Spring 2014, Vol. 12, No. 1

Qualitative & Multi-Method Research

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

Contents

Letter from the Section President

Lisa Wedeen
University of Chicago
lwedeen@uchicago.edu

As the newly elected president of the Qualitative and Multi-Methods section of the American Political Science Association, it is my pleasure to introduce an especially charged issue of the newsletter. I write "pleasure" because although I had nothing to do with the theme or articles selected, I am glad to endorse healthy contention. An idea, like political life, often gains vitality through agonistic debate—through the creative frictions produced when staking out positions or defending commitments in public. It is my hope that subsequent issues will also produce imaginative openings for new kinds of discussion. To welcome ideas that shift the grounds on which our arguments previously found traction—this is our obligation as intellectuals. We are lucky to have a vocation enabling us to do what we love. Whether by generating an elegant game theoretic model, puzzling over a passage of philosophical import, doing fieldwork, mining the archives, solving a math problem, interpreting a film, conducting an experiment, writing questions for a survey, or devising new theories of political change and retrenchment, we have the good fortune of participating in worlds that are sustaining and affirming. Despite our tendencies toward justification, we would do well to acknowledge that our methodological choices are often based on what makes us happy and at ease in our environments. For some, joy comes from destabilizing conventional ways of thinking. For others, it is the activity of establishing new conventions or enriching old ones that invigorates.

We are a large section, encompassing a wide array of viewpoints and intellectual traditions. It is to be welcomed when we have serious scholarly disagreements. Our section's strength rests in part on the ways in which members are willing to listen to one another and to entertain criticism seriously. The section's commitments to embracing different approaches within the qualitative tradition, including recent trends combining quantitative and qualitative research in "multi-method" projects, allow us to generate a broad range of debates without becoming self-satisfied or conformist—or even overly empathic.

Relatedly, let me take this opportunity to draw your attention to the five short courses the section is sponsoring or co-sponsoring at the upcoming APSA meeting. This year we have outstanding offerings, not just in terms of the number of courses, but because of their novel content and breadth of

coverage. I want to thank the scholars who will be leading the short courses (too numerous to be listed here, but described in full on pages 52–53 below) for their efforts.

Thanks are also owed to the outgoing president, Gary Goertz, for his leadership over the past two years, to Robert Adcock for his editorial expertise and his patience (with me, at least), and especially to Colin Elman whose organizational acumen, intelligence, and indefatigable decency make this section so worthwhile. The work of the many colleagues who have served on QMMR's committees this year is also truly appreciated.

Symposium. The Set-Theoretic Comparative Method: Critical Assessment and the Search for Alternatives

Part 1. Concerns about the Set-Theoretic Method

Problematic Tools: Introduction to Symposium on Set Theory in Social Science

David Collier
University of California, Berkeley
dcollier@berkeley.edu

> "To welcome ideas that shift the grounds on which our arguments previously found traction—that is our obligation as intellectuals." Wedeen (2014: 1)

Analysts who developed the set-theoretic comparative method (STCM) have formulated admirable goals for researchers who work in the qualitative and multi-method tradition. This method includes above all Charles Ragin's innovative approach of Qualitative Comparative Analysis (QCA), along with further systematization of the set-theoretic framework by other authors.[1] These colleagues are outstanding scholars and intellectual leaders in the field of methodology, and their advocacy of these goals is a major contribution.

However, the analytic tools employed by STCM have in many ways become an obstacle to achieving these admirable goals. For example, the system of fuzzy-set scoring appears to be problematic, poorly matched to a standard understanding of conceptual structure, and perhaps unnecessary in its present form. Computer simulations suggest that findings suffer from serious problems of stability and validity; and while the choice of simulations that match the method is a matter of some controversy, the cumulative weight of simulation results raises major concerns about STCM's algorithms—i.e., its basic, formalized analytic procedures.

Questions also arise about the cumbersome formulation of findings in what is often a remarkably large number of causal paths. Relatedly, some scholars question the STCM's rejection of the parsimonious findings, in the form of "net effects," routinely reported in other methodological traditions. Regarding applications, readily available software has encouraged publication of dozens of articles that appear to abandon key foundations of the method and rely far too heavily on these algorithms. Finally, STCM appears inattentive to the major, recent rethinking[2] of standards and procedures for causal inference[3] from observational data.

These problems raise the concern that the set-theoretic comparative method, as applied and practiced, has become disconnected from the underlying analytic goals that motivated Charles Ragin to create it.

This symposium explores these problems and seeks to identify promising directions for further work that pursues these same goals. In the symposium, this overall set of methods is referred to as STCM, and the designation QCA is used when the discussion is specifically focused on Ragin's contribution. For the convenience of readers, in anticipation that this essay might be read apart from the symposium, full citations to the other contributions are included in the bibliography.

