Available online at www.sciencedirect.com

ScienceDirect

Policy and Society 33 (2014) 287–298

policy and society

www.elsevier.com/locate/polsoc

QCA is of questionable value for policy research

Sean Tanner

Goldman School of Public Policy, University of California, Berkeley, United States

Abstract

Qualitative Comparative Analysis (QCA) has been championed as a valuable tool for public policy research. Focusing on the field of policy evaluation, this research note assesses QCA by comparing research that uses this method to studies based on standard practices for quantitative policy analysis. While attention is centrally focused on causal inference, questions of measurement are also addressed. The analysis suggests that QCA adds little value to current methods of policy scholarship, and its contribution in fact falls far short, compared with present-day standard practices. For example, a properly defined ''net effects'' framework – which is pointedly rejected by QCA – provides valuable insights regarding the causal effects that are a central concern of policy evaluation. By contrast, as an approach to policy analysis, QCA suffers from severe limitations in both its framework and its findings.

1. QCA and public policy analysis

Qualitative Comparative Analysis (QCA) has been championed as a valuable tool for public policy research (Hudson & Kühner, 2013b; Rihoux & Grimm, 2010; Rihoux, Rezsöhazy, & Bol, 2011). The proponents of this method advance a number of claims about its special relevance for policy studies: QCA focuses on set theoretic relationships, can uncover multiple conjunctural causation, and allows flexible causal modeling (Rihoux et al., 2011, pp. 16–17).[1] A further premise is that it allows analysts to move beyond the constraints of ''net effects thinking'' to consider more complex forms of interactions among explanatory variables (Ragin, 2010, pp. 16–24; Schneider & Wagemann, 2012, pp. 83–89).

The recent symposium in this journal (Hudson & Kühner, 2013a) echoed these arguments and explored QCA's purported advantages for both measurement and causal inference; three articles in the symposium were devoted to measurement, and two focused on causal inference.

This research note explores QCA's contribution to causal inference by examining its performance in the crucial field of policy evaluation – i.e., studies of policy impacts. Substantial attention is also devoted to measurement issues. The focus on evaluation studies is justified for two reasons. First, policy evaluation is the place where ''the rubber hits the road'' in terms of causal inference. Evaluating the causal impact of policies is the field's raison d'être, and any novel method of inference must be able to perform that task at least as well as conventional methods. Second, the evaluation field is singled out for special concern by the symposium's editors, who argue that the dearth of QCA-based

E-mail address: stanner@berkeley.edu.

[1] 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.

http://dx.doi.org/10.1016/j.polsoc.2014.08.003
1449-4035/

Table 1
Overview of standard best-practice studies.

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

policy evaluations is ''unfortunate and limits the potential of QCA as a research tool'' (Hudson & Kühner, 2013b, p. 279). Among the diverse domains of policy analysis, evaluation research is therefore a critical test-case for QCA.

The next section presents six examples of studies that use standard methods of causal inference to deliver credible, useful results. The following section examines the sharply contrasting contributions of five studies that have been offered as exemplars of QCA-based policy analysis. The third section builds on the discussion of these studies to raise broader questions about some of QCA proponents' basic arguments and practices. 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.

A further introductory point must be underscored. This evaluation of QCA is not in any sense offered from the standpoint of ''quantitative methods.'' Quite the contrary; the norms of evidence and inference employed here are also the basis for a major critique of conventional, regression-based quantitative analysis – a critique that has recently led to a fundamental rethinking of methods in social science (Angrist & Pischke, 2010). Further, while ideas about causal inference in experiments and natural experiments are part of this rethinking, the point is definitely not that all researchers should be doing experiments. Rather, these ideas have played a productive role in the wider reconsideration of causal inference (Brady & Collier, 2010).

2. Policy analysis with standard, current practices

In an environment in which the effects of government action are typically small, 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 yield information on these impacts (Allison, 2006, p. 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 of credible studies are listed in Table 1.

To begin with a simple example: Angrist, Dynarski, Kane, Pathak, and Walters (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, & Rockoff, 2013; Hanushek, 2011) finds that differences of this magnitude are associated with substantial increases in lifetime earnings. This average partial (or ''net'') effect of charter schools is 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 increase by only a modest amount overall. Yet strikingly, the effect is largest for the lowest performing students (Ladd & Lauen, 2010). Again, this magnitude of gain is predicted to yield an appreciable increase in lifetime earnings – a matter of enormous policy relevance because of the frequent failure of the US education system to improve student success (Hanushek, 2003).

