Qualitative & Multi-Method Research

Newsletter of the
American Political Science Association
Organized Section for Qualitative and Multi-Method Research

Spring 2014, Vol. 12, No. 1

QCA and Causal Inference: A Poor Match for Public Policy Research

Sean Tanner
University of California, Berkeley
stanner@berkeley.edu

Qualitative Comparative Analysis (QCA) offers distinctive research tools that, according to its practitioners, yield a productive solution to many problems and limitations of conventional quantitative methods. QCA is claimed to combine the strengths of the qualitative and quantitative traditions and to yield distinctive leverage for causal inference.

Among diverse avenues available for evaluating any given method, one approach is close examination of its contribution to the study of a particular substantive area. Such evaluation is especially appropriate if proponents of the method argue that it is indeed highly relevant to that domain.

In fact, proponents of QCA have championed this method as a valuable tool for public policy research,[1] arguing that it is "extremely useful" and has "intriguing potential" for policy analysis.[2] They advance a number of specific arguments about its relevance for policy studies: QCA focuses on set-theoretic relationships, uncovers multiple conjunctural causation, and allows flexible causal modeling (Rihoux et al. 2011: 16–17).[3] A further premise is that the method moves beyond the constraints of causal assessment based on "net effects thinking" to consider more complex interactions among explanatory variables (Ragin 2010: 16–24; Schneider and Wagemann 2012: 83–89).

How should these claims be evaluated—especially the central argument that QCA's approach to causal analysis is especially productive for policy studies? Public policy research obviously encompasses diverse areas, and some of them—for example the politics of policy formation—present analytic challenges relatively similar to those encountered in a broad spectrum of political science topics. A claim by QCA of distinctive value for studying the politics of policy formation would thus be equivalent to a general argument that the method is relevant for political science. Developing such an argument would of course be perfectly appropriate, but it may not capture this idea of the method's special relevance to policy studies that is advanced by QCA scholars.

In fact, something distinctive *is* indeed at stake here. In policy studies, the place where "the rubber hits the road" in terms of causal assessment is the field of evaluation research—i.e., the study of policy impacts. Policy evaluation has in re-

[1] Hudson and Kühner (2013); Rihoux and Grimm (2010); Rihoux, Rezsöhazy, and Bol (2011).

[2] Quotes are from, respectively, Rihoux, Rezsöhazy, and Bol (2011: 17); and Hudson and Kuhner (2013: 284).

[3] Claims about QCA's relevance to policy research are stated in somewhat different ways in other books and articles. These three attributes are the most common and salient across all of these authors.

cent years seen dramatic innovation in tools for causal inference, along with an energetic search for new methods that advance key inferential goals.

Hence, it is valuable to ask: does QCA's distinctive approach to causal assessment help meet the goals of an area of policy analysis that is especially concerned with valid causal inference? Does the method provide special leverage that addresses the concerns of the evaluation field?

These questions are all the more salient because evaluation research is a prominent focus in leading graduate schools of public policy. If QCA's value-added for policy evaluation were demonstrated, this would be a key step in legitimating the method in the policy studies community.

Across the spectrum of topics in the broad field of policy studies, evaluation research is therefore a "crucial case" for assessing QCA.

Organization of the Analysis

The following discussion first examines this crucial case of evaluation research by providing a base line for comparison. Six studies are analyzed that exemplify current practices in the policy evaluation field. The focus is on the kinds of questions asked—which centrally involve causal inference—and the tools employed in answering them. It is argued that these methods deliver the kind of insights sought by policy analysts. Hence, they provide a useful basis for comparison.

It should immediately be emphasized that these six studies—and current norms for acceptable research in leading policy schools—are very different from what might be thought of as "conventional quantitative methods." The social sciences have recently seen a basic rethinking of norms about causal inference, and these norms—which will be noted at various points below—now undergird standard practice in leading schools of public policy. These six studies reflect this standard practice.

The second section of this paper examines five examples of policy evaluation based on QCA—examples that have been offered by QCA scholars to illustrate their approach to policy analysis. The discussion below asks: Do these studies organize their causal findings in a way that is useful for scholars concerned with public policy? Do they meet the norms for justifying causal claims that are standard in current policy research? Is the largely deterministic framework, central to their set-theoretic approach, productive for policy analysis?[4]

The third section raises broader questions about QCA's basic arguments and practices, as applied to policy studies. Topics addressed here include net effects, context and causal heterogeneity, the distinction between case-oriented and variable-oriented analysis, norms for causal inference, and incorporating uncertainty.

In response to this series of questions, the present analysis concludes that QCA is of questionable value for this crucial case of policy evaluation.

Two further introductory points must be underscored. First, although the central focus here is on the value of QCA for public policy research, the wider implications for the method's contribution to causal inference are also of great interest. The norms articulated here for good causal inference are in fact quite general today in the social sciences. It is therefore useful to ask whether QCA meets these norms.

Second, this evaluation of QCA is not in any sense offered from the standpoint of conventional quantitative methods—which, as just noted, is definitely not the preferred approach in policy research today. Quite the contrary, the norms of evidence and inference employed here have also been the basis for the major critique of conventional, regression-based quantitative analysis. Further, while ideas about causal inference in experiments and natural experiments are part of this rethinking, the point is definitely *not* that (a) all researchers should be doing experiments, or (b) valuable causal inferences cannot be made based on observational data. Rather, these ideas have played a productive role in a wider, multifaceted reconsideration of causal inference.