Readers familiar with *Rethinking Social Inquiry: Diverse Tools, Shared Standards* (Brady and Collier 2004, 2010) will recognize the parallel with the present symposium. *Rethinking Social Inquiry* addressed an earlier, constructive initiative to redirect thinking about qualitative methods: King, Keohane, and Verba's (1994) *Designing Social Inquiry*—widely known as KKV. Their book had excellent overall goals, which centrally included a concern with systematizing qualitative research procedures that too often are unsystematic and unstandardized.[4] However, the book advocated specific tools for pursuing these goals that many scholars considered inappropriate, and in some respects counter-productive. *Rethinking Social Inquiry* sought to formulate methodological priorities and analytic tools more appropriate to qualitative research.

This symposium adopts the same perspective on the set-theoretic comparative method. The overall goals are excellent, and they centrally include a concern with systematizing qualitative research procedures that too often are unsystematic and

[1] Ragin 1987, 2000, 2008; and above all Goertz and Mahoney 2012, and Schneider and Wagemann 2012. QCA is understood here to include the crisp-set, multi-value, and fuzzy-set versions—i.e., csQCA, mvQCA, and fsQCA.

[2] This rethinking is discussed in Tanner's (2014) contribution to this symposium and in Collier (2014).

[3] The term causal inference is employed by some STCM authors (e.g. Goertz and Mahoney 2012; Schneider and Wagemann 2012), yet for other authors "causal interpretation" and "causal recipe" are preferred. The present discussion respects these distinctions, and uses "causal inference" as an umbrella term that encompasses these alternatives.

[4] David Laitin (1995), well known as a (creatively) eclectic scholar who is deeply engaged in both the qualitative and quantitative traditions, praised KKV as an important step toward "disciplining" political science.

unstandardized. However, the specific tools advocated for pursuing these goals have again been seen by many scholars as inappropriate, and in some respects counter-productive. Finally, in parallel, the symposium explores alternative tools that hold promise for more effectively pursuing these same goals.

The contributors to this symposium hope their essays will move the discussion forward, thereby seeking to sustain the same constructive spirit that the Brady and Collier volume sought to achieve. For both debates, a central recommendation is a return to more traditional qualitative methods,[5] which have in fact seen valuable innovation in recent years.[6]

Excellent Goals

There should be wide agreement that the set-theoretic comparative method has productively extended the horizon of scholars concerned with qualitative and multi-method research. STCM has introduced important new insights and challenged scholars to think about them carefully.

One example is the focus on asymmetric causation,[7] readily understood in terms of two types of causes: blocking causes that prevent a given outcome—as would occur with the absence of a necessary condition; versus triggering causes that ensure its occurrence—as would occur with the presence of a sufficient condition. The importance of this idea is seen in the fact that for many political scientists, it initially produces puzzlement to argue that the occurrence versus non-occurrence of an outcome could have a different explanation. STCM has taken an idea that is too often seen as puzzling, and shown that this idea is indispensable.[8]

Other key contributions include a new approach to studying equifinality, i.e., multiple causal paths; distinctive tools for assessing causal interactions; an insistence on the importance of context; a strong commitment to mobilizing case knowledge; and a central emphasis on the interplay of theory and case knowledge that brings together deductive and inductive approaches to gaining new insights.

The field has definitely benefitted from STCM's advocacy of these goals. Indeed, rather than declaring qualitative and quantitative methods to be "two cultures" (Goertz and Mahoney 2012), scholars might instead celebrate the contributions of STCM in advocating these goals for the broader field of methodology—though of course, some quantitative researchers have long promoted many of the same goals.

[5] This recommendation echoes the conclusions of Seawright (2005: 41; 2014); Lucas and Szatrowski (2014); and Collier (2014).

[6] Possibly the single most important innovation in qualitative methods of the past several years is Bennett's (2014) reframing of process tracing, summarized in this symposium, which builds on Humphreys and Jacobs (2013) remarkable new framework for multi-method research.

[7] The term asymmetric causation is also used to characterize a unidirectional causal relation between a given pair of variables. That is not the meaning intended here.

[8] Ironically, the idea of asymmetric causal patterns is more standard in other, very different domains. For example, it is presented in as conventional a source as Fahnestock and Secor's (1982: 132–146; 2nd and 3rd editions 1990 and 2003) textbook for teaching undergraduates good writing skills.

Part 1. Concerns about the Set-Theoretic Method

At the same time, questions have arisen about STCM's tools, and the four contributions to Part 1 of this symposium explore these questions. The present essay raises a number of concerns about these methods, provides an overview of the symposium, and poses questions to suggest future directions for more effectively pursuing these same goals.

The other three authors in Part 1: (a) challenge the idea that set-theory can be justified in part because it reflects the structure of meaning in natural language (Lakoff); (b) argue that an emphasis on constructing well-bounded concepts is a separate matter from embracing the logic and procedures of set theory as a guiding framework for research (Sartori); and (c) raise a number of questions about STCM's approach to causal inference (Tanner).

