	0.50	0.01
1.00	5	20.00	4	0.50	0.01
0.80	2	30.00	4	0.50	0.38
0.90	2	30.00	4	0.50	0.36
1.00	2	30.00	4	0.50	0.34
0.80	3	30.00	4	0.50	0.17

continued on next page

Table A1 – continued from previous page

Consistency Score	Configurational N	N	# of Variables	Outcome Distribution	Predicted p
0.90	3	30.00	4	0.50	0.15
1.00	3	30.00	4	0.50	0.13
0.80	4	30.00	4	0.50	0.07
0.90	4	30.00	4	0.50	0.05
1.00	4	30.00	4	0.50	0.05
0.80	5	30.00	4	0.50	0.02
0.90	5	30.00	4	0.50	0.02
1.00	5	30.00	4	0.50	0.01
0.80	2	10.00	5	0.50	0.26
0.90	2	10.00	5	0.50	0.25
1.00	2	10.00	5	0.50	0.24
0.80	3	10.00	5	0.50	0.06
0.90	3	10.00	5	0.50	0.05
1.00	3	10.00	5	0.50	0.05
0.80	4	10.00	5	0.50	0.01
0.90	4	10.00	5	0.50	0.01
1.00	4	10.00	5	0.50	0.01
0.80	5	10.00	5	0.50	0.00
0.90	5	10.00	5	0.50	0.00
1.00	5	10.00	5	0.50	0.00
0.80	2	20.00	5	0.50	0.31
0.90	2	20.00	5	0.50	0.30
1.00	2	20.00	5	0.50	0.29
0.80	3	20.00	5	0.50	0.09
0.90	3	20.00	5	0.50	0.08
1.00	3	20.00	5	0.50	0.07
0.80	4	20.00	5	0.50	0.02
0.90	4	20.00	5	0.50	0.02
1.00	4	20.00	5	0.50	0.01

continued on next page

Table A1 – continued from previous page

Consistency Score	Configurational N	N	# of Variables	Outcome Distribution	Predicted p
0.80	5	20.00	5	0.50	0.00
0.90	5	20.00	5	0.50	0.00
1.00	5	20.00	5	0.50	0.00
0.80	2	30.00	5	0.50	0.37
0.90	2	30.00	5	0.50	0.36
1.00	2	30.00	5	0.50	0.34
0.80	3	30.00	5	0.50	0.13
0.90	3	30.00	5	0.50	0.11
1.00	3	30.00	5	0.50	0.10
0.80	4	30.00	5	0.50	0.03
0.90	4	30.00	5	0.50	0.03
1.00	4	30.00	5	0.50	0.02
0.80	5	30.00	5	0.50	0.01
0.90	5	30.00	5	0.50	0.01
1.00	5	30.00	5	0.50	0.01
0.80	2	10.00	3	0.90	0.56
0.90	2	10.00	3	0.90	0.64
1.00	2	10.00	3	0.90	0.71
0.80	3	10.00	3	0.90	0.33
0.90	3	10.00	3	0.90	0.39
1.00	3	10.00	3	0.90	0.45
0.80	4	10.00	3	0.90	0.16
0.90	4	10.00	3	0.90	0.19
1.00	4	10.00	3	0.90	0.22
0.80	5	10.00	3	0.90	0.07
0.90	5	10.00	3	0.90	0.08
1.00	5	10.00	3	0.90	0.09
0.80	2	20.00	3	0.90	0.62
0.90	2	20.00	3	0.90	0.69

continued on next page

Table A1 – continued from previous page

Consistency Score	Configurational N	N	# of Variables	Outcome Distribution	Predicted p
1.00	2	20.00	3	0.90	0.75
0.80	3	20.00	3	0.90	0.42
0.90	3	20.00	3	0.90	0.48
1.00	3	20.00	3	0.90	0.55
0.80	4	20.00	3	0.90	0.25
0.90	4	20.00	3	0.90	0.29
1.00	4	20.00	3	0.90	0.33
0.80	5	20.00	3	0.90	0.13
0.90	5	20.00	3	0.90	0.15
1.00	5	20.00	3	0.90	0.16
0.80	2	30.00	3	0.90	0.68
0.90	2	30.00	3	0.90	0.74
1.00	2	30.00	3	0.90	0.79
0.80	3	30.00	3	0.90	0.52
0.90	3	30.00	3	0.90	0.58
1.00	3	30.00	3	0.90	0.64
0.80	4	30.00	3	0.90	0.37
0.90	4	30.00	3	0.90	0.41
1.00	4	30.00	3	0.90	0.45
0.80	5	30.00	3	0.90	0.23
0.90	5	30.00	3	0.90	0.26
1.00	5	30.00	3	0.90	0.28
0.80	2	10.00	4	0.90	0.47
0.90	2	10.00	4	0.90	0.55
1.00	2	10.00	4	0.90	0.63
0.80	3	10.00	4	0.90	0.19
0.90	3	10.00	4	0.90	0.24
1.00	3	10.00	4	0.90	0.30
0.80	4	10.00	4	0.90	0.06

continued on next page

Table A1 – continued from previous page

Consistency Score	Configurational N	N	# of Variables	Outcome Distribution	Predicted p
0.90	4	10.00	4	0.90	0.08
1.00	4	10.00	4	0.90	0.10
0.80	5	10.00	4	0.90	0.02
0.90	5	10.00	4	0.90	0.02
1.00	5	10.00	4	0.90	0.03
0.80	2	20.00	4	0.90	0.55
0.90	2	20.00	4	0.90	0.63
1.00	2	20.00	4	0.90	0.70
0.80	3	20.00	4	0.90	0.29
0.90	3	20.00	4	0.90	0.35
1.00	3	20.00	4	0.90	0.41
0.80	4	20.00	4	0.90	0.12
0.90	4	20.00	4	0.90	0.14
1.00	4	20.00	4	0.90	0.17
0.80	5	20.00	4	0.90	0.04
0.90	5	20.00	4	0.90	0.05
1.00	5	20.00	4	0.90	0.06
0.80	2	30.00	4	0.90	0.64
0.90	2	30.00	4	0.90	0.71
1.00	2	30.00	4	0.90	0.77
0.80	3	30.00	4	0.90	0.40
0.90	3	30.00	4	0.90	0.46
1.00	3	30.00	4	0.90	0.53
0.80	4	30.00	4	0.90	0.20
0.90	4	30.00	4	0.90	0.24
1.00	4	30.00	4	0.90	0.28
0.80	5	30.00	4	0.90	0.09
0.90	5	30.00	4	0.90	0.10
1.00	5	30.00	4	0.90	0.11

continued on next page

Table A1 – continued from previous page

Consistency Score	Configurational N	N	# of Variables	Outcome Distribution	Predicted p
0.80	2	10.00	5	0.90	0.37
0.90	2	10.00	5	0.90	0.46
1.00	2	10.00	5	0.90	0.55
0.80	3	10.00	5	0.90	0.11
0.90	3	10.00	5	0.90	0.14
1.00	3	10.00	5	0.90	0.18
0.80	4	10.00	5	0.90	0.02
0.90	4	10.00	5	0.90	0.03
1.00	4	10.00	5	0.90	0.04
0.80	5	10.00	5	0.90	0.00
0.90	5	10.00	5	0.90	0.01
1.00	5	10.00	5	0.90	0.01
0.80	2	20.00	5	0.90	0.49
0.90	2	20.00	5	0.90	0.57
1.00	2	20.00	5	0.90	0.66
0.80	3	20.00	5	0.90	0.18
0.90	3	20.00	5	0.90	0.23
1.00	3	20.00	5	0.90	0.28
0.80	4	20.00	5	0.90	0.05
0.90	4	20.00	5	0.90	0.06
1.00	4	20.00	5	0.90	0.08
0.80	5	20.00	5	0.90	0.01
0.90	5	20.00	5	0.90	0.01
1.00	5	20.00	5	0.90	0.02
0.80	2	30.00	5	0.90	0.60
0.90	2	30.00	5	0.90	0.68
1.00	2	30.00	5	0.90	0.75
0.80	3	30.00	5	0.90	0.29
0.90	3	30.00	5	0.90	0.35

continued on next page

Table A1 – continued from previous page

Consistency Score	Configurational N	N	# of Variables	Outcome Distribution	Predicted p
1.00	3	30.00	5	0.90	0.42
0.80	4	30.00	5	0.90	0.10
0.90	4	30.00	5	0.90	0.12
1.00	4	30.00	5	0.90	0.15
0.80	5	30.00	5	0.90	0.03
0.90	5	30.00	5	0.90	0.04
1.00	5	30.00	5	0.90	0.04