In sum, given this fundamental rethinking of methods, the overall question here is two-fold: does QCA yield valuable substantive findings for policy researchers, and also for social scientists in general?

Policy Evaluation with Standard, Current Methods

The effects of government action are often small, and relatively modest impacts can be of great interest to policy makers. Since the first schools of public policy were founded in the late 1960s, conventional policy analysis has rested on tools that effectively and directly yield information on these impacts (Allison 2006: 68). Policy research is also attentive to contextual effects, subgroup differences, and interactions in the impact of policies—phenomena that are effectively addressed within the conventional analytic framework. To anticipate the discussion, the six examples that serve to illustrate these arguments are listed in Table 1.

To begin with a simple example: Angrist et al. (2012) exploit a random lottery to find a modest but palpable impact of charter schools on student reading scores. The effect is not large, yet other research (Chetty, Friedman, and Rockoff 2013; Hanushek, 2011) finds that differences of this magnitude are associated with substantial increases in lifetime earnings. Identification of this average partial (or "net") effect of charter schools is therefore an important insight for research on education policy.

The concern with how policy affects disadvantaged groups is a recurring theme. For instance, with the introduction of new teacher performance standards in North Carolina, student math scores increased, overall, by only a modest amount. Yet strikingly, the effect is largest for the lowest performing students (Ladd and Lauen 2010). Again, this magnitude of gain is predicted to yield an appreciable increase in lifetime earnings—a matter of enormous policy relevance, given the frequent failure of the U.S. education system in improving the success of disadvantaged students (Hanushek 2003).

By contrast, in another domain the more at-risk population is *not* similarly advantaged. Sen (2012) finds that people

[4] QCA can contain some probabilistic elements, such as quasi-necessity and quasi-sufficiency, but the framework is still largely deterministic.

Table 1: Overview of Studies Based on Standard, Current Methods

Study	Substantive Focus	Type of Analysis
Angrist et al. 2012	Charter Schools	Random Lottery
Ladd and Lauen 2010	Teacher Performance Standards	Fixed Effects Regression
Sen 2012	Gas Prices and Exercise	Fixed Effects Regression
Reardon et al.2012	School Re-segregation	Interrupted Time Series
Mauldon et al. 2000	Educational Attainment of Teen Mothers	Randomized Control Trial
Datar and Nicosia, 2012	School Nutrition	Instrumental Variables Regression

tend to get more physical exercise—a desirable health outcome—when gas prices increase, but that this effect is quite heterogeneous across socioeconomic status. On average, a dollar increase in gas price increases exercise by 2.4 percent. However, there was no detectable increase for the lowest socioeconomic group,[5] whereas for the middle income group the increase is 3.7 per cent (Sen 2012: 357). This suggests that a gas tax is unlikely to affect the physical activity of those people comprising the lowest socioeconomic—and also the least-healthy—group.

A context-dependent effect uncovered by Reardon et al. (2012) is of great salience to analysts concerned with the impact of court decisions on public policy. From the early 1990s to the present, Southern school districts re-segregated far more than their Northern counterparts, after being released from desegregation orders. This trend is likely to be highly consequential, given that desegregated school districts have improved the long-term income and health of African-American students (Johnson 2011).

Though each of the studies focuses on one intervention, or "treatment," policy researchers additionally care about interactions among interventions. If a given policy has two components, analysts routinely ask if either is valuable, if one is more valuable than the other, and whether they are most effective when pursued jointly. Mauldon et al. (2000) is an excellent example of research addressing such interactions. The authors conduct a social welfare experiment seeking to promote high school completion for teenage mothers. In the experiment, some mothers receive financial incentives for pursuing further education, some receive case management, some receive both, and some receive neither. The researchers find that financial incentives by themselves have a marginal effect, case management by itself has no effect, and the truly significant effect occurs when the two interventions are combined. This finding is of great interest to analysts designing future welfare policy.

Of course, not all policies produce causal effects. Datar and Nicosia (2012), for example, find that junk food availability does not increase obesity or decrease exercise in a cohort of fifth grade students. These null results have important policy consequences. As debates about school nutrition remain highly visible at the national level, having analytic tools that can establish the *absence* of an effect is of great importance.

[5] The point estimate of a .8 percent increase is not distinguishable from zero.

Summary of Standard Methods

Table 2 summarizes key features of these six studies. All of them seek to meet current, very exacting, standards for good causal inference—though certainly some succeed more fully than others. These standards are centrally concerned with potential weakness of any inferences based on observational data, and they sharply question the adequacy of naive regression analysis. Two of these articles are based on policy experiments—and they show that randomized experiments can indeed address major substantive questions. The remaining four use combinations of natural experiments and careful statistical analysis, and in all instances they employ sensitivity analysis and other simulation tools to assess the robustness of findings.

In substantive terms, policy analysts care about average partial effects and these studies directly tackle that issue. Of course, in the net effects framework, there are routinely subgroup differences and interactions, and these examples show that analysts frequently examine them to great advantage. Whether the focus is on subgroups or the full set of cases, the policy researcher cares crucially about the net impact of policies. This is the fundamental basis for embracing, modifying, or rejecting policies. Methods that evaluate net effects directly address that high priority.