Set Theory and Natural Language (Lakoff). A recurring theme in discussions of set-theoretic methods is that this approach is compelling in part because it reflects the structure of natural language.[9] The relationship to natural language is addressed here in an interview with the prominent cognitive linguist George Lakoff, whose work is periodically evoked in arguments that justify set-theoretic approaches—including fuzzy-set analysis.[10]

Lakoff dissents from these arguments, drawing on strong evidence that the organization of meaning in natural language is not based on classical categorization, with necessary and sufficient conditions for category membership. While some concepts are well-bounded, prototype theory suggests that a great many are not. Even for those that are well-bounded, prototype theory points to the importance of not reifying these boundaries. This raises serious concerns about the set-theory template.

In addition, Lakoff discusses Zadeh's fuzzy logic, expressing admiration for its application to engineering—yet arguing it is not generally a good match for the structure of meaning in natural language. He also notes the large difference between Zadeh's fuzzy logic and Ragin's procedure for scoring fuzzy-set membership. Lakoff suggests, given the fixed numerical values assigned in Ragin's fuzzy-set scoring, that this analytic procedure should in fact not be considered a fuzzy method.

The Quest for Well-Bounded Concepts (Sartori). Crafting well-bounded concepts has long been a central priority for methodologists and also for applied researchers. Giovanni Sartori is a leading advocate of this practice, and his work is evoked by advocates of set theory and associated systems of logic.[11] For the purposes of social science, Satori insists on classical categorization, based on necessary and sufficient criteria for category membership. However, he rejects the application of set theory as a central technique in qualitative re-

[9] Ragin (2008: especially 38; also 2, 13, 97); Goertz and Mahoney (2012: 11–12, 16–18); Schneider and Wagemann (2012: 7; 2013: 21–22).

[10] See Ragin (2000: 6, 171; 2008: 98); Goertz and Mahoney (2012: 16, 18); Schneider and Wagemann (2013: 21–22).

[11] Ragin (2000: 321, 328; 2008: 98); Goertz and Mahoney (2012: 6, 12, 16, 139, 148).

search. He draws on his long-standing distinction between the unconscious thinker, who fails to reflect on concepts and methods, and the over-conscious thinker, who is counter-productively focused on techniques that may well be more complex than is productive for the task at hand (Sartori 1970: 1033–1040).

Sartori seeks to follow a middle path. For example, in developing his well-known ladder of abstraction, he did make limited use of ideas from logic, and he does think that scholars should be familiar with the tools of logic. However, he rejects the proposal that the set-theoretic approach and associated forms of logic should become a dominant framework, because they may lead the researcher to become bogged down in unproductive techniques.

Tools for Causal Inference (Tanner). Sean Tanner, a scholar of public policy, makes a two-fold argument. Using many examples, he reviews QCA's tools for causal inference and finds them problematic. Tanner also addresses the argument advanced by QCA scholars that their distinctive approach to causal inference is valuable for the study of public policy. Tanner suggests that evaluation research is an area of policy analysis that places especially strong demands on tools of causal inference, and policy evaluation is therefore an appropriate "crucial case" for assessing QCA's value for public policy studies.

Tanner offers a detailed comparison between policy analyses based on QCA, as opposed to studies that follow today's standard norms for policy research. He finds that the conventional studies yield the kind of insights policy analysts urgently need, whereas the QCA analyses offer findings that are too often unhelpful and uninterpretable. The discussion includes such topics as gaining insight into causal interactions; the importance of "net effects thinking," an approach strongly questioned by QCA scholars; the problem that QCA scoring too often yields measurements that are difficult for policy analysts to interpret; the extremely large—and therefore, again, hard to interpret—number of causal paths often yielded by QCA; and the exceedingly small number of cases per causal path, sometimes just one or two.

Throughout, Tanner emphasizes current standards for causal inference that mandate careful choices in making inferences from observational data—and indeed, from all kinds of data. By these standards, he finds QCA to be seriously deficient.

Part 2. Where Do We Go From Here?

Building on these commentaries, the second part of the symposium asks: "Where do we go from here?" Topics include: (a) the challenge of developing simulations that are appropriate for evaluating the stability and validity of findings derived from STCM; (b) alternative procedures for analyzing partial membership in categories; (c) contrasting approaches to the study of interactions among different combinations of explanatory factors; and (d) a new approach to process tracing that moves beyond earlier criteria of necessity and sufficiency for evaluating causal inference.

Evaluating Simulations (Krogslund and Michel). The use of computer simulations to evaluate alternative methods is now standard across the social sciences. Correspondingly, the most important area of assessment and innovation in discussions of STCM currently involves simulations that evaluate the stability and validity of findings. This topic merits close attention here.