By contrast, in another domain the more at-risk population is *not* similarly advantaged. Sen (2012) finds that people 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%. However, there is no detectable increase for the lowest socioeconomic group,[2] whereas for the middle income group

[2] The point estimate of a 0.8% increase is not distinguishable from zero.

Table 2
Detailed summary of standard best practices studies.

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
Ladd and Lauen (2010)	Teacher performance standards	Fixed effects regression	Small	Greater gains in the tails of the distribution	Detailed	Moderate
Sen (2012)	Gas prices and exercise	Fixed effects regression	Small	No effect for lower SES group	Detailed	Moderate
Reardon et al. (2012)	School re-segregation	Interrupted time series	Medium	Greater impact in South than North	Detailed	Moderate
Mauldon et al. (2000)	Educational attainment of teen mothers	Randomized control trial	Small	Best results when both policies applied	Detailed	Strong
Datar and Nicosia (2012)	School nutrition	Instrumental variables regression	None	No effect	Detailed	Moderate

the increase is 3.7% (Sen, 2012, p. 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, Grewal, Kalogrides, and Greenberg (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, as desegregated school districts have improved the long-term health and incomes of African-American students (Johnson, 2011).

Though each of the studies focus on one intervention, or "treatment," policy researchers also 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, Malvin, Stiles, Nicosia, and Seto (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 itself has no effect, and the truly significant effect occurs when the two interventions combine. This finding is of great interest to analysts designing 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 also 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 vital.

2.1. Summary of standard best practices

Table 2 summarizes key features of these six studies. All of them seek to meet current, exacting standards for good causal inference – though certainly some are more successful 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 appropriate tools to assess the robustness of findings.

In substantive terms, policy analysts care about average partial effects, and these studies directly tackle that issue. Subgroup differences and interactions routinely exist within this framework, and these examples show that analysts often examine them to great advantage. Whether the focus is on subgroups or entire samples, the policy researcher

Table 3
Overview of QCA studies.

Study	Substantive focus	Type of QCA
Befani and Sager (2010)	Environmental impact assessments	csQCA
Balthasar (2006)	Evaluation use	mvQCA
Lee (2013)	Employment policy	fsQCA
Pennings (2005)	Welfare expenditures	fsQCA
Warren et al. (2013)	Health policy	csQCA

cares most about the net impact of policies. This is the fundamental basis for embracing, modifying, or rejecting policies. The methods that evaluate net effects directly address that high priority.

Finally, these studies generally do well 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, randomness 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 treatment was experienced by one subgroup, but not by the other subgroup, ''as-if'' by random assignment. Through explicit discussion of the treatment assignment mechanism, researchers bolster confidence in their causal inferences.

3. Policy analysis with QCA

QCA scholars who recommend applying their method to policy analysis have offered many illustrations of their approach. The discussion below focuses on five examples (Table 3): the chapter-length study by Befani and Sager (2010) that is included in the book *Innovative Comparative Methods for Policy Analysis* (Rihoux & Grimm, 2010); two examples from the extensive review by Rihoux et al. (2011)[3]; and two examples from the symposium on QCA in the journal *Policy and Society* (Hudson & Kühner, 2013a). The five selected for comparison are specifically policy evaluations and, therefore, provide a reasonable comparison group for the six articles that use standard methodological tools.[4]

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

3.1. *Befani and Sager (2010) on environmental impact assessments*

These authors investigate the circumstances under which Swiss environmental impact assessments are effectively implemented.[5] Impact assessments are an enormously important aspect of environmental policy-making, and improperly implemented assessments undermine a fundamental tool of environmental regulation.

Using crisp-set QCA 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

[3] Of the 144 articles reviewed by Rihoux, Rezsohazy, and Bol, only six were categorized by the authors as ''policy evaluations'' (2011, p. 21). Five of those six articles used the results from a single QCA-based study of Swiss Environmental Impact Assessments, so that only two different topics are analyzed among the six articles. Among the five articles that dealt with the Swiss case, the Befani and Sager (2010) chapter in the book *Innovative Comparative Methods for Policy Analysis* (2010) offers the most fine-grained analysis, and therefore is a plausible choice for exemplifying QCA to best advantage. Balthasar's (2006) analysis of agency reviews and institutional distance is the other policy evaluation from this list. Pennings (2005) was included as well.

[4] As is the case in conventional policy research, many other QCA-based policy articles address measurement issues or the process of policy formation.

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

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) the size of the project.

The authors find that the 15 cases can be completely accounted for by the 12 distinct causal paths listed below.[6] Assessments are well-implemented if there are:

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 output, while the absence of a clear project definition leads to a negative output (Befani & Sager, 2010, p. 275). Should policy makers base their policy decisions on a result such as this?

In fact, policy makers should be wary of reading too much into this result, as the finding is based on only two cases. Moreover, a number of the other paths reported in this study are based only on a single case. Though one of QCA's goals is to offer context-specific results, such results based on only one or two cases are often overly specific and hence not robust to sensitivity tests, such as adding or dropping cases.

Moreover, with regard to measurement, 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. But, the dependent variable also takes a value of zero if the impact assessment displayed "complete non-compliance" (Befani & Sager, 2010, p. 274), which is left undefined but clearly meant to convey a case of extremely poor implementation.

The problem with this dichotomization is that the six deterministic paths to an outcome value of zero do not distinguish between complete non-compliance and merely one missed deadline. For example, 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? The logical solution may be to consider adding systematic project management, but such a solution 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.

3.2. *Balthasar (2006) on effective use of evaluation studies*

Multi-value QCA is intended to overcome the limitations of the approach to measurement in the crisp set version, specifically by moving beyond the dichotomies of 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),[7] (ii) whether evaluations are routine in

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

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

each context, (iii) the potential usefulness of the evaluation to the agency under review,[8] and (iv) the 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.[9]

Balthasar finds that seven different combinations of conditions explain institutional evaluation use (2006, pp. 364–365).

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-oriented (1 case).
2. Routine, potentially useful, performed by institutionally distant organizations, and goal-oriented (1 case).
3. Routine, not potentially useful, performed by institutionally close organizations, and process-oriented (1 case).
4. Not routine, potentially useful, performed by institutionally close organizations, and either both process- and goal-oriented, or only process-oriented (2 cases).

Conversely, 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 (2 cases).
6. Routine, performed by institutionally distant organizations, both process- and goal-oriented (1 case).
7. Potentially useful, performed by institutionally close organizations, and goal-oriented (2 cases).

Just as in the Befani and Sager (2010) piece, the number of cases per path – one or two in each of the seven paths – is alarming. 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 component to the more standard goal-oriented component 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 more plausible measurement of the purpose of evaluations, it does not rescue the analysis from the limitations that QCA imposes.

3.3. *Lee (2013) on employment policy*

Might fuzzy-set QCA, which allows for even finer gradations than mvQCA in measuring conditions and outcomes, 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.[10] The fsQCA algorithm ''identifies'' two causal pathways.

[8] Usefulness is defined by Balthasar as the ability of the findings to be implemented by the agency (2006, p. 362).
[9] These values are nominal as there is no natural ordering to the scale.
[10] A full explication of the fuzzy-set scoring and analysis procedure can be found in Schneider & Wagemann (2012).

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.

Unfortunately, analysts learn few policy lessons from this fsQCA analysis. Just as in the crisp-set and multi-valued versions, fuzzy-set scaling 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 ''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 very 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.

As another step in the process that leaves policy makers without meaningful measures, after scaling variables and establishing membership scores for logical combinations of conditions,[11] a researcher designates a ''sufficiency threshold,'' and the fsQCA algorithm calculates ''consistency scores'' for the combinations of conditions.[12] During this process, the analysis reverts back to a dichotomous treatment, meaning that it loses the ''improvement'' over csQCA and mvQCA provided by the fuzzy set measurement gradations.

To understand the implications of this loss of information in the process of measurement, 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, pp. 1–2). By contrast, a corresponding, 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 exceptionally vague and, more importantly for policy makers, devoid of 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 impossible to believe that assignment of minimum wages and employment protections occurs with any approximation at all of ''as-if'' random assignment.

By contrast, the canonical minimum wage study in the United States provides far more detail on the assignment mechanism, does not obscure the raw data by basing measurement on fuzzy-set membership scores, and includes robustness checks on the modeling assumptions (Card & 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, pp. 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. In stark contrast to suggestions that fsQCA produces results that are especially relevant to policy analysts, such efforts yield little of value to the policy research community.

3.4. *Pennings (2005) on welfare expenditures*

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

[11] 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.

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

> 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. (322)

The explanatory conditions are measured in a similar manner in order to get fuzzy-set membership scores in (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 this analysis, 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 claims 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.

3.5. *Warren et al. (2013) on Health Policy*

Each of the QCA studies identified thus far conducts analysis on a small number of cases. Because causal inference is always difficult with small sample sizes, though, might QCA offer lessons to policy makers *if* conducted on a larger sample? Warren, Wistow, and Bambra (2013) employ the measurement framework of 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,[13] (iv) skill level, and (v) frequency of social interactions with neighbors.

Rather than leveraging the large sample to distinguish between real patterns and idiosyncrasies of the sample, QCA merely increases the number of complex interactions that ''explain'' improved health. With five conditions and an approach to measurement once again restricted to dichotomous variables, there are 32 (2^5) potential causal pathways. Remarkably, 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 plausible causal combination.

This finding of a large number of causal pathways is not useful to a policy maker. 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

[13] Health conditions that involve the musculoskeletal system or not.

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 et al. (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.

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 identified causal pathways and no clear mechanism, policy makers cannot use the results of this method for policy prescription.

With standard tools of policy analysis, increased sample sizes will, ceteris paribus, increase the precision of results in random samples and allow for more confident policy implications. As this example demonstrates, increased sample sizes do not have the same advantage in QCA. The algorithm and deterministic framework combine to produce untenable results with little policy relevance, even in large samples, as is evident in Table 4.

4. Questionable QCA arguments and practices

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

4.1. 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 criticizes standard, quantitative methods of social science as adhering to ''net-effects thinking,'' which he describes in a representative section of *Redesigning Social Inquiry: Fuzz Sets and Beyond* (2008):

> 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. (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 policy research. Relative explanatory power is indeed one of the pieces of information yielded by multivariate regression (Angrist & Pischke, 2009, pp. 34–35; Greene, 2012, pp. 28–30; Wooldridge, 2010, pp. 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 simply and transparently by finding plausibly random variation in the assignment of charter school status.

4.2. 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 Ragin's quotation above 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 by 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.

4.3. Context and causal heterogeneity

Ragin 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. (181–182)

Ragin is correct that it is important to know whether certain sub-groups within the target population respond to the treatment more than others, but he ignores 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 clearly describing the treatment assignment mechanism, and *without* discarding information on effects due to the approach to measurement, 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 nonparametric 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, & Hoynes, 2006). What this corpus of techniques shares is the ability to carefully estimate the precise effect, whether net or distributional, of a policy, either on the sample as a whole or various subcomponents therein. 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 wholesale transformation of variables into fuzzy set membership scores.

4.4. 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 their general arguments and in the discussion of net effects, that (1) conventional quantitative research is seen as ''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. (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 education policy as discussed above – it is standard for quantitative researchers to have extremely detailed knowledge of schools and districts. Such knowledge has been used 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. (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, pp. 37–70). These methods are standard practice for rigorous policy research.

4.5. Norms for causal inference

Another issue concerns current standards for causal inference. QCA scholars appear to be unaware of the rigorous new skepticism about causal assessment with observational data. Technical specification issues aside, searching for the variable with the greatest explanatory power in observational data does 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, & 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 ignore 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.

4.6. 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, ignore this emerging perspective and instead unproductively view the world as deterministic. As the six examples of conventional policy research reveal, the average impact of an explanatory variable is typically small. As a proportion of the full variability in outcomes, the explanatory variables routinely change the outcomes by less than one tenth of their full ranges. 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 large enough samples.

5. Conclusion: an unsuitable method

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 research methods. Conceptualizing policy outcomes in terms of bounded sets and then basing measurement on membership in those sets forces causal inference into a set-theoretic 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.

References

Allison, G. (2006). Emergence of schools of public policy: Reflections by a founding dean. In M. Moran, M. Rein, & R. E. Goodin (Eds.), *The Oxford handbook of public policy* (pp. 58–79). New York, USA: Oxford University Press.

Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A., & Walters, C. R. (2012). Who benefits from KIPP? *Journal of Policy Analysis and Management, 31*(4), 837–860.

Angrist, J. D., & Pischke, J. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives, 24*(2), 3–30.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press (p. 373).

Balthasar. (2006). The effects of institutional design on the utilization of evaluation: Evidenced using qualitative comparative analysis (QCA). *Evaluation, 12*(3), 353–371.

Befani, B., & Sager, F. (2010). QCA as a tool for realistic evaluation: The case of the Swiss environmental impact assessment. In B. Rihoux & H. Grimm (Eds.), *Innovative comparative methods for policy analysis: Beyond the quantitative–qualitative divide* (pp. 263–284). New York: Springer.

Bitler, M. P., Gelbach, J. B., & Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review, 96*(4), 988–1012.

Bloom, H. (Ed.). (2006). *Learning more from social experiments: Evolving analytic approaches*. New York: Russell Sage Foundation Publications.

Brady, H. E., & Collier, D. (Eds.). (2010). *Rethinking social inquiry: Diverse tools, shared standards* (2nd ed.). Lanham, MD: Rowman & Littlefield Publishers.

Card, D., & Krueger, A. B. (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, R., Friedman, J. N., & Rockoff, J. E. (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. (Online). February, 2014. Available: http://www.cbo.gov/publication/44995. Referenced: April 15, 2014.

Datar, A., & Nicosia, N. (2012). Junk food in schools and childhood obesity. *Journal of Policy Analysis and Management, 31*(2), 312–337.

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

Greene, W. H. (2012). *Econometric analysis* (7th ed.). Saddle River, NJ: Prentice Hall.

Hanushek, E. A. (2003). The failure of input-based schooling policies. *Economic Journal, 113*(485), F64–F98.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30*(3), 466–479.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies, 64*(4), 605–654.

Hudson, J., & Kühner, S. (2013a). Innovative methods for policy analysis: QCA and fuzzy sets. *Policy and Society, 32*(4), 279–356.

Hudson, J., & Kühner, S. (2013b). Qualitative comparative analysis and applied public policy analysis: New applications of innovative methods. *Policy and Society, 32*(4), 279–287.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics, 86*(February), 4–29.

Johnson, R. 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, H. F., & Lauen, D. L. (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management, 29*(3), 426–450.

Lalonde, R. J. (1986). Evaluating econometric evaluations of training programs with experimental data. *American Economic Review, 76*(4), 604–620.

Lee, S. S. (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, C. F. (1995). *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press (p. 172).

Mauldon, J., Malvin, J., Stiles, J., Nicosia, N., & Seto, E. Y. (2000). *The impact of California's cal-learn demonstration project final report*. Berkeley: UC DATA, University of California. Retrieved from: http://escholarship.org/uc/item/2np332fc.

Pennings, P. (2005). The diversity and causality of welfare state reforms explored with fuzzy-sets. *Quality and Quantity, 39*(3), 317–339.

Ragin, C. C. (2008). *Redesigning social inquiry: Fuzzy sets and beyond*. Chicago: University of Chicago Press.

Ragin, C. C. (2010). The limitations of net-effects thinking. In B. Rihoux & H. Grimme (Eds.), *Innovative comparative methods for policy analysis: Beyond the quantitative–qualitative divide* (pp. 13–41). New York: Springer. Paperback.

Reardon, S. F., Grewal, E. T., Kalogrides, D., & Greenberg, E. (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.

Rihoux, B., & Grimm, H. (2010). *Innovative comparative methods for policy analysis: Beyond the quantitative–qualitative divide*. New York: Springer (p. 344).

Rihoux, B., Rezsöhazy, I., & Bol, D. (2011). Qualitative comparative analysis (QCA) in public policy analysis: An Extensive Review. *German Policy Studies, 7*(3), 9–82.

Rothstein, R. (2004). *Class and schools: Using social, economic, and educational reform to close the black–white achievement gap*. Washington, DC: Economic Policy Institute (p. 203).

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association, 100*(469), 322–331.

Schneider, C. Q., & Wagemann, C. (2012). *Set-theoretic methods for the social sciences: A guide to qualitative comparative analysis*. Cambridge: Cambridge University Press (p. 392).

Sen, B. (2012). Is there an association between gasoline prices and physical activity? Evidence from American time use data. *Journal of Policy Analysis and Management, 31*(2), 338–366.

Warren, J., Wistow, J., & Bambra, C. (2013). Applying qualitative comparative analysis (QCA) to evaluate a public health policy initiative in the north east of England. *Policy and Society, 32*(4), 289–301.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge: MIT Press.