Warrantable and Unwarranted Methods: The Case of QCA

Jason Seawright
Department of Political Science, Northwestern University

Abstract

A methodologically pluralistic version of the social sciences requires norms to bound the set of methods that may be seen as admissible in debates about causal inference. This essay proposes a standard of warrantability, which is essentially the idea that methods should only be seen as usable if there is at least some class of research design situations in which a good argument can be made that the method in question works. After developing the nuances of this standard, the essay applies it by building a case that one important family of methods (crisp-set QCA) is currently unwarrantable. This case is made by showing that cs-QCA lacks justifying analytic results, performs relatively poorly in simulations, and frequently cannot replicate known causal results. By contrast, other methods including the use of regression in experiments and CART to analyze Boolean-type causal complexity, come closer to meeting standards of warrantability.

Introduction

In the wake of debates between quantitative and qualitative methods (e.g., King, Keohane and Verba 1994), the social sciences in general appear to be adopting methodological pluralism as a default outlook (e.g., Brady and Collier 2004). Yet pluralism confronts each analyst time and again with the problem of choosing a method out of the wide range that have been developed. It appears that there is no easy solution to this quandary.

One kind of solution would be to believe that there is a best method for making causal inferences with observational studies, to raise a familiar version of the problem. Yet this view faces serious difficulties. In the first place, if there is a universally best method for causal inference in observational contexts, scholars have not yet developed any kind of consensus about what that method is. Furthermore, if as is plausible different causal domains are characterized by different kinds of causal relationships, then there may well be no single best method because different inferences would be beset by divergent categories of challenges. Finally, there is the problem that even if there is a best method and we have already invented it, scholars would not necessarily be able to find enough or the right kind of evidence to prove that this happy state of affairs held. For these reasons among others, it seems unlikely that dilemmas of methodological choice will be resolved through the emergence of one obviously superior alternative.

Another potential resolution, which enjoys some popularity, is the claim that the specifics of a given research question provide scholars with the evidence necessary to choose the best available method for answering that question. This proposal seems deficient both as a characterization of how scholars choose research methods and of the kind of knowledge scholars could conceivably have of their research question at the moment of methodological choice. After all, as an observational matter, scholars tend to bring the same methodological toolkit to each question that they tackle. This is either an unaccountably unlikely fact about the research questions people choose to follow — i.e., people just happen to choose problem after problem that imply the use of the same set of methods — or it is evidence that people use the methods that they like and have mastered across research questions, and that the choice of method is somewhat independent of the problem at hand. This conclusion fits nicely with the more abstract point that, in principle, determining the best method for a given domain without prior knowledge of the method's superiority would require pre-answering all the important causal questions about that domain so that the results of a given method could be systematically. Since such in-depth causal prior knowledge is scarce to say the least, scholars generally do not know enough to choose a best method, even if they were not in practice driven by methodological habit and acquired expertise.

Instead, scholars seem to be left with a choice among a huge range of methods. Sensible navigation requires some means of narrowing down that range; it is implausible for scholars to know the details of all or even most possible (or even actually proposed) methods for causal inference. Can this pruning be done in a principled way, or are scholars ineluctibly reduced to arbitary, personal heuristics?

In fact, there are grounds for optimism. While scholars lack the knowledge to narrow the set of appropriate methods to one or a handful of choices, even for a defined research problem, it is nonetheless clear that social scientists have near-consensus about the usefulness of a wide variety of methods. For example, no prominent scholars use or advocate the method in which every causal effect is said to be equal to 0.53; a vast array of equally arbitrary methods is categorically rejected without a second thought. Near the opposite extreme, relatively few scholars object to using natural experimental methods in contexts in which the treatment of interest is subjected to a true randomization (Dunning 2012: 48-53).

Of course, for any given research problem, methods anywhere on this implicit spectrum of credibility may produce the right answer. After all, some causal effects really are (at least approximately) equal to 0.53. By the same token, even the most credible methods for making causal inferences in observational studies can go badly wrong. So the idea here cannot be to separate truth from error, or completely reliable methods from hopelessly incorrect ones.

Instead, a more modest and plausible question is pursued: is there a principled way to draw a dividing line separating methods lower on this implied spectrum from those higher on the spectrum? More rigorously, are there publicly observable and debatable criteria that differentiate between a set of methods that — to the best of our current knowledge — are unlikely to result in valid causal inference from a second set that are more likely? Such criteria cannot solve the problem of methodological choice in a world of pluralism, but they can bound the problem and help render its dimensions at least marginally more manageable.

No doubt any number of such divisions might be proposed, some more controversial than others. In this essay, I argue for a rather minimal decision rule. I propose, as a first cut for dividing plausible and usable methods from those which are unacceptable, distinguishing between *warrantable* and *unwarranted* methods.[1] A method is warrantable if knowledge and broadly per-

[1] Alston's helpful analysis of justification and warrant draws the conclusion that justification fundamentally involves an epistemic evaluation of the acceptability of holding a particular belief (Alston 1989: 83-84). Of course, people hold beliefs about research methods, and those beliefs themselves may be justified or not. However, the issue of greater importance is almost always whether the use of a given method in a particular context results in justifiable beliefs about that context; that is, the beliefs that may be warranted by use of a method are ultimately more important than the beliefs that are warranted about the method itself. Hence the use here of the alternative term, warrantable,

suasive arguments exist that give an informed outsider a reasonable basis for believing that the method would work under some set of causally generic circumstances that might plausibly be non-trivially present in the domain of inquiry. Thus a warrantable method can at least sometimes give an outsider philosophical warrant or justification for believing that a certain causal claim is true; unwarrantable methods cannot, and therefore their utility is rather limited.[2] Clearly several elements of this standard need further discussion.

Informed outsiders are an essential element of the proposed standard of warrantability, in keeping with the widespread recognition in the philosophy of science and in epistemology that the quality of arguments cannot be judged in the abstract but only in a social, persuasive context — because that quality depends on such background factors as the perceived plausibility of the premises in an argument and the relative value to be placed on different elements in a trade-off (see, e.g., Goldman 1999). If such background factors line up in such a way that a method can only be seen as good by people who currently advocate it, then it follows that no one outside that community should be persuaded by the results of using that method, and hence the method is unwarrantable. But the distribution of such background factors can only be explored in interaction with people outside the community of users of a method; such outsiders thus become the ultimate judges of warrantability. These individuals need not be objective in any particular sense — a difficult and contested standard in any case (see, e.g., Novick 1988, Daston and Galison 2007, Gaukroger 2012). Nonetheless, the ideal interlocutor for an analysis of warrantability would be intellectually honest in the sense that she is open to surprising evidence about the method in question and willing to revise her views in line with evidence when it is made available.

Such outsiders should consider direct evidence about the method itself, rather than about specific applications. This is essential to avoid a potential pitfall in reliable process accounts of warranted belief (Vogel 2000, Cohen 2002). Epistemic circularity arises if we use standards for

to separate those methods whose use can in at least some context warrant specific beliefs from those that cannot.

[2] In discussing warrant, I am adopting the term's widespread usage in epistemology in which it is essentially interchangeable with justification. I thus do not intend to invoke Plantinga's (1993) more specialized but nonetheless influential theological usage, in which a belief can only be warranted if the mind holding the belief was designed by a creator with the intention of correctly handling information of a given kind. Nonetheless, Plantinga's views are one position within the tradition of process reliabilism, which claims that beliefs degree of warrant is determined primarily by the traits of the argumentation or reasoning that caused the belief (Goldman 2011); warrantability as discussed here obviously draws on this tradition.

evaluating methods based on specific contested examples. When such examples are allowed to be the basis for warrant, it can be warranted to believe in a collection of findings because of the method used to produce them and simultaneously warranted to believe that the method gets things right because of the collection of findings it has produced. Such unhelpful circularity can be avoided by requiring direct evidence that the method at least sometimes works.[3] By contrast, attempts to bolster a method via appeals to the results of using that method to answer contested research questions are generally caught in this kind of epistemic circularity.

Warrantability requires relevant evidence that the method in question works to recover at least some knowledge about *causally generic* situations. Causally generic situations are those in which a particular kind of causal relationship may be known to hold, but the specifics of existing causal patterns are not known in advance. This is a way of delimiting the kind of prior knowledge about the situation that it is reasonable to demand that the idealized outsider consider. If no limit is imposed, then we might stipulate that our set of relevant circumstances is just exactly those with a causal effect of 0.53; for this set, what we know about the method of concluding that all causal effects are equal to 0.53 implies perfect success. The requirement that circumstances be causally generic is intended as a solution to this difficulty: a causally generic situation is one that may stipulate a given category of causal model, but not a specific causal finding. Thus, we might discuss contexts in which causation is known to be additive, linear, and constant across cases, but not contexts in which each unit of the treatment is known to cause a 0.425 unit increase in the outcome.

An important implication follows from this discussion of causally generic situations: warrantable methods must be able to recover causal knowledge even when the user of the method falsely believes the causal effect of one or more variables to be non-zero. Stipulating in advance that all true effects for all included variables be non-zero is just another way of defining the context of applicability by specific findings rather than by generic models, and therefore should be

[3] Because it attends to these kinds of evidence in particular, warrantability side-steps some other criticisms of process reliabilism accounts of justification. Warrantability is not an account of when belief in general is justified, but rather of whether there is reason to think that a particular method can ever warrant a specific belief. As such, it is immune to criticisms of reliable process as unnecessary (e.g., Cohen, 1984) or insufficient (e.g., BonJour 1980, Lehrer 1990).

inadmissable. Thus, warrantability imposes the necessary requirement that a method be at least somewhat robust to the erroneous inclusion of irrelevant variables.[4]

All of this taken together suggests that informed outsiders should feel free to reject as unwarrantable the use of methods that fail in the following ways, among others. First, any methods for which we have no body of well-established knowledge about how the method works, apart from applications to contested causal inferences, are unwarranted. Second, methods also are likely to fail the standard if the knowledge we have about them cannot pick out any causally generic domain in which the method is likely to work. Third, observers should feel free to classify as unwarrantable methods for which there is such a domain but no group of scholars believes that domain to have a nontrivial set of exemplars within the research context of interest. In any of these cases, scholars have a sound basis for concluding that no use of the method could ever reasonably support belief in any concrete proposition — and therefore the method is unwarrantable, and thus largely lacking in practical value.

Warrantability is not a radical new idea by any means. It boils down to the idea that there is at least some class of research design situations in which a good argument can be made that the method in question works. As long as there is some such specialized situation, an argument by analogy is available for scholars who wish to use the method outside its specialized class of situations. In general, such arguments by analogy are rather weak and will be controversial. Warrantability does not by any means settle questions of the general value or specific applicability of a method; instead, it is a low-bar threshold that simply helps decide whether a method can *ever* reasonably be used.

The remainder of this essay is devoted to two more specialized tasks. First, I illustrate three key forms of knowledge that are useful in evaluating a method's warrantability: analytic results, simulations, and replications. While this list is surely not exhaustive, these are the most common

[4] Some observers, in evaluating methods, will want to impose a related but stronger condition, requiring the method to be viable even when the user of the method erroneously believes that the effect of a given variable is zero and therefore incorrectly discards it from the analysis. Thus, in this more stringent view, warrantable methods must necessarily be at least in some contexts reliable in the face of confounders/omitted variables. This is clearly a standard of the same kind as the overall requirement of causally generic situations and thus cannot be rejected out of hand; on the other hand, many real-world informed outsiders in the social sciences will obviously act from a knowledge that many widely used methods fail this more stringent standard and may therefore not require it for warrantability.

sources of knowledge about methods' properties and therefore the most reasonable starting place for any argument about warrantability. Analytic results, simulations, and replications reflect different blends of a pair of fundamental ways of knowing that a method might be useful: reasoning from the generally-known properties of the method and knowledge of the research situation to the results of using that method, and empirically demonstrating that the method reproduces a set of previously known truths.

Analytic results — either formal mathematical proofs or less formal verbal logical arguments — are the prototype of the first way of knowing. They start with known facts about a method (prototypically properties of the mathematical machinery it uses, but also in more qualitative forms considerations such as psychological knowledge regarding human response to a category of stimulus, historical information about the origins and intended audience of a document, and so forth). These facts are then combined with a description of the research situation — for example, random assignment in a laboratory with subjects isolated from each other, or a causal pathway that is known to be isolated from other causal factors and an exhaustive account of the connection between the treatment and the outcome — to come up with a result that justifies causal inference. The central insight is that there are facts about methods, and learning those facts can sometimes support rigorous analysis of the contexts in which a method works.

The other two methods work primarily via empirical demonstration that the method in fact works in the sense of replicating causal findings that are known in advance. These are useful tools because analytic results are not always available. After all, sometimes the relevant facts about a method are unknown or are too complex to reason with effectively. In these contexts, scholars can make a convincing argument that a method works for causal inference at least in a bounded domain by showing that the method gets a known answer right. The challenge here is finding a causal answer to serve as the criterion that can be reliably known to be right. Simulations resolve this by letting the scholar design and implement the true causal process in computer code. In such a circumstance, all the relevant causal knowledge about the process in question is known, and it is entirely feasible to compare a method's causal inference to the truth, and to score the method as right or wrong.

Replications use real-world causal processes instead of computer-generated ones, so the challenge of finding a credible criterion is more rigorous. One useful approach is to take advantage of analytic results to identify a method — such as bivariate regression based on laboratory experimentation — that is known to produce good causal inferences in a given context. Then that method and the method of interest can both be applied to check whether the method of interest in fact gets the answer right. Obviously, this approach is only as good as the criterion used; comparing a new method to an established method of dubious warrant contributes little.

These approaches to arguing warrantability thus variously use facts about the internal workings of a method and facts about the external usefulness of the method to make a case for or against that method. While other kinds of arguments could surely be used, these have historically been the most common ways of arguing for what I call warrantability — and thus should probably be the first resources in any debate of this sort. The best established methods, qualitative and quantitative, are defended in just these ways; a central component of the idea of warrantability is that methods in general should be regarded with distrust until there is a reasonably substantial group of outside observers who feel that a compelling argument of this kind has been made in favor of the method.

The second task of the remainder of this essay is to demonstrate that the standard of warrantability is rigorous enough for at least one informed outsider[5] to exclude at least one real-world method as unwarrantable, while competing methods are in fact warrantable. I make this case by showing that one widely discussed method in the social sciences — crisp-set, deterministic QCA — is at present unwarrantable. This paper will offer no general overview of QCA methods. Instead basic familiarity with them will be assumed. Readers who need further background on QCA are referred to Ragin (1987) and Schneider and Wagemann (2012). In parallel to this, I sketch the ways that analytic results can render regression analysis warrantable in the context of experiments, and show that simulation data provide a starting point for a potential argument that CART is a warrantable approach to analyzing Boolean-type causal complexity.

[5] The author is of course the evaluator in question, although the reader is invited to consider the evidence and render a verdict in parallel.

Analytic Results

Proofs are a hugely important and influential approach to satisfying the requirements of warrantability, although they are not the only relevant set of tools. If a valid formal argument can be constructed connecting realistic assumptions and the known properties of a method through to the goal of causal inference in some plausible empirical context, then that method should generally be seen as warrantable. Hence, an important starting point for any method is to ask what can be proven about its workings.

There are, for example, well-known proofs that justify using bivariate regression to draw causal inferences with experimental data. Without going into any detail, these proofs revolve around one feature of experimental design and one statistical law: random assignment of cases to the treatment and control groups and the law of large numbers. Because cases are randomly assigned, the treatment group and the control group are two independent random samples from the same population of cases. The law of large numbers essentially tells us that large, independent random samples from the same population have the same distributions on all variables, measured and unmeasured.

Let us use a bivariate regression-like model to sketch the causal process involved in a true (i.e., randomized and manupulated) experiment, with Y_i as the mean-deviated outcome variable, T_i indicating a case's assignment to the treatment or control group, β_1 representing the average causal effect of the treatment in the population, and ϵ_i representing both the effects of variables other than the treatment on the outcome and cases' deviations from the population average causal effect for the treatment of interest:

$$Y_i = \beta_1 T_i + \epsilon_i \tag{1}$$

The law of large numbers in combination with random assignment and an independence assumption[6] means that the mean of ϵ_i among treatment cases will be zero; the same will be true for the mean of ϵ_i among control cases. This fact justifies a family of common proofs showing that

[6] This assumption, often given the unhelpful name of SUTVA (Rubin 1980), is complex and not of central concern in this discussion. Nonetheless, it bears mention that this assumption can be met by design in experiments as long as subjects in the experiment are isolated from each other.

a regression estimate of β_1 will be an unbiased estimate of the population parameter. But because that parameter is causally defined, β_1 in this context can be causally interpreted (e.g., Freedman 2008). Furthermore, for finite (or even small) sample sizes, techniques such as randomization inference help differentiate between estimates of β_1 that probably represent real causal effects and those that may simply be flukes of small sample size. Thus, for this method there are key analytic results that make causal inference altogether plausible in the real-world context of experiments with random assignment.

For some methods, including a range of case-study approaches, there are few if any relevant formal proofs of this sort. Nonetheless, scholars have sometimes constructed analogous lines of evidence based on rigorous and compelling — but non-mathematical — arguments from assumptions about the research context and about the properties of the method in question. For example, in debates regarding process tracing scholars have quite reasonably argued that evidence about information flows can rule out some causal propositions. Qualitative methods are sometimes capable of showing that a certain piece of information or idea was unavailable to decision-makers in a particular time or place. This demonstration may take multiple forms: documents may show that the information in question was not developed until a later date, for example, or transcripts of key meetings may show decision-makers expressing frustration that they do not know the information. Even though such evidence does not absolutely prove that the information was unavailable, it is reasonable to propose that it sometimes suffices to make credible, or even highly probable, the claim that it was unavailable. This, in combination with the relatively uncontroversial assertion that unknown information has no causal effects, can suffice to justify a causal inference. While this argument is obviously not a mathematical proof, it is nonetheless a relevant analytic argument that proceeds from properties of the method and the data to conclusions justifying causal claims. That is to say, this kind of evidence counts for warrantability.

Of course, if warrantabiity is to be a standard with teeth, it must be true that not all methods are justified by relevant analytic results. This is in fact the case: QCA in particular appears to lack justification of this sort. Like most methods of empirical research, QCA has initial phases involving variable selection, measurement, decisions about functional form and transformations, and so forth. Yet the analytic and inferential core of the method is the use of the Quine-McCluskey algorithm to carry out Boolean minimization. This algorithm accepts a non-contradictory truth function, or Boolean equation translating true-false scores on a number of right-hand-side variables into a true-false score on the left-hand side, and returns a version that is logically equivalent but as simple as possible.

Truth functions obviously may not be causal; they may for example be definitional, classificatory, or expressions of such non-causal regularities as the relation between a circle's radius and its area. Yet they also may have causal content. For instance, Quine-McCluskey minimization has historically been primarily an algorithm used in designing electrical circuits (e.g., Roth and Kinney 2004) — a context that generates an obvious causal meaning for the truth function, in that the conditions that make the function true become the conditions that cause a signal in the circuit. Thus, causal functions capture at least some kinds of causal structure.

What, however, can be said about the range of contexts in which Quine-McCluskey minimization of truth functions will produce meaningful causal inferences? Prior work has determined some key boundaries. QCA can only be useful with causal relations that are known to be identical across cases, deterministic, and that involve only known variables that are measured without error (see, e.g., Seawright 2005). If the causal relations in question are not identical across all cases, then there is no single truth function that can fit all cases, and hence empirical research will eventually identify a contradiction and thereby violate the scope conditions for Quine-McCluskey minimization. If the relationships in question are fundamentally probabilistic, then the truth function is an inadequate representation of the process: it is a model that allows for no randomness. Finally, if relevant variables are omitted from the analysis or measured with error, then the original truth function is not in fact true — and therefore the minimization will not be true either.

Some further conditions are relevant, as well. First of all, because truth functions are built around the mathematics of logical equivalence, the relations that they represent have no inherent directionality. This is in obvious contrast with causation, which involves inherently directed relationships between cause and effect. This directionality needs to come into a QCA analysis from some other source, since Quine-McCluskey minimization has no tools with which to test for it.

That is to say, QCA can only be proven to work for causal inference when the direction of all causal relationships can be demonstrated using other forms of evidence.[7]

A final point involves an issue that QCA scholars broadly recognize, but which has surprisingly far-reaching implications.The analytic results that formally justify Quine-McCluskey minimization involve complete truth functions, i.e., functions for which the truth outcomes associated with all possible combinations of conditions (equivalent to independent variables) are known. A truth function that gets some of these combinations wrong (because they were not observed in the data) is not in fact true, and a minimization of an untrue truth function will also be untrue. Ragin recognizes this issue, discussing it as the problem of "limited diversity" (Ragin 1987: 104-13; Ragin 2000: 81-87; Schnieder and Wagemann 2012: 151-77). What is the proposed solution? Effectively, to guess: scholars should use any knowledge they have in hand to decide what they think would have happened in unobserved combinations — or, if that is too much to ask, they can simply let QCA software fill in guesses for them with the criterion of minimizing complexity. It seems uncontroversial to assert that neither of these procedures will be infallible.

Because of these many issues, the properties of Quine-McCluskey minimization when applied to a truth function that is only approximately correct become relevant. Unfortunately, the original proofs for this procedure do not consider the properties of minimized approximately correct truth functions. Since the classic proofs in this domain do not apply to QCA in practice, does that mean that QCA cannot be warrantable? It does not, but it does mean that some other source of knowledge is needed to meet that standard. It is worth noting that many standard techniques in quantitative social science, including regression analysis, logit/probit models, LISREL-type models, and matching similarly would have to look beyond analytic results for evidence of warrantability when applied to observational studies; in general, theorems justifying these methods for causal inference rely on assumptions that are not known to apply — and indeed are often essentially known not to apply — to non-experimental research contexts. Hence, while QCA cannot really draw much credibility from analytic results, it is far from unusual in this limitation.

[7] This point is not unique to QCA; regression analysis similarly lacks inherent directionality, and any information about causal direction in a regression study must come from the research design or other outside sources.

Simulations

When analytic results are unavailable or inapplicable, statisticians and methodologists routinely use simulation evidence, often called Monte Carlo studies, to discover the properties of a method. The idea is straightforward. A simulation creates a situation in which the scholar knows everything about the data, including underlying causal structures and parameter values. Thus, the simulation gives the scholar a solid baseline against which to judge the performance of a particular method or estimator: if the method produces results that do a good job of matching the known truth, then it works well. Otherwise, it fails the test. Many quantitative methods owe their main practical justification to simulation results (Robert and Casella 2004). Simulation thus deserves consideration alongside analytic results as a major means of showing that a method is warrantable.

Simulations have to date played a limited role in discussions of QCA. Hug (2013), for example, uses simulations to show that measurement error can lead QCA to badly misrepresent the relationships in the data, and Lucas and Szatrowski (2013) use similar techniques to demonstrate (among a wealth of other results) that findings of causal asymmetry in QCA are a methodological artifact of setting consistency thresholds greater than 50. From a more supportive point of view, Marx (2010) relies simulations to show that QCA can distinguish between purely random data and data with a meaningful structure as long as the analyst pays attention to the proportion of combinations of conditions for which cases show contradictory outcomes, and Ragin and Rihoux (2004: 23) supportively discuss a hypothetical simulation in which QCA is applied to two random subsamples from a given initial sample. Thus, simulation techniques have been used by both critics and advocates of QCA; if QCA works in some context, simulations can provide supporting evidence and help identify the relevant context.

At the same time, existing simulation studies suffer from important limitations. Several of them focus on very specific issues and do not address the larger question of whether QCA in fact produces meaningful results in contexts of limited diversity, and perhaps omitted relevant variables. The specifics vary across studies: Hug's work addresses measurement error; Marx's simulation looks only at purely random data, showing that consistency tests can often prevent what would necessarily be purely spurious findings with such data, but never asks whether QCA findings for more structured data in fact capture the right relationships among variables; and Ragin and Rihoux are interested only in discussing the logical relation between subsample QCA results and the results for the sample as a whole. While these studies advance methodological knowledge of QCA in important ways, they barely speak to the central issue of warrantability: is there good reason to think that QCA will produce correct findings in any plausible research context?

Lucas and Szatrowski use a series of fascinating simulation studies to address exactly this issue, showing time and again that QCA results fail to replicate the relationships among variables in the genuine data-generating process. However, as far as I can tell, only one of these simulations uses a deterministic, QCA-style data generating process — and that appears to have been a one-shot simulation and thus vulnerable to criticisms that its results may have been unrepresentative of the performance of the method in general. In the remaining simulations, data are generated using a logit model. These results provide the valuable insight that QCA fails, often quite badly, when asked to analyze data that are generated through some process other than Boolean truth functions. However, QCA supporters may complain with some justification that the simulations simply do not take their ontological commitments seriously. Simulation studies are needed that speak to issues of warrantability, as Lucas and Szatrowski do, but that follow Ragin and Rihoux's methodological lead in basing the analysis in a QCA-style truth table as the true data-generating process.

The previous section raised a number of issues that make practical applications of QCA potentially unlike the motivating proofs for Quine-McCluskey minimization. This section will explore these issues via a simulation analysis, as an illutration of the kind of simulation evidence that can be used to determine a method's warrantability. In the spirit of taking seriously the views of users of a given method when evaluating that method, a key concern for the concept of warrantability, it seems fitting to focus on the one issue that is most widely agreed to be important for QCA: limited diversity, or the problem of determining the outcome for unobserved combinations of independent variables.

The simulations will involve creating data from a known causal formula, and then applying QCA to it in order to determine how close the method comes to recovering the original causal formula. By way of comparison, classification and regression trees (CART) will be applied in each instance, as well. CART is a machine-learning algorithm that involves iterative splits of a data set according to values on a collection of predictor variables, with the divisions selected in order to maximize homogeneity on an outcome variable (Breiman et al. 1984).[8] The result is a series of nodes dividing cases by their value on a single independent variable at a time. When such nodes are connected to form a tree, the result is a description of the data in terms of configurations of values across the independent variables that best sort the cases with respect to the outcome. Thus, while the computational details of CART are substantially different from those of QCA, the final result is quite comparable.

In order to isolate this issue of limited diversity from other possible challenges, the simulation reported here will adopt standard QCA assumptions essentially across the board. That is to say, it will involve deterministic causal relations, with a strictly Boolean functional form, that are constant across all cases, among variables measured without error. Furthermore, the direction of causality will be taken to be known in advance. This is carried out by calculating the complete truth table implied by the Boolean function:

$$Y = V_1 + V_2 * v_3 * V_4 + V_3 * V_5 \tag{2}$$

The truth table also includes three variables which are not in the Boolean function that serves as the data generating process, and which are therefore causally irrelevant by design: V_6, V_7, and V_8. The key test in the simulation is to determine whether it is likely that QCA will identify the three genuinely relevant combinations of independent variables, while not including the three irrelevant variables in the solution.

For a given iteration of the simulation, a random sample with replacement of a given size is drawn from that truth table. This means that, in these data, the true data-generating process for Y is just exactly that given in the equation above. Furthermore, there is no measurement error and there are no contradictions. The only messy issue in the data is limited diversity, a problem

[8] Newer extensions of CART, including random forests and Bayesian variants of the technique, may perform even better than CART (Hastie, Tibshirani, and Friedman 2009: Chapter 15). Thus, results below for CART provide a baseline for what machine learning can do in comparison with QCA; CART is conceptually and computationally simpler than the newer variants, and so it is used here.

— as discussed in the previous section — that QCA advocates and critics all agree is widespread in practice.

Each simulated data set is then subjected to Boolean minimization, implemented using the eqmcc procedure within the QCA library in R (Thiem and Dusa 2013). All unobserved combinations of conditions are treated as logical remainders; this is equivalent to asking QCA to produce the most parsimonious solution. Because the data contain no contradictions, consistency parameters are irrelevant. However, the simulation does require a rather complex approach to the minimum number of cases with a given combination of scores on the independent variables in order for that combination to be treated as meaningful. For small sample sizes, setting the minimum number of cases within a combination greater than one tends to delete all the cases. On the other hand, for larger samples, setting this number less than two frequently results in exponentially explosive computation costs for the Boolean minimization process.[9] Hence, the simulation adopts the conditional approach of setting this threshold at two cases unless that deletes most or all of the data, in which case the fallback threshold of 1 is used. This pragmatic solution seems in keeping with guidance from QCA methodologists, who encourage a higher threshold but obviously do not hold scholars to that rule if the total number of cases is small.

Sample sizes are set to every multiple of 10 between 20 and 100. For each sample size, 1000 independent samples are drawn from the truth table. Figure 1 shows the results. The top line in the figure shows QCA's success rate at identifying the stand-alone sufficient cause, V_1, as part of the final causal solution. Evidently the method works well for such simple causal patterns, although there is a dip in the middle of the plot representing a tendency for medium-sized samples to over-complexify V_1's relationship with Y.

The results are far less promising for the second two causal combinations, $V_2 * v_3 * V_4$ and $V_3 * V_5$. Success rates for these two combinations vary in a range from about 35% to about 70%, but mostly fall near 40%. Across the simulation as a whole, only 20% of analyses correctly identify both of these combinations. These unimpressive results are made all the more troubling by the fact that 30.4% of analyses include at least one irrelevant variable in the final minimized solution. In

[9] One error message explained that the code could not continue because 26 gigabytes of RAM would be required.

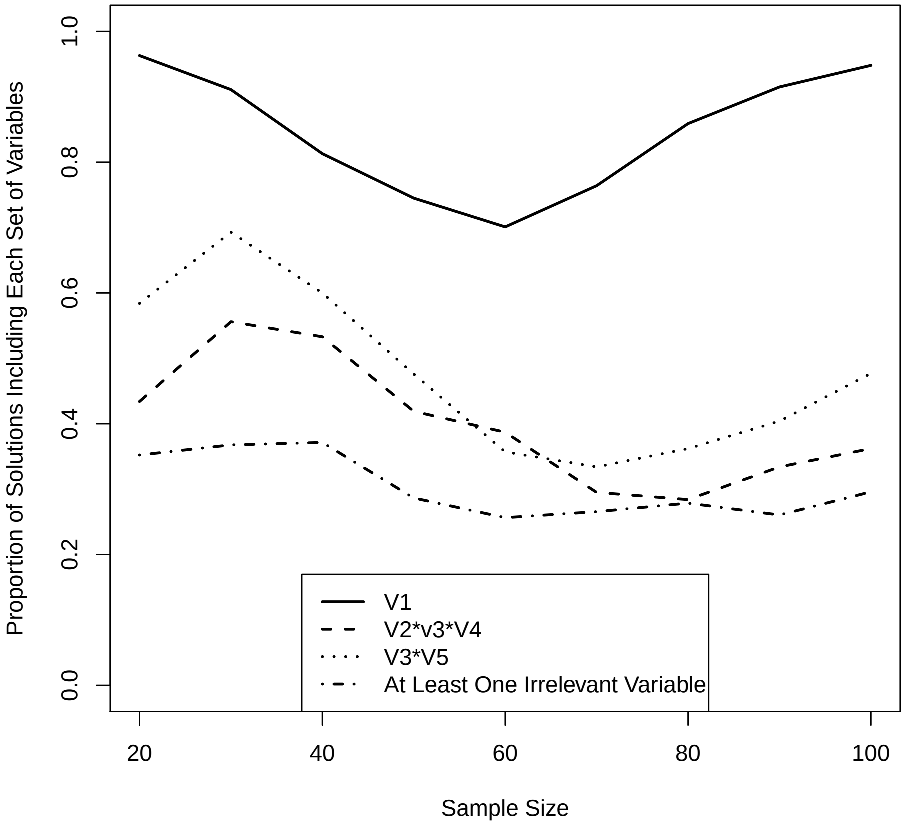

Figure 1. Baseline QCA Success Rates.

other words, given the reality of limited diversity, QCA is in practice not particularly powerful at detecting genuine causal complexity, yet it is fairly prone to false-positive results.

Two plausible objections deserve attention at this point. First, the simulation was set to find the most parsimonious solution. Does this perhaps bias the results against QCA? In fact, the answer is no. On the one hand, less parsimonious solutions are more likely to include irrelevant variables — a problem that is already substantial in these results. On the other hand, less parsimonious solutions sometimes pick up correct clusters that more parsimonious solutions miss, but they also sometimes overcomplexify clusters that the parsimonious solution gets right. Preliminary simulation work suggests that the overall picture is not substantively different if a less parsimonious solution is used.

Second, it could be claimed that QCA requires analysts to use case knowledge to correctly discern the true outcome associated with combinations of conditions that cannot be observed due to limited diversity — and therefore that these results are irrelevant in practice. This argument would of course be an overstatement. QCA permits analysts to assign values to unobserved combinations, but it certainly does not require this (see Ragin 1987: 104-13). Furthermore, analysts in practice do not always make any attempt to use substantive knowledge as a way of filling in missing configurations. To take a recent and representative example, Avdagic's (2010) fs-QCA analysis of the emergence of European social pacts uses the QCA software's two default settings ("complex" and "parsimonious" assumptions about unobserved causal conditions) for dealing with limited diversity — and makes no use of theory or case knowledge to fill in outcomes for unobserved combinations.

Applied researchers' evident reluctance to assign outcomes to hypothetical cases seems sensible enough. Nonetheless, it is true that, if the data-generating process was of a QCA character as in these simulations and if analysts correctly fill in unobserved configurations, QCA will produce the right answer. While this is true, it is also essentially meaningless. If an analyst knows all of the outcomes associated with observed and unobserved combinations in an a priori way, then the analyst obviously knows everything about the causal process of interest before QCA is even used. To say that QCA works whenever the researcher knows this much to begin with is to simply sidestep the central issue of warrantability: can a contested — not a settled — claim ever be reasonably believed on the basis of results from this method? These simulation results suggest that, even in best-case scenarios, prospects are rather grim.

While QCA's warrantability appears, on these simulation results, to be extremely limited, one might noneless imagine a "best of a bad lot" defense. The idea would be that Boolean truth functions are simply very hard causal structures to investigate and that QCA, for all its evident frailties, does better with these data generating processes than other methods would. To test the plausibility of this line of argument, I replicate the simulation above using CART in place of QCA.

Results of the simulation for CART are shown in Figure 2. Several features of the findings deserve discussion. In the first place, the fact that all three lines for relevant combinations of variables collapse to the same value for samples of size 20 implies that CART was totally unable

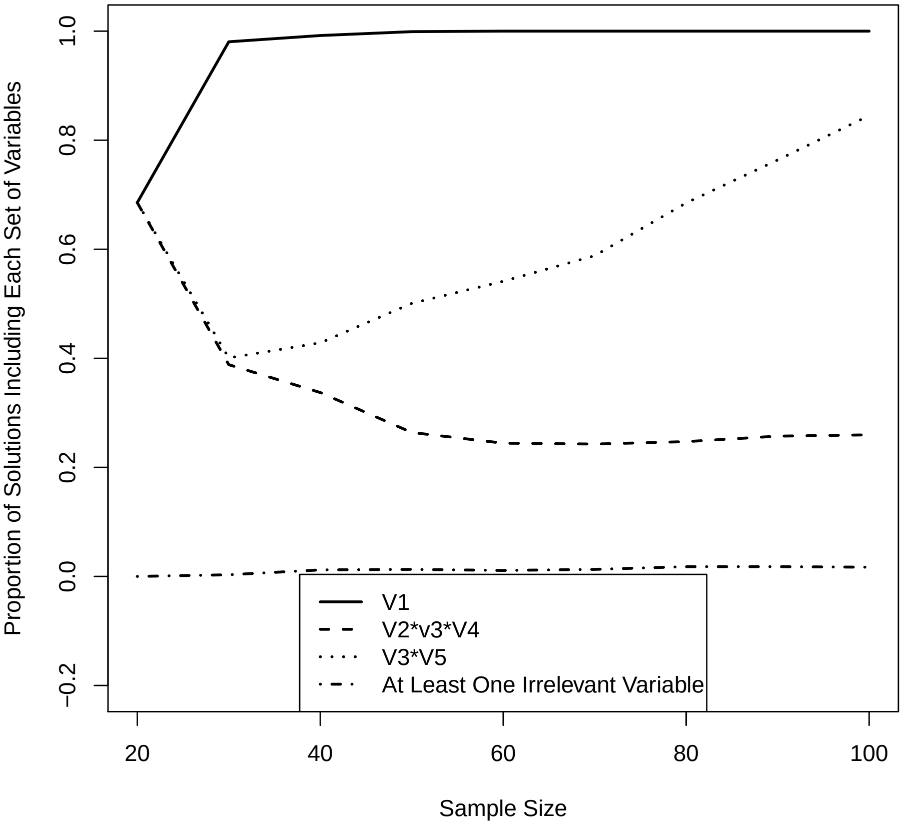

Figure 2. Baseline CART Success Rates.

to make sense of the data for this small sample size. For samples of size 30 or larger, CART produces differentiated solutions, and is generally reasonably successful in comparison with QCA. For sample sizes 30 and 40, QCA is somewhat more likely than CART to identify the third causal configuration, but for larger configurations CART clearly out-identifies QCA on this combination. For samples of size 30 and above, CART also clearly outperforms QCA in identifying the simple, stand-alone sufficient cause, V_1. With respect to the second and most difficult causal combination, $V_2 * v_3 * V_4$, QCA is persistently, but not substantially, more likely to correctly identify this cluster.

Yet these differences are really swamped by the contrast between QCA and CART in terms of the probability of including at least one irrelevant variable in the final solution. For QCA, this probability fluctuates around 30%, but for CART — using exactly the same simulation process —

the probability is pegged under 2% for every simulation condition. Simply put, even in this ideal test case for QCA, CART findings are always far more credible. QCA finds solutions at a sample size of 20, while CART does not, and QCA is somewhat more likely to capture the second causal configuration across the board. Yet these gains are counterbalanced by at least a 13-fold increase in the probability of false-positive findings. Indeed, the probability of false-positive findings is so high that any skeptic of any QCA finding is always justified in rejecting the findings as plausibly false positives; this is clearly not the case for CART.

These results become even starker if we adopt the more stringent version of warrantability and require methods to be resilient to at least modest effects of missing variables. To allow for a very small amount of missingness, the above simulation is repeated except that 5% of cases are drawn from a truth table generated by the following Boolean equation, rather than the one introduced earlier:

$$Y = V_1 * V_9 + V_2 * v_3 * V_4 * v_9 + V_3 * V_5 * V_{10} \tag{3}$$

The two new variables, V_9 and V_{10} are unobserved by the researcher and hence omitted from the QCA analysis. The two variables are also independent of all the included independent variables. How well do QCA and CART perform in the face of this very minor admixture of missing variable influence? Figures 3 and 4 show the results.

CART is essentially unaffected by the slight admixture of cases with omitted variables; the results are qualitatively and quantitatively almost impossible to distinguish from the simulation with no missing variables. QCA on the other hand is badly affected by even a small amount of missingness. The proportion of analyses that successfully identify the stand-alone sufficient cause drops substantially, and QCA is now more likely to include an irrelevant variable in the final causal equation than to identify either of the two causally complex conjunctions.

These results show how warrantability can be tested, and sometimes rejected, using simulation methods. Perhaps surprisingly, the results suggest that CART methods are arguably warrantable for sample sizes of 30 or larger — at least when there is some reason to think that a QCA-type causal structure characterizes the research domain of interest. After all, these methods do a fair job

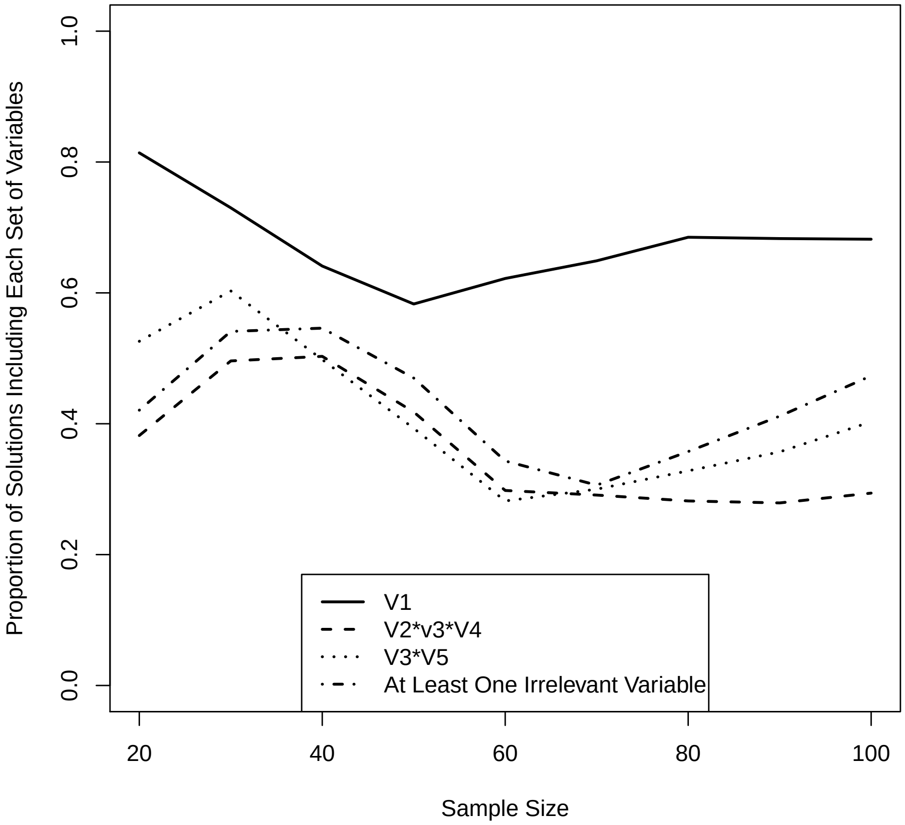

Figure 3. QCA Success Rates with Omitted Variables in 5% of cases.

of recovering a kind of causal structure that a substantial research community — QCA scholars, among others — believes to be prevalent in the real world while running almost no risk in this context of producing false positives. By contrast, the simulations fail to establish warrantability for QCA: the method produces false positives at a rate similar to its genuine findings for causally complex combinations.

These examples show how simulation can be used to test warrantability, by providing a research context that allows methodologists to ensure that the causal situation of interest to advocates of a given method in fact holds, and permitting a correspondence check between the method's results and the known causal facts of the matter in the simulated world. If the results correspond reasonably well and false positives are rare, then the method is warrantable; otherwise, other kinds

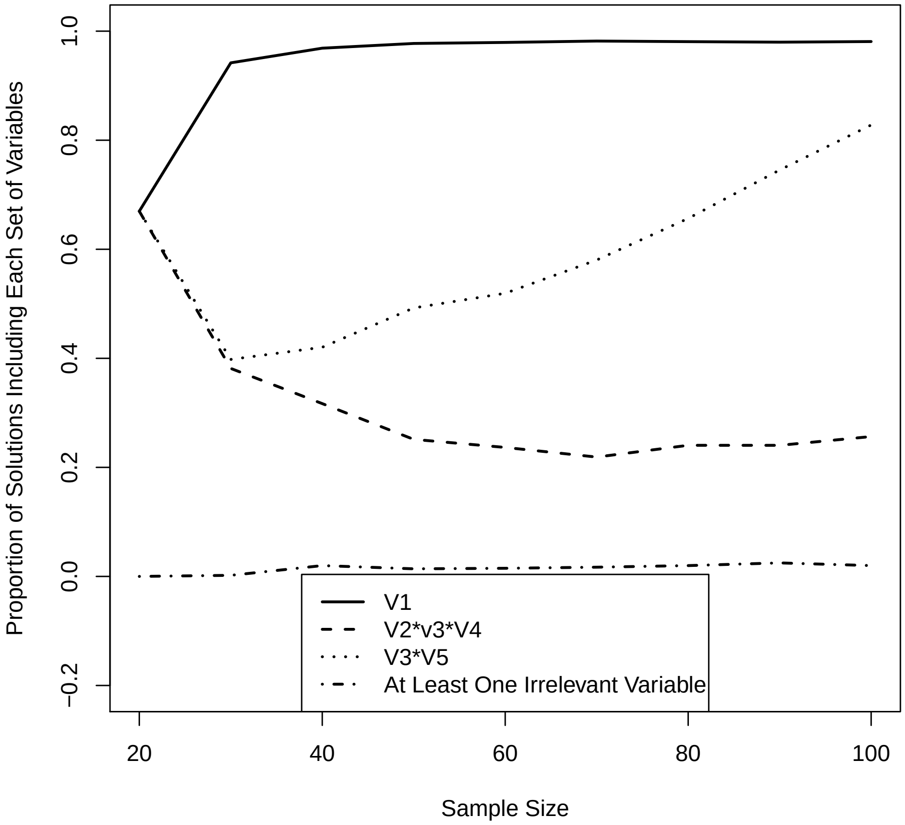

Figure 4. CART Success Rates with Omitted Variables in 5% of cases.

of evidence would be needed to make a case for warrantability. While CART looks good in the simulations presented above, it is important to note that further inquiry could raise problems for it, as well. It appears to be a better method than QCA in that its use seems more warranted in the very situations for which QCA was designed; however, there are certainly situations that cause trouble for CART as well.

Replication

For some methods, neither proofs nor simulations provide evidence for warrantability — but such methods may still be warrantable if they have a proven track record of replicating known causal findings from other studies. This is the logic, for example, in the Lalonde et al. studies attempting to use observational studies to replicate experimental results, and also for Cook's similar study.

Advocates of QCA have argued for the method's usefulness largely via an unhelpful variant of this approach. Whereas the studies discussed in the last paragraph attempt to replicate findings that are known to be causal because of experimental research design, QCA methodologists have generally attempted to provide evidence for their technique by replicating observational studies (e.g., Ragin 1987: 125-63; Ragin 2000: 120-45; Schneider and Wagemann 2012: 178-90) — whose results are obviously not known to be causal. When the original study that provides a benchmark in a replication analysis is not known to be causal, the only objective conclusions to be drawn involve whether the new technique produces the same results. Nothing can be said about whether the new technique captures the *right* results, the key issue for warrantability.

Hence, to demonstrate warrantability, QCA would have to be compared with known, causal findings. This is best done by benchmarking QCA against experimental results. One important issue complicates this task: experiments are generally designed to estimate the sample average treatment effect via a difference in means (Neyman 1923/1990; Rubin 2005), whereas QCA emphatically does not estimate a sample average treatment effect. Instead, QCA provides what purports to be a full causal formula for the outcome. Much has been made of the incomparability of these causal quantities (Goertz and Mahoney 2012: 75-82) — in fact, too much has been made of this. Indeed, a direct translation from QCA results to an average treatment effect is possible.

The reason is that a full causal formula logically implies a sample average treatment effect. The average treatment effect can be calculated from QCA results as follows. Create a copy of the data in which every case is reassigned to the treatment group. Then use the QCA formula to figure out the outcome each case should have if in the treatment group. Repeat the process for a copy of the data in which every case is reassigned to the control group. Calculate the average of the two outcome vectors and take the difference. If QCA is recovering real causal formulas to some useful degree of approximation, then this process should recover a value that corresponds fairly closely with the unbiased estimate of the average treatment effect given by a difference in means estimator on the experimental data.

To try this out, I use data from Mutz's (2007) study of the effects of uncivil discourse on attitudes about the political opposition. Specifically, I look at the effect of being randomly assigned to watch a civil or an uncivil debate between fictional candidates on research subjects' thermometer ratings of their less-preferred candidate.[10] For the experimental results, the estimator is just a difference in means between civil and uncivil treatment groups; for QCA, I ask the Boolean minimization process to deal with a data set including the outcome, the treatment, and all other available variables (respondents' income, education, sex, degree of interest in politics, partisanship, feeling thermometer rating toward their preferred candidate, and evaluation of the legitimacy of the arguments on their own side and on the other side).

For the original data, the difference in means estimates a sample average treatment effect of 0.15; for a dependent variable rescaled to be bounded between 0 and 1, this is a substantial effect. QCA, by contrast, estimates a sample average treatment effect of 0.06. While this marked underestimate is not promising, it is of course possible that it is a fluke of this particular sample.

To test for that possibility, I bootstrap (Efron and Tibshirani 1994) both analyses. Out of 1000 bootstrapped subsamples, QCA was unable to provide an effect estimate due to excessive numbers of contradictions in fully 486 iterations. The difference in means estimator, by contrast, was available for all 1000 iterations. Histograms of the resulting sample average treatment effect estimates for both methods are provided in Figures 5 and 6. These results confirm that the difference in means is a well-behaved estimate, with a distribution centered at the true value in the sample from which these subsamples are drawn and with modest spread consistent with a result that is significantly different from zero, although not overwhelmingly so.

The QCA results, by contrast, reflect a remarkably badly behaved estimate. More than half of the subsamples that QCA was able to analyze at all returned treatment effect estimates at or very near zero. The rest were spread quite broadly and seemingly arbitrarily up to an estimate of 0.6. In other words, QCA produces results that imply a typically quite incorrect sample average treatment effect for these data and that are remarkably variable. This attempted replication, at least, clearly provides no evidence that QCA is warrantable.

[10] Much may depend on measurement and variable scaling. To put the methods on as equal a footing as possible, I simply linearly rescaled all included variables to range from 0 to 1 and used fs/QCA to calculate QCA estimates of the sample average treatment effect.

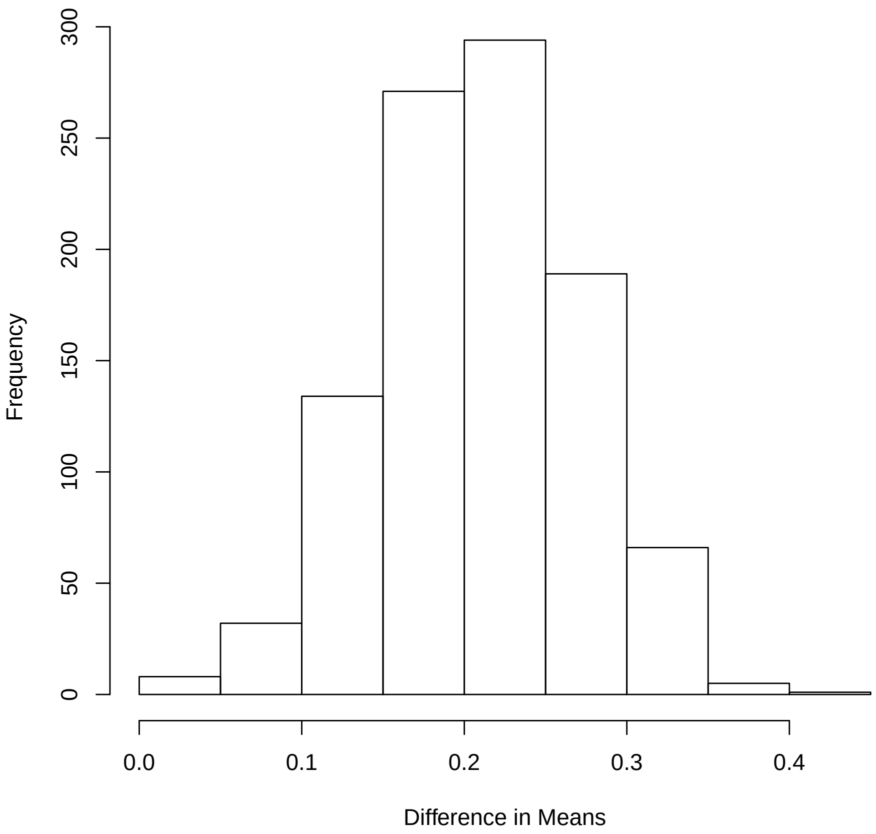

Figure 5. Sample Average Treatment Effect Estimates, Difference in Means.

Of course, much may depend on the details. Perhaps different scaling decisions for the variables, the inclusion of other independent variables in the analysis, and so forth could have led QCA to behave better as an estimate of the sample average treatment effect. Or perhaps there are other experimental data for which QCA turns out to function more reliably. An analysis demonstrating that such was the case would go a long way toward establishing QCA as a warrantable method. For the time being, unfortunately, this must stand as yet another piece of evidence that QCA is not at present warrantable.

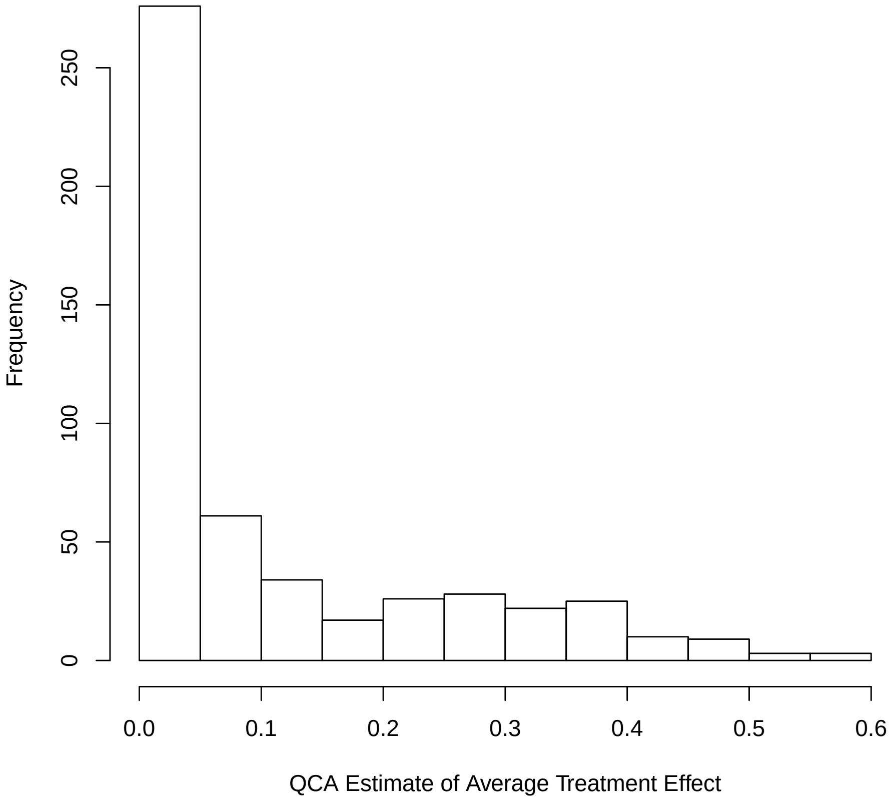

Figure 6. Sample Average Treatment Effect Estimates, QCA.

Conclusions

The standard of warrantability is thus a sufficiently rigorous one to reject QCA, given current knowledge about the method. For a wide range of other methods, the issue of warrantability needs investigation and perhaps debate. I have suggested above that regression in the specific context of true experiments and some forms of process tracing are warrantable. Furthermore, there is some reason to think that CART may be warrantable in the context of some QCA-like causal scenarios. However, for a wide range of commonly-used qualitative and quantitative research tools, questions of warrantability are at least unclear. The implication is that at least some research communities are likely to be using methods that fail to pass the relatively low standard of warrantability

In such future discussions, the kinds of evidence discussed above — as well as other forms of knowledge about the methods in question — will no doubt play important roles. Yet two forms of argument that do *not* help resolve questions of warrantability are common enough to bear separate mention.

First, hypotheses about the nature of causality have limited bearing on discussions of warrantability. For example, QCA advocates sometimes argue that important strands of causation in the social world simply do take on set-theoretic or Boolean-algebriac forms (e.g., Ragin 2000: 88-119; Ragin 2008:13-14). Even granting this assertion, though, scholars still need evidence that QCA can recover true causal patterns when those patterns look like Boolean equations. The simulation evidence above implies that this issue is nontrivial. Questions of warrantability are about whether a method can reliably recover a given kind of causal information, not whether the kind of causal information associated with a method is particularly common in the social world; hence, debates on the latter point are not only difficult to resolve but also ultimately somewhat beside the point.

Second, as discussed earlier, examples of a method's use in the context of observational studies are not helpful for thinking about warrantability. The reason is that, with such examples, the true causal pattern is really unknown — and therefore there is no standard against which to check the results of any given method. Really, the best that can be said of such results is that they are interesting and plausible. While it is worth knowing whether a method can produce interesting and plausible results, this is a fundamentally different question from whether the method is warrantable and hence worth relying on. Thus, illustrative examples of QCA without a meaningful benchmark (e.g., Ragin 2000:123-39; Schneider and Wagemann 2012: 135-38) really do nothing to resolve questions of warrantability; such examples will be equally unhelpful in discussions of other methods.

By contrast, analytic results make a credible case that regression produces good causal inference in a well-designed and -executed laboratory experiment. Simulation data provides grounds for optimism about CART with respect to Boolean-type causal patterns. Further analysis of this sort will no doubt find qualitative techniques that can likewise play a meaningful if delimited role in methodologically pluralistic debates about causation. The idea of warrantability does not rule any specific qualitative techniques in or out of discussion; instead, it suggests a research agenda that advocates of such techniques should use as a way of justifying their preferred technique.

References

Alston, W. P. (1989). *Epistemic justification: Essays in the theory of knowledge.* Cornell: Cornell University Press.

Avdagic, S. (2010). When are concerted reforms feasible? explaining the emergence of social pacts in western europe. *Comparative Political Studies*, *43*(5), 628-57.

Bonjour, L. (1980, Sept.). Externalist theories of empirical justification. *Midwest Studies in Philosophy*, *5*, 53-74.

Brady, H. E., & Collier, D. (Eds.). (2004). *Rethinking social inquiry: Diverse tools, shared standards.* Rowman and Littlefield and Berkeley Public Policy Press.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees.* Boca Raton: Chapman and Hall/CRC.

Cohen, S. (1984). Justification and truth. *Philosophical Studies*, *46*(3), 279-95.

Cohen, S. (2002, Sept.). Basic knowledge and the problem of easy knowledge. *Philosophy and Phenomenological Research*, *65*, 309-29.

Daston, L., & Galison, P. (2007). *Objectivity.* Brooklyn: Zone Books.

Dunning, T. (2012). *Natural experiments in the social sciences: A design-based approach.* Cambridge: Cambridge University Press.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap.* Boca Raton: Chapman and Hall/CRC.

Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, *40*, 180-93.

Gaukroger, S. (2012). *Objectivity: A very short introduction.* Oxford: Oxford University Press.

Goertz, G., & Mahoney, J. (2012). *A tale of two cultures: Qualitative and quantitative research in the social sciences.* Princeton, N.J.: Princeton University Press.

Goldman, A. (1999). *Knowledge in a social world.* Oxford: Oxford University Press.

Goldman, A. (2011). Reliabilism. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy.* http://plato.stanford.edu/archives/spr2011/entries/reliabilism/.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer.

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

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research.* Princeton: Princeton University Press.

Lehrer, K. (1990). *Theory of knowledge.* Boulder: Westview Press.

Marx, A. (2010, Aug.). Crisp-set qualitative comparative analysis (csqca) and model specification: Benchmarks for future csqca applications. *International Journal of Multiple Research Approaches*, *4*, 138-58.

Mutz, D. C. (2007, Nov.). Effects of "in-your-face" television discourse on perceptions of a legitimate opposition. *American Political Science Review*, *101*, 621-35.

Neyman, J. S. (1923/1990, Nov.). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, *5*(4), 465-72.

Novick, P. (1988). *That noble dream: The "objectivity' question" and the american historical profession.* Cambridge University Press.

Plantinga, A. (1993). *Warrant and proper function.* Oxford: Oxford University Press.

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

Ragin, C. C. (2000). *Fuzzy-set social science.* Chicago: University of Chicago.

Ragin, C. C., & Rihoux, B. (2004, Fall). Qualitative comparative analysis (qca): State of the art and prospects. *Qualitative Methods*, *2*, 3-13.

Robert, C. P., & Casella, G. (2004). *Monte carlo statistical methods.* New York: Springer Science+Business Media.

Roth, J., Charles H., & Kinney, L. L. (2004). *Fundamentals of logic design.* Stamford: Cengage Learning.

Rubin, D. B. (1980). Discussion of basu's "randomization analysis of experimental data: The fisher randomization test." ". *Journal of the American Statistical Association*, *75*, 591-93.

Rubin, D. B. (2005, March). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 322-31.

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

Seawright, J. (2005, Spring). Qualitative comparative analysis vis-a-vis regression. *Studies in Comparative International Development*, *40*, 3-26.

Thiem, A., & Dusa, A. (2013). *Qualitative comparative analysis with r: A user's guide.* New York: Springer.

Vogel, J. (2000, Nov.). Reliabilism leveled. *Journal of Philosophy*, *97*, 602-23.