STCM employs a complex set of algorithms,[12] and a growing number of studies have raised concerns that these algorithms are highly sensitive to small changes in measurement decisions and to shifts in the parameters that must be set for causal inference.[13] A recurring finding has been a tendency to generate false positives.

From within the STCM tradition, Schneider and Wagemann's overview of simulations and robustness is more encouraging, but they conclude with great caution: "QCA is not vastly inferior to other comparative methods in the social sciences" (2012: 294). This is faint praise, given the numerous, sharp critiques of the stability of findings based on conventional quantitative analysis—obviously a key method of comparison.

Considerable attention is being devoted to the crucial question of how to design tests that reproduce the algorithms employed in these methods, and Thiem (2013), for example, has made a key contribution. Readers should be alerted to forthcoming debates that are unfortunately not yet available (including to this author) at the time of this newsletter's publication.[14]

The jury is definitely still out in terms of assessing specific simulation tests—at the same time that the cumulative evidence raises very strong concerns about STCM's tools. These concerns are reinforced by the fact that STCM has not incorporated into its analytic procedures a recognition of the recent transformation of thinking about causal inference, as practiced in all social science methods, which is discussed—as noted above—by Tanner and by Collier (2014).

[12] Algorithms are understood as systematized procedures for making calculations, often implemented with computer software. QCA's ensemble of algorithms includes, for example, procedures that address contradictions, logical remainders, minimization, sufficiency scores, minimum frequency, consistency, coverage, and the probabilistic criteria for causal inference.

[13] Hug 2013; Seawright 2013, 2014; Kurtz 2013; Krogslund, Choi, and Poertner 2013; Krogslund and Michel 2014; Lucas and Szatrowski 2014.

[14] Some time ago, Ragin and Rihoux (2004: 22–24) argued that an earlier evaluation based on hypothetical data was not suitable to QCA. Readers should watch for important, forthcoming exchanges. Thus, an important debate has been generated by Lucas and Szatrowski's (2014) simulations. Ragin (2014) has written a commentary on their article, and they in turn have responded to his commentary. Given the norms of the journal *Sociological Methodology* where this exchange will appear, these comments are not available to other contributors to that symposium until the time of their actual publication. In addition, Thiem's (2014) commentary on Hug (2013) is scheduled to be included in the next issue of the present newsletter, accompanied by a response from Hug. When these publications become available, they should and will play an important role in these debates.

Krogslund and Michel's contribution to this symposium seeks to advance this important search for appropriate simulations by evaluating results from a "drop-one" sensitivity test of QCA. Their initial finding from this test—intriguingly—inverts Arend Lijphart's (1971: 686) traditional formulation of the "many variables, small-N problem" in comparative research. Thus, Krogslund and Michel's findings suggest QCA might, ironically, have a "fewer variables, larger-N" problem. That is to say, findings appear more unstable, to the degree that the analyst focuses on more cases and a smaller number of explanatory variables. This counter-intuitive result leads them to scrutinize the properties of this test, as well as its appropriateness to the analytic procedures of QCA—with a particular emphasis on how logical remainders (i.e., empty rows in the truth table) are treated.

Simulation tests will be crucial in the ongoing evaluation of QCA, and Krogslund and Michel illustrate the kind of painstaking, fine-grained analysis needed to adequately assess the appropriateness of specific tests.

Measuring Partial Membership in Categories (Elkins). Fuzzy-set analysis was introduced as a tool for measuring partial membership in categories, and Zachary Elkins' contribution evaluates alternative approaches to the study of partial membership. He raises some of the same concerns advanced in this symposium by Lakoff (2014) and Tanner (2014) about the ambiguities of scoring fuzzy-sets. On the basis of these concerns, Elkins advocates attention to three issues: the conceptual structure of the categories, homogeneity within categories, and degree of membership.

Elkins focuses on a substantive example that is highly salient to the present discussion: the comparative study of constitutions. The categories of presidentialism, semi-presidentialism, and parliamentarism are widely recognized and—presumably—extremely well defined, yielding an excellent opportunity to explore ideas about full membership and partial membership. Elkins applies three scaling techniques to his data—MIMIC Modeling (Multiple-Indicators Multiple Causes), similarity-based measures and latent-class analysis—using them to assess the structure and heterogeneity of the categories and degrees of membership. He suggests that these methods have many advantages over fuzzy-set scoring and they yield important new insights into the complexities of these extremely well-known categories of legislative-executive relations.

Studying Interactions (Braumoeller). A major concern of social science methodology is with how causal patterns differ when distinct combinations of causes interact. This topic is addressed under various rubrics, including the study of interactions and of contextual effects. Concern with causal patterns such as these is a hallmark of the qualitative tradition. Bear Braumoeller's contribution systematically compares four approaches to the study of interactions: QCA's focus on causal combinations, interaction terms in regression analysis, stochastic frontier modeling, and Boolean logit.

These methods differ in their approach to measurement, the role of coefficients, the treatment of thresholds, conceptualization of the interactions themselves, and data requirements. Clearly, none of these models is the "correct" one, and Braumoeller's contribution is an excellent point of departure for understanding the opportunities, limitations, and trade-offs that arise in studying interactions. It establishes a broad agenda for future work on this crucial topic.

Process Tracing: Beyond the Criteria of Necessity and Sufficiency (Bennett). Process tracing is a fundamental tool relevant to all forms of research, both qualitative and quantitative, and Andrew Bennett's contribution is a major stride forward.[15] A central goal of ongoing research on process tracing has been to provide criteria for evaluating the four process-tracing tests proposed some time ago by Van Evera (1997: 31-32): straw-in-the-wind, hoop, smoking-gun, and doubly-decisive tests.

One approach to mapping the relationship among these tests, adopted by Bennett (2010) and Collier (2011), had been to differentiate according to whether each test is necessary and/or sufficient for affirming a given causal inference. Goertz and Mahoney (2013: 279) have noted this approach as an example of applying set theory and related forms of set logic to qualitative methods.

However, Bennett has now moved the process-tracing literature well beyond this approach, based on his application of Bayesian analysis. In Bennett's new formulation, the overall goal is the same: evaluating the power of process-tracing evidence for testing a given hypothesis. The Bayesian approach systematizes insights into this probative power based on three criteria: (1) whether positive evidence is found, (2) the researcher's prior confidence that the hypothesis is correct, and (3) the likelihood ratio, i.e. the odds of finding positive evidence if the explanation is correct versus those of finding the same evidence even if the explanation is false.

In this framework, the four traditional tests can be situated within a spectrum of possibilities. This spectrum is established based on the degree to which (a) finding or not finding the evidence has (b) a modest or strong effect on (c) the posterior assessment of whether the hypothesis is supported.

Bennett argues that the ideas of necessary and sufficient are superseded because they are too categorical, and he advocates this more flexible approach. In a sense, the differentiation of four process-tracing tests has also been superseded, and these tests can now most usefully be seen simply as useful benchmarks within this continuous spectrum of alternative inferences.

With the Bayesian method, a key question is whether researchers should "fill in the numbers," using the Bayesian algorithms to actually make calculations; or instead employ the method as a useful heuristic for reasoning about process tracing. This second version would correspond to McKeown's (2004: 158–162) idea of "folk Bayesianism." Bennett is open to both alternatives—and in either case, a central point is that the categorical framing of necessity and sufficiency as a basis for evaluating process-tracing tests is replaced by a framing based on the idea of continuous gradations.

It should be added that scholars identified with STCM

[15] More broadly, see the major new book on process tracing by Bennett and Checkel (2014).

have also recently introduced new perspectives on process tracing, and this method promises to be a key area of future innovation.[16]

Some Further, Exploratory Questions

The essays in this symposium call for a fundamental rethinking of the tools employed by the set-theoretic comparative method. This section poses several exploratory questions—some of which correspond to innovations currently being developed by STCM scholars—that may be useful points of reference in this rethinking. The first three questions focus on measurement, the others causal inference.

Is fuzzy-set scoring really fuzzy? Lakoff observes that given the fixed values assigned in fuzzy-set scoring, the method is in fact not fuzzy. He notes the contrast with Zadeh's scoring procedure, in which the assigned values are themselves fuzzy, rather than fixed.

Is fuzzy-set scoring viable? Lakoff, Tanner, and Elkins are all concerned about the lack of clear standards for assigning scores, and hence for problems with the interpretability of scores. For example, if one finds the initial designation of full membership in the set to be ambiguous, then the rest of the scale becomes ambiguous. Tanner is also concerned that even with the fuzzy-set version, the method yields aggregated causal paths that lose a great deal of information.

These concerns point to a further question: do the standard indicators on which fuzzy-set scores are often based convey more useful information than the fuzzy-set scores themselves? These indicators have the advantage of being formulated in terms of familiar and more readily interpretable units of measurement.

Is fuzzy-set scoring necessary? Correspondingly, might it be preferable if STCM scholars made greater use of standard indicators, rather than fuzzy-set scoring? If one accepts the possibility that fuzzy sets are in fact not fuzzy, then perhaps not a great deal would be lost. Analysts would retain the more readily intelligible units of measurement. At the final step in causal inference, when the fuzzy-set findings are dichotomized for entry into the truth table, STCM scholars could draw on the insights gained in the course of the analysis to make what might be better-informed judgments about establishing the cut-points for dichotomization.

This approach embraces Ragin's (2000: 171) recommendation that recoding fuzzy membership scores in the course of the analysis should be considered standard practice. In addition, the important goal of eliminating "irrelevant variation" (Ragin 2000: 161–63) would be achieved through the dichotomization introduced at this final step in the process—again, hopefully guided by these better-informed judgments that emerge in the course of the study.

A final point: In addition to encouraging greater use of standard indicators, this framework could incorporate a broader set of tools. For example, Elkins' methods for analyzing partial membership in categories might be considered just as appropriate as fuzzy-set scoring for these STCM applications.

What are the next steps in analyzing interactions? Tanner expresses the concern that QCA's approach to studying interactions routinely does not yield productive insights, and Braumoeller takes a large stride toward extending the discussion by noting major tradeoffs among four alternative models of interactions, including QCA.

Braumoeller's line of analysis must be developed much further—for example, analysts might wish to add a fifth approach. Tanner's examples include the compelling analysis of an interaction that relies on a simple two-by-two table, involving welfare interventions for teen-aged mothers. Tools for analyzing standard cross-tabulations are much neglected in today's social science, and obviously there is a specific "algorithm" that one can follow for such analysis. This algorithm might be added as a fifth model for studying interactions.

Have the algorithms and software taken over, and has case knowledge been eclipsed? This symposium has suggested the answer is yes, and this question merits continuing attention. Consider the many empirical studies using QCA that are analyzed by Tanner and by Krogslund and Michel—along with dozens of additional empirical articles using the method. The concern does indeed arise that the method is in effect reduced to the algorithms—which are all too readily applied using QCA software. The intensive use of case knowledge is often not in evidence.

If we look at the trajectory of other innovative methods, we see that the widespread availability of software can be both a blessing and a curse. In the case of structural equation modeling, it opened a Pandora's Box of bad applications (Steiger 2001). One worries that the same distortion of the method has occurred with QCA.

How should one think about problems of stability and validity of findings that emerge in simulations? Error and the "DGP." As scholars conduct more simulation tests, it is essential to ask why these tests often reveal problems of stability and validity. One possibility is that STCM lacks adequate tools for dealing with a number of issues, including measurement error, problems of model specification, and potentially a random element in the data.

In evaluating these issues of the stability and validity of findings, STCM scholars should consider a concept that is crucial to methodological discussions today: the underlying "data-generating process" (DGP), which focuses attention on what is now the standard insight that causal assessment involves reaching conclusions about the DGP on the basis of the particular set of observed values. This is a useful way of bringing into sharp focus the daunting challenges of causal inference. Some simulation tests specifically evaluate STCM within this DGP framework (Krogslund and Michel 2014), and these tests reinforce concerns about the stability and validity of findings. Additional work along these lines will make a large contribution to evaluating STCM's analytic tools.

Two more questions about the stability and validity of

[16] See Mahoney (2012); Schneider and Rohlfing (2013); Rohlfing and Schneider (2013); Beach and Pedersen (2013); Rohlfing (2014).

findings also merit attention.

Is there an asymmetry in the treatment of false negatives and false positives? STCM has procedures such as its probabilistic criteria to help guard against false negatives. Is it possible that parallel procedures to guard against false positives are lacking? Regarding false negatives: in the real world, given problems that include measurement error, a true causal relationship of necessity and/or sufficiency may be present, yet it might be imperfectly reflected in the data. For example, if the analysis employs a two-by-two table, some cases might be located in the "wrong" cells, vis-à-vis the true causal pattern. It appears that STCM has procedures for addressing this problem, thereby guarding against a potential false negative.

However, it is not clear that there is a procedure for addressing the opposite problem, which could yield a false positive. Thus, cases might be located in the "right" cells for inferring necessity and/or sufficiency, yet in an analysis free of error they might in fact be located in the "wrong" cells. The resulting inference runs the risk of being a false positive.

The question of whether STCM has appropriate tools for addressing both false negatives and false positives requires much further discussion.

How should the empty rows in the truth table—the logical remainders—be addressed? For scholars who are not part of the STCM tradition, the treatment of empty rows—i.e. combinations of conditions not found in the empirical cases—is puzzling. The problem of "limited diversity" addressed in STCM is definitely important, and the counterfactual reasoning that underlies causal inference routinely involves empty rows. Yet STCM appears, overall, problematic in meeting the challenge posed by standard norms of causal inference; and employing an extremely complicated analytic procedure to address what may well not be the most compelling aspect of this deficit seems questionable.

The truth table, a foundation of QCA, is a logical construct, and as a logical construct it encompasses all possible combinations of explanatory conditions that could be matched with both the occurrence and non-occurrence of the outcome. As is widely noted, with additional explanatory conditions this leads to an exponential increase in the overall number of rows, and also to a dramatic increase in empty rows. This yields a cascade of complications for the method, along with the need for complex algorithms to address these complications.

In the actual practice of STCM, the truth table might more usefully be treated not as a logical construct, but as a valuable form of data display. Correspondingly, the focus on the empty paths could be dropped from the method.

In the field of comparative-historical analysis—for which STCM is intended to have great value—a focus on empty paths would be non-standard. For example, in *Shaping the Political Arena* (R. Collier and D. Collier 1991), we could not possibly have been able to—or wanted to—address a large number of empty rows. Many books have a concluding chapter that somewhat speculatively places the cases analyzed in a wider comparative perspective—including potentially some comments on empty rows. This is valuable, and it strengthens causal inference. But elaborate attention to empty rows is emphatically not a cornerstone of the comparative-historical method.

If the truth table were simply treated as a valuable data display, perhaps the concern with empty rows could simply be dropped, and attention might usefully focus on other limitations of STCM's procedures for causal inference.

To reiterate, these are exploratory questions that seek to advance the discussion. Some questions correspond to innovations currently being developed by STCM scholars—although the findings of the present essay suggest that the overall goals of the method may be more effectively served by turning to different tools.

Conclusion: Restoring Ragin's Dialogue between Ideas and Evidence

Taken together, these questions point to the need for a fairly drastic reevaluation of the tools employed in the set-theoretic method. To place this reevaluation in perspective, a concluding point should be made about a central foundation of the method, which grows out of the work of Charles Ragin.

One of Ragin's fundamental scholarly contributions is his conception of social research as a dialogue between ideas and evidence. This conception is important for QCA, but very crucially also in his many non-QCA books and articles on methods (Ragin and Zaret 1983; Ragin and Becker 1989, 1992; Ragin 1994, 2004; Ragin and Amoroso 2010). This trajectory in his work can readily be seen as a creative extension of the long sociological tradition that includes, for example, the Lazarsfeld elaboration model and the constant comparative method of grounded theory. This tradition remains a cornerstone of good research.

As noted, in much of the applied work using QCA, this component may well have disappeared. It seems that the role of case knowledge—and the dialogue between ideas and evidence—is being eclipsed by the algorithms and the computer software. Setting aside the algorithms—combined with concentrating on case studies and process tracing, as is advocated here—would bring the focus back to Ragin's larger contribution.

Next Steps: Following through on this Symposium

The *Qualitative and Multi-Method Research* newsletter is committed to publishing in its next issue a comment on this symposium from the perspective of STCM, a response from the standpoint of the symposium, a comment on the recent Hug (2013) article, and a response from Hug. This should be a valuable exchange, and we strongly welcome it.

Regarding this evolving discussion, an observation should be made about a direction the debate hopefully will *not* take. In earlier exchanges, the skeptical evaluation of QCA by two commentators was challenged as reflecting the limited perspective and "defensive reactions" of quantitative researchers.[17] In another exchange, a skeptical evaluation by a third

[17] See Ragin and Rihoux's (2004: 22) comments on the evaluations by Lieberson and Seawright.

commentator was dismissed as "what one would expect a quantitative researcher to believe."[18]

These responses by STCM scholars took the discussion down the wrong path, particularly because all three of these commentators are specifically well-known as strong critics of conventional quantitative methods.[19] In the spirit of this newsletter's commitment to multi-method research, we have sought to meet a key standard: to the extent possible and feasible, contributors should have a broad view of methodology that transcends limitations such as these.

In that framework, we greatly look forward to the continuing discussion.

References

Beach, Derek, and Rasmus Brun Pedersen. 2013. *Process-Tracing Methods: Foundations and Guidelines*. Ann Arbor: University of Michigan Press, 2013.

Bennett, Andrew. 2010. "Process Tracing and Causal Inference." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, eds. Henry E. Brady and David Collier. Lanham, MD: Rowman and Littlefield, 207–219.

Bennett, Andrew. 2014. "Process Tracing with Bayes: Moving beyond the Criteria of Necessity and Sufficiency." *Qualitative & Multi-Method Research* 12 (1): 46–51.

Bennett, Andrew and Jeffrey Checkel. 2014. *Process Tracing in the Social Sciences: From Metaphor to Analytic Tool*. Cambridge: Cambridge University Press.

Brady, Henry E. 1995. "Doing Good and Doing Better." *The Political Methodologist* 6 (2): 11–19.

Brady, Henry E. and David Collier, eds. 2004, 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman and Littlefield.

Braumoeller, Bear. 2014. "Analyzing Interactions: Four Alternative Models." *Qualitative & Multi-Method Research* 12 (1): 41–46.

Collier, David. 2011. "Understanding Process Tracing." *PS: Political Science and Politics* 44 (4): 823–830.

Collier, David. 2014. "QCA Should Consider Abandoning the Algorithms." *Sociological Methodology* 44: Forthcoming.

Collier, Ruth Berins and David Collier. 1991. *Shaping the Political Arena: Critical Junctures, the Labor Movement, and Regime Dynamics in Latin America*. Princeton: Princeton University Press.

Elkins, Zachary. 2014. "Measuring Partial Membership in Categories: Alternative Tools." *Qualitative & Multi-Method Research* 12 (1): 33–41.

Fahnestock, Jeanne, and Marie Secor. 1982. *A Rhetoric of Argument*. 2nd and 3rd eds, 1990 and 2003. New York: McGraw-Hill.

Goertz, Gary, and James Mahoney. 2012. *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton, New Jersey: Princeton University Press.

Goertz, Gary and James Mahoney. 2013. "For Methodological Pluralism: A Reply to Brady and Elman." *Comparative Political Studies* 46 (2): 278–285.

Hug, Simon. 2013. "Qualitative Comparative Analysis: How Inductive Use and Measurement Error Lead to Problematic Inference." *Political Analysis* 21 (2): 252–265.

Humphreys, Macartan and Alan Jacobs. 2013. "Mixing Methods: A Bayesian Unification of Qualitative and Quantitative Approaches." Paper presented at the Annual Meeting of the American Political Science Association, Chicago, IL.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

Krogslund, Chris and Katherine Michel. 2014a. "Evaluating Set-Theoretic Methods: Finding Appropriate Simulations." *Qualitative & Multi-Method Research* 12 (1): 25–33.

Krogslund, Chris and Katherine Michel. 2014b. Recovering the Data-Generating Process: How Well Do Set-Theoretic Methods Perform? Revised version of a Paper Presented at the 2014 annual meeting of the Midwest Political Science Association, April 3-6, Chicago IL.

Krogslund, Chris, Donghyun Danny Choi, and Matthias Poertner. 2013. "Fuzzy Sets on Shaky Ground: Parameter Sensitivity and Confirmation Bias in fsQCA." Revised version of a Paper Presented at the Annual Meeting of the American Political Science Association, Chicago.

Kurtz, Marcus. 2013. "The Promise and Perils of Fuzzy-set/Qualitative Comparative Analysis: Measurement Error and the Limits of Inference." Presented at the Annual Meeting of the American Political Science Association, Chicago.

Laitin, David D. 1995. "Disciplining Political Science." *American Political Science Review* 89 (2): 454–456.

Lakoff, George. 2014. "Set Theory and Fuzzy Sets: Their Relationship to Natural Language. Interview with George Lakoff, Conducted by Roxanna Ramzipoor." *Qualitative & Multi-Method Research* 12 (1): 9–14.

Lieberson, Stanley. 1985. *Making It Count: The Improvement of Social Research and Theory*. Berkeley: University of California Press.

Lijphart, Arend. 1971. "Comparative Politics and the Comparative Method." *American Political Science Review* 65 (3): 682–693.

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

McKeown, Timothy J. 2004. "Case Studies and the Limits of the Quantitative Worldview." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Henry E. Brady and David Collier, eds. (Lanham, MD: Rowman and Littlefield), 139–167.

Mahoney, James. 2012. "The Logic of Process Tracing Tests in the Social Sciences." *Sociological Methods & Research* 41 (4): 570–597.

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

Ragin, Charles C. 1994. *Constructing Social Research*. Thousand Oaks, CA: Pine Forge Press.

Ragin, Charles C. 2000. *Fuzzy-Set Social Science*. Chicago: University of Chicago Press.

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

Ragin, Charles C. 2014. "Lucas and Szatrowski in Critical Perspective." *Sociological Methodology* 44: Forthcoming.

Ragin, Charles C. and Lisa M. Amoroso. 2010. *Constructing Social Research: The Unity and Diversity of Method*. 2nd ed. Los Angeles: Sage Publications.

Ragin, Charles C. and Howard S. Becker. 1989. "How the Microcom-

[18] See Goertz and Mahoney's (2013: 279–80) response to the evaluation by Brady.

[19] Regarding the critiques of conventional quantitative methods by these three authors, see Lieberson (1985); Brady (1995), Brady and Collier (2004); and for Seawright, Chapters 1, 2, 9, and 10 in Brady and Collier (2004), and subsequently Seawright (2007). It also merits note that to the extent that scholars might be classified as "quantitative" researchers, they routinely also have a strong background in mathematics and logic—tools essential to these discussions.

puter Is Changing Our Analytic Habits." In *New Technology in Society: Practical Applications in Research and Work*. (New Brunswick, NJ: Transaction Publishers), 47–55.
Ragin, Charles C. and Howard S. Becker. 1992. *What Is a Case? Exploring the Foundations of Social Inquiry*. Cambridge: Cambridge University Press.
Ragin, Charles C. and Benoît Rihoux. 2004. "Qualitative Comparative Analysis (QCA): State of the Art and Prospects." *Qualitative Methods* [subsequently renamed *Qualitative and Multi-Method Re search*] 2 (2): 3–13, 22–24.
Ragin, Charles and David Zaret. 1983. "Theory and Method in Comparative Research: Two Strategies." *Social Forces* 61 (3): 731–754.
Rohlfing, