May 17, 2016

4100 Calit2 Building
UCI, California, 92617
Phone: (251) 510-0864
Email: cbgibson@uci.edu

Tim Futing Liao
Sociological Methodology
University of Illinois

Dear Editors:

Please find the responses to reviewers below for submission SMX-15-0022. Please also note that, at the suggestion of reviewer 3, we have changed the title of the paper to "The Bootstrapped Robustness Assessment for Qualitative Comparative Analysis. Due to these changes, we have also changed the name of the package (braQCA) and it's components (baQCA and brQCA).

We have also changed formatting (text size, font, notes) to match SMX's requirements.

Thank you.

Reviewer 1: Thank you for your comment. As for the minor comment about including counterfactuals in this assessment, we address this in the extended answer to reviewer 2.

Reviewer 2: We appreciate Reviewer 2's comments. Though the reviewer calls the method 'seriously defective,' he or she seems to desire a method with different goals than the method proposed by the authors. The reviewer criticizes the method for not comparing the observed result with other causal configurational combinations, but that would require another robustness method entirely. Instead, we propose a principled method to determine whether existing thresholds (i.e. configurational n and consistency score thresholds) are effective in eliminating observed combinations of causal conditions that are highly likely to be due to chance. The method proposed by the reviewer may be useful, but the reviewer's goals clearly do not align with ours. We concede that perhaps we were not clear in the paper on the goal of this method; to attempt to eliminate ambiguity, we have fleshed out more theory in the body of the paper, which is marked in green.

Though we feel that the above response is a sufficient defense of the paper, an extended response is below.

To begin, I will follow the reviewers data example:

> Let me end by emphasizing a radically inadequate consequence of iaQCA – that ultimately is due to its import of a RAM-inspired reliability test into QCA. Take the following table:

A	B	C
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	0	1
1	0	1

```
0 1 1
0 1 1
0 0 0
0 0 0
0 0 0
0 0 0
0 0 0
0 0 0
0 0 0
```

This is ideal data by the standards of QCA. It feature no limited diversity and no inconsistencies. Moreover, the outcome C can be covered to a maximal degree. The perfectly consistent and maximally covering QCA solution for this table is

$$A + B < - > C$$

There simply is no better configurational data than this. Nonetheless, the iaQCA R-package calculates a probability of 0.4075 that this result is spurious (with consistency set to 1 and frequency to 2)!

To explain our reply, we explain why we take issue with the idealness of this set of data. Though loaded with rows with only 0s and only 1s, the case for $A + B < - > C$ is not founded due to the few cases suggesting $A < - > C or B < - > C$, respectively. Without just one row in each of these configurational subsets, the solution changes. For example, without row seven and n.cut=2, it changes from $A + B < - > C$ to $A < - > C$ or $A + B < - > C$; without row nine and n.cut=2, it changes to $A < - > C$ or $A + B < - > C$. This is not due to the general sensitivity of QCA given changes in data; it is due to the lack of evidence to change the solution by nature of having only two cases. The application of baQCA suggests not including those rows in the final set due to the uncertainty that those observations were not due to random chance. The retort here is that rows 7-10 observed together signal a causal process, so an 'n.cut' should be applied to all four rows (and rows 1-6), not 7-8 or 9-10 alone. But this process is not known *a priori*, and each set of two rows shifts the causal process observed. The idea, thus, is to take each set of observations alone, as each unique set of observations between the two variables can alter the solution found by QCA. If we were to only consider the solution informed by spurious data versus all others, we would miss the fact that the solution itself was not robust. If we only consider the observed causal process, we would only be able to confirm a spurious solution.

The brQCA is not determining whether the data collected by a researcher is not indicative of a causal process; it is determining whether QCA would return a solution when given random data with the same parameters as the observed data. Comparing a result with random chance is a useful technique in nearly all data summaries, from t-tests to BIC, but like those tests, its interpretation is limited. This method simply applies the same technique to QCA. Our technique does not imply any conclusion about data quality, just that a solution returned by QCA is more robust than the extreme case of completely random data as a very low baseline; more specifically, it determines whether configurations that do not have many cases or low agreement in the direction of the outcome should be used in the solution proposed by QCA.

Though it is true (in theory) that 'for every thus drawn data table there exists a causal structure that systematically generates this very table,' that doesn't preclude random error in QCA data sets. Applying QCA in practice would not be complete without a consideration of configurational N and consistency score thresholds, which are intended to account for some error in observations. This is precisely why the "null model" is a useful technique in the QCA case. If there does not exist a causal structure that systematically generates the data, but an application of QCA would indeed suggest such a causal structure, then QCA is flawed and should not be used. The bootstrapped assessment for QCA provides additional evidence that such a structure would not be detected by QCA in the case that a random data set is drawn.

By keeping marginal distributions intact, we assert that a purely random data set would maintain similar distributions of 'random variables' without a causal mechanism producing dependencies between random variables (on average). If we assumed dependencies between random variables in our baseline, then we would not be comparing against a random data set, by definition. Relatedly, keeping marginal distributions intact does produce random configurations by nature of simply choosing multiple values in the same row. Finally, the marginal distributions of the causal conditions are not shown have much effect on the end result of brQCA, so are not totally necessary for computing the method; however, the marginal distribution of the outcome is hugely predictive of the end result, and should be used.

Also, if we take a second look at the data provided, we see that brQCA finds this to be a robust solution (p ¡ .1) with just an n.cut of 3 (and p ¡ .01 at n.cut=4, which does not change the final solution):

```
mod<-eqmcc(dat, "C",n.cut=2,incl.cut=1)
baQCA(mod)

dat<-dat[c(1:6,11:17),] #remove the n.cut<2 rows, QCA returns an error if we don't...

mod<-eqmcc(dat, "C",n.cut=3,incl.cut=1)
with3<-baQCA(mod)
with3
$Probability
[1] 0.092

mod<-eqmcc(dat, "C",n.cut=4,incl.cut=1)
with4<-baQCA(mod)
with4
$Probability
[1] 0.008
```

In many cases, simply using the observed cases with a minimum of three or more cases per solution is sufficient to assure that a returned configuration is robust.

Finally, we emphasize two points, and in fact, have edited the paper to reinforce these points. The first point is that this method is merely informing the robustness thresholds already present in Ragin's initial book. QCA does in fact allow for some error in its configurations, which is the idea behind configurational n and consistency score thresholds. This method merely asserts a principled basis to use existing robustness tools. Secondly, this method is intended to inform, not replace, human observation. If there exists compelling qualitative evidence of a causal process without sufficient data, then that will need to be justified. However, our method warns against (in many cases, not all) changing a causal process because of the observation of two cases.

Reviewer 3: Thank you for the constructive comments. First, we've changed the name to "Boostrapped Robustness Assessment for QCA," which is more descriptive. Second, we've included some comments on the predictions from the regression analysis, as well as some ideal cases for a 'robust' configuration at the standard .05 level. For the sake of clarity, we rely primarily on predicted values, rather than coefficient size. Though the regression analysis plots do not show it, QCA is generally robust when using a configurational N threshold of 4-5.

With that caveat included, we have qualified the paper to include some reservations when using QCA, as in many cases it does not achieve what is traditionally considered 'robust' by other statistical methodology. Edits to the manuscript in response to reviewer 3 is denoted in blue.

C. Ben Gibson
Department of Sociology
University of California, Irvine
4100 Calit2 Building
UCI, California, 92617
Phone: (251) 510-0864 (cell)
Email: cbgibson@uci.edu

Sincerely,

C. Ben Gibson