Finally, these studies generally do well in in defending the plausibility of causal inferences because they explicitly discuss the treatment assignment mechanisms. Specifically, they bolster the as-if random assignment assumption required to identify plausible counterfactuals. With experiments, treatment assignment is unambiguous: random assignment is achieved by the experimental design. In other research designs, random assignment is approximated by comparing groups that would, save for the policy treatment in question, be expected to have similar outcomes. The challenge in these designs is to defend the critical assumption that the policy was differentially implemented "as-if" by random assignment. Through explicit discussion of the treatment assignment mechanism, researchers

Table 2: Detailed Summary of Studies Based on Standard Methods

Study	Substantive Focus	Type of Analysis	Size of Main Effect	Interactions/ Subgroup Differences	Analysis of Treatment Assignment	Plausibility of Causal Inference
Angrist et al. 2012	Charter Schools	Random Lottery	Medium	Greater impact for less skilled students	Detailed	Strong
Datar and Nicosia 2012	School Nutrition	Instrumental Variables Regression	None	No effect	Detailed	Moderate
Ladd and Lauen 2010	Teacher Performance Standards	Fixed Effects Regression	Small	Greater gains in the tails of the distribution	Detailed	Moderate
Mauldon et al. 2000	Education of Teen Mothers	Randomized Control Trial	Small	Best results when both policies applied	Detailed	Strong
Reardon et al. 2012	School Re-segregation	Interrupted Time Series	Medium	Greater impact in South than North	Detailed	Moderate
Sen 2012	Gas Prices and Exercise	Fixed Effects Regression	Small	No effect for lower SES group	Detailed	Moderate

bolster confidence in their causal inferences. This step is relevant and valuable, even if they are not carrying out experiments or natural experiments.

Policy Analysis with QCA

QCA scholars who recommend applying their method to policy analysis have offered many illustrations of their approach. In the framework proposed here—of focusing on policy evaluation as a crucial case—the following discussion reviews five examples that QCA scholars have identified as strong illustrations of their method, as applied to policy evaluation. Specifically:

a. Rihoux and Grimm's (2010) book *Innovative Comparative Methods for Policy Analysis* includes one chapter-length, substantive study that is offered to exemplify the method. In this chapter, Befani and Sager (2010) focus on the conditions under which environmental impact assessments will be effectively implemented.

b. Two examples are from the review essay by Rihoux, Rezsohazy, and Bol (2011). Balthasar (2006) explores the features of oranizational evaluations that lead them to be effective, and Pennings (2005) analyzes welfare expenditures. While Pennings' analysis includes macro variables, he also looks at the impacts of policies per se, including outcomes that derive from the mix of welfare policies (Rihoux et al., 2011: 31), as well as from prior policy choices about economic openness.

c. The final two examples are drawn from the symposium on QCA published in 2013 by the journal *Policy and Society*—where they were included with the goal of illustrating "the intriguing potential of QCA for policy analysis and evaluation…" (Hudson and Kühner 2013: 284). Lee (2013) evaluates the impact of alternative labor policies on patterns of employment; and Warren, Wistow, and Bambra (2013) evaluate the circumstances under which a health intervention yields the desired health improvement.

These five studies, to which QCA advocates have particularly called attention, provide a suitable comparison with the policy evaluations, discussed above, that use standard methodological tools. Further, these five appear an appropriate basis for some broader observations about QCA as a method.

As with the articles above, the main question of concern here is: Do these QCA policy studies deliver useful insights for the policy research community? Table 3 provides an overview of the five studies. The third column in the table indicates the type of QCA utilized: the dichotomous crisp-set version (csQCA), the multi-value version (mvQCA), or the fuzzy-set version (fsQCA).

To begin, Befani and Sager (2010) investigate the circumstances under which Swiss environmental impact assessments are effectively implemented.[6] Impact assessments are an enormously important aspect of environmental policy-making, and improperly implemented assessments undermine a fundamental tool of environmental regulation.

Using csQCA and focusing on 15 cases, Befani and Sager (2010) consider six conditions that may influence effective implementation: (i) a clear definition of the project being evaluated, (ii) early discussion of all relevant questions, (iii) systematic project management by the relevant public agency, (iv) early integration of all stake-holders, (v) socio-political sensitivity to environmental concerns, and (vi) size of the project.

The authors find that the 15 cases can be completely accounted for by the 12 distinct causal paths.[7] Assessments are well-implemented if there are:

[6] Implementation is defined primarily by compliance with regulations regarding environmental impact assessments.

[7] The exact number of cases in each path could not be inferred from the data presented in the article.

Table 3: Overview of Five Studies Offered by QCA Scholars as Illustrations of the Method

Study	Substantive Focus	Type of QCA
Balthasar 2006	Evaluation Use	mvQCA
Befani and Sager 2010	Environmental Impact Assessments	csQCA
Lee 2013	Employment Policy	fsQCA
Pennings 2005	Welfare Expenditures	fsQCA
Warren, Wistow, and Bambra 2013	Health Policy	cs/QCA

1. Clear project definitions and early discussion
2. Early discussion and low environmental sensitivity
3. Early discussion and a small project
4. Clear project definitions, high environmental sensitivity, and a large project
5. Clear project definitions, systematic project management, and a large project
6. Clear project definitions, systematic project management, and high environmental sensitivity

Conversely, assessments are not well-implemented if there are:

7. Unclear project definitions and a large project
8. Unclear project definitions and high environmental sensitivity
9. Unclear project definitions and lack of early discussion
10. Lack of early discussion and lack of systematic project management
11. Lack of early discussion and low environmental sensitivity
12. Lack of early discussion and a small project

To cite an example of one finding, where there is an environmentally sensitive context, a clear project definition is responsible for a positive outcome, while the absence of a clear project definition leads to a negative output (Befani and Sager 2010: 275). Should policy makers base their policy decisions on a result such as this?

In fact, policy makers might want to be cautious about reading too much into this result, as the finding is based on only two cases. Moreover, a number of other paths reported in this study are based on only a single case. Though one of QCA's goals is certainly to take each case seriously in its own terms, results based on only one or two cases too often inadequately reflect underlying causal patterns and routinely are not robust to sensitivity tests.

Moreover, the dichotomization necessary to perform csQCA forfeits potentially relevant variations in the concepts of interest. For example, the dependent variable in this analysis takes on a zero if the impact assessment has some implementation deficits, such as missed deadlines or failure to follow certain procedures. However, the dependent variable also takes a value of zero if the impact assessment displayed "complete non-compliance" (Befani and Sager 2010: 274), which is left undefined but clearly meant to convey a case of extremely poor implementation.

The problem with this dichotomy is that the six deterministic paths to an outcome value of zero do not distinguish, for example, between complete non-compliance and merely one missed deadline. Further, the tenth path in the list above yields poor implementation when there is a lack of early discussion and a lack of systematic project management. How should an agency avoid this outcome? One solution may be to add systematic project management, but this is likely to impose a significant cost. If it is unclear whether this cost will result in avoiding a *single missed deadline* or in *complete non-compliance*, the agency will likely want to reevaluate the implied deterministic relationship to see if the relationship disappears when considering *only* cases of complete non-compliance. These dichotomies are ineffective for making useful policy recommendations.

Multi-value QCA is intended to overcome some of the limitations of dichotomies in csQCA. Balthasar (2006) employs mvQCA to answer a crucial question for evaluation studies: Under what circumstances are evaluations of organizations actually used by the agency being assessed? Focusing on ten cases, the analysis includes four explanatory conditions: (i) the overall focus of the evaluation (organizational process versus overall organizational goals),[8] (ii) whether evaluations are routine in each context, (iii) potential usefulness of the evaluation to the agency under review,[9] and (iv) institutional distance between the agency and the evaluating organization. While the outcome and three of the four conditions remain dichotomous, the author allows three discrete values for condition (i), the overall focus: a value of zero indicates purely process oriented evaluations, a value of one indicates purely goal-oriented evaluations, and a value of two indicates a combination of process- and goal-oriented evaluations.[10] Balthasar (2006: 364–365) finds that seven different combinations of conditions explain institutional evaluation use.

[8] Balthasar (2006: 362) employs the commonly used terms formative and summative to refer to evaluations that focus on process and goals, respectively.

[9] Usefulness is defined by Balthasar (2006: 362) as the ability of the findings to be implemented by the agency.

[10] These values are nominal as there is no natural ordering to the scale.

Agencies that have been evaluated make use of the resulting reports if they are:

1. Routine, potentially useful, performed by institutionally distant organizations, and process-focused
2. Routine, potentially useful, performed by institutionally distant organizations, and goal-focused
3. Routine, not potentially useful, performed by institutionally close organizations, and process-focused
4. Not routine, potentially useful, and either both process- and goal-focused, or only goal focused not exclusively process-focused

Agencies *do not* make use of the resulting reports if they are

5. Not potentially useful, performed by institutionally distant organizations, and both process- and goal-oriented
6. Routine, performed by institutionally distant organizations, both process- and goal-oriented
7. Potentially useful, performed by institutionally close organizations, and goal-oriented

Just as in the Befani and Sager (2010) article, the number of cases per path—one or two in each of the seven paths—is worrisome to a policy maker. It is highly likely that some of these results are due to idiosyncrasies that are not replicable or valid in drawing policy lessons. Additionally, in substantive terms, is it plausible that adding a process-oriented portion to routine goal-oriented evaluations will guarantee that an agency with close institutional distance from the evaluator will not use the evaluations? This is precisely what path six suggests. These problems indicate that, though the mvQCA framework allows for a more natural categorization of the goal condition, it does not rescue the analysis from the limitations that QCA imposes.

Might fuzzy-set QCA, which allows for even finer gradations of conditions and outcomes than mvQCA, be useful for policy analysis? Lee (2013) employs this algorithm to compare employment policy in 18 OECD countries, particularly focusing on South Korea and Japan. She explores what combination of policies cause a high rate of non-standard—temporary or otherwise unreliable—employment. Because workers employed in these settings are economically vulnerable and often without the social welfare protection enjoyed by their standardly employed peers, it is important to understand which labor policies encourage employers to rely on non-standard employment.

Lee's analysis considers four policy variables that may influence this type of employment: (i) minimum wage, (ii) unemployment benefits, (iii) employment protection for temporary workers, and (iv) employment protection for permanent workers. In contrast to the dichotomous and multi-valued versions of QCA discussed above, the values range from zero to one for any given condition, with the values of one representing full membership, zero representing full non-membership, and intermediate values representing varying degrees of partial membership. For example, membership in condition (iv), strong employment protection for permanent workers, will be near zero for countries that have very weak protection and near one for countries that have very strong protection.[11] The fsQCA algorithm identifies two causal pathways.

A nation will experience high non-standard employment if it has:

1. Low statutory minimum wage and strong protections for permanent workers
2. Low statutory minimum wage and weak protections for temporary workers

Two of the cases, South Korea and Japan, are examined in greater detail. In South Korea, a low minimum wage in combination with strong protection of permanent workers is sufficient for high non-standard employment; in Japan, a low minimum wage in combination with weak protection of temporary workers is sufficient for high non-standard employment.

Just as in the crisp-set and multi-valued cases, the fuzzy-set scaling system eliminates the units of measurement that are meaningful to policy makers. In order to scale variables, an analyst must first transform raw variables into fuzzy-set membership scores, but this process is often opaque and ill-defined. For example, the proportion of the South Korean temporary workforce is approximately 30 percent. Lee considers South Korea to have nearly full membership in the condition of high temporary employment, giving South Korea a fuzzy-set score of 0.95 for this condition. Japan's temporary workforce is also around 30 percent and considered to have full membership in the condition of high temporary employment, but Lee chooses to give Japan a score of only 0.58 for this condition. This large difference in fuzzy-set scores between South Korea and Japan is perplexing and the author fails to provide an explanation for why the scores are so drastically different.

Yet another step in QCA also contributes to depriving policy makers of meaningful measures. After scaling variables and establishing membership scores for different logical combinations of conditions,[12] a researcher designates a sufficiency threshold and the fsQCA algorithm calculates consistency scores for the combinations of conditions.[13] The analysis thus reverts back to a dichotomous treatment, thereby losing the improvement vis-a-vis csQCA and mvQCA that is provided by the fuzzy set measurement of gradations.

To understand the implications of this loss of information, imagine two possible versions of a Congressional Budget Office report on the impact of a change in minimum wage. In fact, a recent report argued that raising the minimum hourly wage to $10.10 "would reduce total employment by 500,000, or .3 percent….The increased earnings for low wage workers resulting from the higher minimum wage would total $31 billion" (Congressional Budget Office 2014: 1–2). By contrast, a corre-

[11] A full explication of the fuzzy-set scoring and analysis procedure can be found in Schneider and Wagemann (2012).

[12] The lowest score that a given case displays for any of the conditions included in the combination is its membership score for the combination. For instance, if Korea has individual membership scores of 0.8, 0.7, and 0.35 for non-standard employment, welfare benefits, and temporary employment protection, then the membership score for the combination of those conditions is 0.35.

[13] The consistency score measures the strength of sufficiency of each combination of conditions for the outcome.

sponding, hypothetical report based on fsQCA might read: "Raising the minimum wage in countries with strong protection for permanent employees would be sufficient to cause full membership in high unemployment and high low wage income." Such conclusions are vague and, more importantly for policy makers, they lack meaningful units of measurement. These problems are compounded by the fact that the author devotes little space to examining the treatment assignment mechanism—and, without justification of this mechanism, it is unclear if the assignment of minimum wages and employment protections occurs with any approximation of "as-if" random assignment.

By contrast, the canonical minimum wage study in the United States—a study based on observational data—provides far more detail on the assignment mechanism, does not obscure the raw data with fuzzy-set membership scores, and includes simulation checks on the modeling assumptions (Card and Krueger 2000). Notwithstanding the caution of these authors, the as-if random assignment assumption in that paper has been criticized as being implausible (Dunning 2012: 250–251). However, Lee's QCA analysis does not include any defense whatsoever of the assumptions required for a causal interpretation of the already precarious multiple interaction terms derived from the scoring and minimization algorithms. Contrary to suggestions that fsQCA produces results that are especially relevant to policy analysts, such efforts yield little value to the policy research community.

Pennings (2005) likewise applies fuzzy-set QCA to investigate the causes of welfare state reforms in 21 countries. Starting with eight variables from the OECD's Social Expenditures Database, Pennings constructs fuzzy-set membership scores for one of the outcomes of interest, social welfare spending:

> The Z-scores of the expenditures in the first eight SOCX-categories are calculated per category for each single year and multiplied with the share of spending as a percentage of GDP in each category in that year. After this the fuzzy-set scores are calculated for every year and subsequently divided into three periods of five years: 1980–1985, 1986–1991, 1992–1998. (Pennings 2005: 322)

The explanatory conditions are scaled in a similar manner in order to get fuzzy-set membership scores for (i) degree of corporatism, (ii) left-party governance, (iii) economic openness, and (iv) elderly population. The fsQCA algorithm is applied and the results suggest that a high degree of social expenditure will result from the following cluster of conditions.

For all three periods (1980–1985, 1986–1991, 1992–1998), high social expenditure results from:

> 1. A high degree of openness and a high degree of left-party governance
> 2. A high degree of openness and a high degree of elderly population

For 1980-1985, high social expenditure results from:

> 3. A low degree of left-party governance and a high degree of corporatism

For 1986-1991, high social expenditure results from:

> 4. A high degree of openness and a low degree of corporatism

For 1992-1998, high social expenditure results from:

> 5. A low degree of left-party governance and a high degree of elderly population

According to these results, high social expenditures will result with near certainty if a country has an open economy and either left-party governance or an elderly population. However, absence of left-party governance is also sufficient for high social expenditures if there is a high degree of corporatism (only in the early 1980s) or an elderly population (only in the 1990s). The exact form of social expenditures cannot be recovered from this analysis, because the original variables are transformed. Pennings argues that the fuzzy-set scoring has the advantage of measuring gradations, but this feature brings a loss of interpretability. Moreover, the fsQCA algorithm ultimately dichotomizes findings, thereby losing the key advantage vis-à-vis the crisp-set and multi-valued alternatives.

Each of the QCA studies identified thus far conducts analysis on a small number of cases. Given the challenges of causal inference with a small N, might QCA offer lessons to policy makers if conducted on a larger N? Warren, Wistow, and Bambra (2013) use csQCA to study 90 individuals who are unemployed due to ill health. The authors focus on the impact of a welfare intervention designed to improve health outcomes and consider five explanatory conditions: (i) age, (ii) sex, (iii) type of ill health,[14] (iv) skill level, and (v) frequency of social interactions with neighbors.

In a study like this, QCA might leverage the large N to distinguish between real patterns in the cases analyzed and patterns that result from measurement error or from possible idiosyncrasies in the data. Instead, the study focuses on a surprisingly large number of complex interactions that, it is argued, explain improved health. With five explanatory conditions, there are 32 (2^5) potential causal pathways. This study concludes that 30 of these are in fact pathways to the outcome, meaning that csQCA identifies nearly every possible interaction of conditions as a causal combination.

This large number of causal pathways is hard for a policy maker to interpret. To understand why this is the case, consider these two sufficiency results: (1) improved health is a result of being a younger man of high skill who is not likely to talk to his neighbors, and (2) improved health is a result of being an older man of low skill who is not likely to talk to his neighbors. What is the appropriate policy response? What is the mechanism through which neighbor avoidance is a catalyst to good health for younger (but not older) high-skilled men and older (but not younger) low skilled men? With so many causal pathways and no clear mechanism, policy makers cannot use the results of this method for policy prescription.

With standard tools of policy analysis, larger N will increase the precision of results and allow for more confident policy implications. As this example suggests, an increased N

[14] The study distinguished between mental ill health and musculoskeletal problems.

may not have the same advantage in QCA. The algorithm and deterministic framework combine to produce questionable results with little policy relevance.

To summarize these QCA studies: A series of questions have been posed about their value for public policy analysis, and more broadly about their contributions to basic empirical research. The answers have been exceedingly disappointing. These articles do *not* yield insights of interest or relevance to policy researchers; and the norms and practices of QCA illustrated here also appear highly questionable from the standpoint of wider norms about research methods.

Broader Concerns about QCA

These examples point to wider issues regarding basic methodological recommendations and practices of QCA.

Net Effects. What does this comparison between conventional and QCA studies tell us about the criticism of the "net-effects" framework that is a central and valuable feature of conventional policy research? Ragin (2008) criticizes standard, quantitative methods of social science as adhering to "net-effects thinking," which he describes in a representative section of *Redesigning Social Inquiry: Fuzzy Sets and Beyond*:

> In what has become "normal" social science, researchers view their primary task as one of assessing the relative importance of causal variables drawn from competing theories…. The key analytic task is typically viewed as one of assessing the relative importance of the relevant variables. If the variables associated with a particular theory prove to be the best predictors of the outcome (i.e., the best "explainers" of its variation), then this theory wins the contest. (Ragin 2008: 177)

This description, as evidenced by the exemplary studies in the first section, is not reflective of either the goals or the rigorous standards for causal inference in good evaluation research. Relative explanatory power is indeed one of the pieces of information yielded by multivariate regression (Angrist and Pischke 2009: 34–35; Greene 2012: 28–30; Wooldridge 2010: 15–25), but it is rarely the focus of rigorous policy analysis. For example, Angrist et al. (2012) do not focus on the power of charter schools to predict student test scores vis-à-vis the explanatory power of demographic and economic variables. Rather, they focus on estimating the impact of charter schools in a transparent and simple manner by finding plausibly random variation in the assignment of charter school status.

Focus Is Not on Comparing Causal Influence of Several Variables. More broadly, research on public policy generally evaluates the impact of at most one or two policies. The key analytic task is not assessing the relative strength of a host of variables, but rather estimating the impact of each relevant policy variable (again, usually one or two). In this sense, the characterization in the quotation above from Ragin (2008) does not correspond to standard practices. For example, in five of the six quantitative articles discussed above, the primary focus is on a single variable. In the sixth article, Mauldon et al.'s (2000) study of high school graduation for teenage mothers, the focus is on two subcomponents of one policy and their interaction. Though it is a useful benchmark, this article does not focus on whether a demographic variable such as family background is a better predictor of high school graduation than participation in the Cal Learn program. Rather, the authors, funders of the program, and policy community at large need to know how participation in the two sub-components of Cal Learn impacts the target group.

Context and Causal Heterogeneity. Ragin (2008) argues that quantitative research methods ignore context and heterogeneity. He states:

> Consider also the fact that social policy is fundamentally concerned with social intervention. While it might be good to know that education, in general, decreases the odds of poverty (i.e., it has a significant, negative net effect on poverty), from a policy perspective it is far more useful to know under what conditions education has a decisive impact, shielding an otherwise vulnerable subpopulation from poverty. (Ragin 2008: 181–182)

Ragin is correct that it is important to know whether certain sub-groups in the target population respond to the treatment more than others, but he overlooks the fact that standard policy research routinely searches for these heterogeneous treatment effects. As Ladd and Lauen (2010), Sen (2012), and Reardon et al. (2012) demonstrate, conventional methods are able to identify differential effects by describing the treatment assignment mechanism and *without* discarding information on effects through measurement coding strategies, or because the policies are neither necessary nor sufficient for an outcome.

Certain methods are even more flexible. For instance, if policy variables are binary, researchers have a host of non-parametric estimation methods that recover the average treatment effect with very few of the assumptions required by the ordinary least-squares estimator (Imbens 2004). Some of these techniques allow researchers to go beyond average effects. For example, kernel density estimators can be used to analyze the effect of a policy on the distribution of an outcome, while quantile regression can be used to analyze impacts at specific points in a distribution (Bitler, Gelbach, and Hoynes 2006). What this corpus of techniques shares is the ability to estimate the precise effect of policy, whether net or distributional, either for the full N or for subgroups. These techniques do not discard information on effects merely because the policies are neither necessary nor sufficient for an outcome; nor do they require the transformation of variables into fuzzy set membership scores.

Case-Oriented versus Variable-Oriented. The case-oriented versus variable-oriented framework is likewise not helpful for thinking about policy effects. Consider the frequently repeated QCA thesis, both in the general arguments and in the discussion of net effects, that (1) conventional quantitative research is "variable oriented"; by contrast, (2) QCA is "case-oriented"—i.e., focused on "kinds of cases," on "cases as configurations." This distinction is evoked in depicting the contrast between the analysis of net-effects in quantitative research, as opposed to causal configurations in QCA.

However, both of these characterizations are inadequate.

Table 4: Overview of QCA Studies

Study	Substantive Focus	Type of QCA	Number of Explanatory Conditions	Number of Paths	Number of Cases	Average Cases per Path	Analysis of Mechanisms	Plausibility of Causal Inference
Balthasar 2006	Evaluation Use	mvQCA	4	7	10	1.4	Absent	Weak
Befani and Sager 2010	Environmental Impact Assessments	csQCA	6	12	15	1.3	Absent	Weak
Lee 2013	Employment Policy	fsQCA	4	10	18/144	1.8/14.4[a]	Absent	Weak
Pennings 2005	Welfare Expenditures	fsQCA	4	5	21	4.2	Absent	Weak
Warren, Wistow, and Bambra 2013	Health Policy	csQCA	5	30	90	3	Absent	Weak

[a] In this panel design, 18 countries are analyzed over 8 years, yielding 144 country years.

(1) With regard to variable-oriented: The causal conditions analyzed in QCA are variables—by any conventional meaning of that term. Variables that have been rescaled into dichotomous, multichotomous, or "fuzzy" forms are still variables, regardless of the reference to them as causal conditions. (2) With regard to case knowledge—taking for example the field of education policy as discussed above—it is standard in this field for quantitative researchers to have extremely detailed knowledge of specific schools and districts. Such knowledge has been used, for example, to debunk sloppy empirical conclusions regarding the Heritage Foundation's "No Excuses" schools that have high performing, high poverty students. Rather than attributing these schools' success to frequency of testing, ease of firing teachers, and resistance to bureaucracy, contextual knowledge allows Rothstein (2004) to identify confounding variables that explain away the Heritage Foundation's thesis.

> This kind of analysis yields some ludicrous results. One Heritage no excuses school, with high poverty and high scores, enrolled children of Harvard and M.I.T graduate students. Graduate stipends may be low enough for subsidized lunches, but these children are not those whose scores are cause for national concern, nor is their performance a model for truly disadvantaged children. (Rothstein 2004: 73)

A recent book on conducting social experiments emphasizes context heterogeneity in randomized control trials and devotes a chapter to methods that estimate such effects (Bloom 2006: 37–70). These methods are standard practice for rigorous policy research.

Norms for Causal Inference. Another issue concerns current standards for causal inference. In the QCA examples considered here, the authors are completely inattentive to the rising concern about challenges of causal assessment with observational data. Technical specification issues aside, searching for the variable with the greatest explanatory power in observational data would not provide compelling evidence of a causal effect. Observational data are plagued by the problem of endogenous explanatory variables, as has been recognized for decades (Heckman, Ichimura, and Todd 1997; Lalonde 1986). The primary focus of top tier policy research is the identification of exogenously determined variation in one or two policy variables and its consequent effect on outcomes. Entire sections of articles are devoted exclusively to this question, and properly so. Without a persuasive account of why a variable is distributed as if by random assignment, the causal results returned by any algorithm, including both QCA and regression, are not compelling (Rubin 2005). QCA scholars do not use this framework and describe causal results from observational data without any discussion of the treatment assignment mechanism. None of the five QCA policy evaluations discuss treatment assignment.

Uncertainty and Random Variability. Policy research should be centrally concerned with uncertainty and random variability. For more than a decade, scholars have been urging the policy research community, including non-academic institutions like the Congressional Budget Office, to incorporate uncertainty into policy analysis (Manski 1995). Set theoretic frameworks, although they note error and uncertainty, have not embraced this emerging perspective and instead basically view the world as deterministic. As the above examples of conventional policy research show, the average impact of an explanatory variable is typically small. As a proportion of the

full range in possible outcomes, the explanatory variables routinely have at most a modest impact. Yet as was shown, even this modest impact can have important consequences for other outcomes. If scholars are to successfully detect these small effects, it is mandatory to parse out the effects themselves, as opposed to error and uncertainty. QCA's Boolean framework is not designed to distinguish between large and small effects, nor to parse out error and uncertainty versus the effects themselves.

The method misses precisely the kind of finding that interests policy researchers. By contrast, standard tools of causal inference can find effects of any size, given a large enough N.

Conclusion: An Unsuitable Method

This discussion has focused on the field of policy evaluation—a crucial case, as it was framed in the introduction, for evaluating the relevance of Qualitative Comparative Analysis to policy research. Public policy analysts seek insights into the real-world impact of policies, which are often marginal changes in human behavior and well-being. Such insights are yielded by well-established methods of policy evaluation.

By contrast, conceptualizing policy outcomes in terms of bounded sets and scoring cases according to set membership forces causal inference into a framework ill equipped to uncover meaningful variation in outcomes. Policy research should be able to reveal modest effects at the margin, which is precisely the focus of established research methods.

More broadly, this analysis has raised serious concerns about QCA's wider contribution to good causal inference. The method's major shortcomings merit close, ongoing scholarly attention.

References

Allison, Graham. 2006. "Emergence of Schools of Public Policy: Reflections by a Founding Dean." In *The Oxford Handbook of Public Policy*, eds. Michael Moran, Martin Rein, and Robert E. Goodin. New York: Oxford University Press.

Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters. 2012. "Who Benefits from KIPP?" *Journal of Policy Analysis and Management* 31 (4): 837–860.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Balthasar, Andreas. 2006. "The Effects of Institutional Design on the Utilization of Evaluation: Evidenced Using Qualitative Comparative Analysis (QCA)." *Evaluation* 12 (3): 353–371.

Befani, Barbara and Fritz Sager. 2010. "QCA as a Tool for Realistic Evaluation: The Case of the Swiss Environmental Impact Assessment." In *Innovative Comparative Methods for Policy Analysis: Beyond the Quantitative-Qualitative Divide*, eds. Benoît Rihoux and Heike Grimm. New York: Springer.

Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes. 2006. "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments." *American Economic Review* 96 (4): 988–1012.

Bloom, Howard, ed. 2006. *Learning More from Social Experiments: Evolving Analytic Approaches*. Thousand Oaks, CA: Russell Sage Foundation Publications.

Card, David and Alan Krueger. 2000. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply." *American Economic Review* 90 (5): 1397–1420.

Chetty, Raj, John Friedman, and Jonah Rockoff. 2013. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." (No. 19424). Retrieved from http://www.nber.org/papers/w19424.

Congressional Budget Office. 2014. "The Effects of a Minimum-Wage Increase on Employment and Family Income." Retrieved from http://www.cbo.gov/publication/44995.

Datar, Ashlesha and Nancy Nicosia. 2012. "Junk Food in Schools and Childhood Obesity." *Journal of Policy Analysis and Management* 31 (2): 312–337.

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-based Approach*. Cambridge: Cambridge University Press.

Greene, William H. 2012. *Econometric Analysis*, 7th ed. Saddle River, NJ: Prentice Hall.

Hanushek, Eric A. 2003. "The Failure of Input-based Schooling Policies." *The Economic Journal* 113 (485): F64–F98.

Hanushek, Eric A. 2011. "The Economic Value of Higher Teacher Quality." *Economics of Education Review* 30 (3): 466–479.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1997. "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *The Review of Economic Studies* 64 (4): 605–654.

Hudson, John and Stefan Kühner. 2013. "Qualitative Comparative Analysis and Applied Public Policy Analysis: New Applications of Innovative Methods." *Policy and Society* 32 (4): 279–287.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *The Review of Economics and Statistics* 86 (February): 4–29.

Johnson, Rucker C. 2011. "Long-run Impacts of School Desegregation and School Quality on Adult Attainments" (No. 16664). Retrieved from http://www.nber.org/papers/w16664.

Ladd, Helen and Douglas Lauen. 2010. "Status versus Growth: The Distributional Effects of School Accountability Policies." *Journal of Policy Analysis and Management* 29 (3): 426–450.

Lalonde, Robert J. 1986. "Evaluating Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76 (4): 604–620.

Lee, Sophia Seung-yoon. 2013. "High Non-standard Employment Rates in the Republic of Korea and Japan: Analyzing Policy Configurations with Fuzzy-Set/QCA." *Policy and Society* 32 (4): 333–344.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Mauldon, Jane, Janet Malvin, John Stiles, Nancy Nicosia, and Eva Seto. 2000. "The Impact of California's Cal-Learn Demonstration Project, Final Report." Retrieved from http://escholarship.org/uc/item/2np332fc.

Pennings, Paul. 2005. "The Diversity and Causality of Welfare State Reforms Explored with Fuzzy-Sets." *Quality and Quantity* 39 (3): 317–339.

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

Ragin, Charles C. 2010. "The Limitations of Net-Effects Thinking." In *Innovative Comparative Methods for Policy Analysis: Beyond the Quantitative-Qualitative Divide*, eds/ Benoît Rihoux and Heike Grimm. New York: Springer.

Reardon, Sean F., Elena Grewal, Demetra Kalogrides, and Erica Greenberg. 2012. "Brown Fades: The End of Court-Ordered School Desegregation and the Resegregation of American Public Schools." *Journal of Policy Analysis and Management* 31 (4): 876–904.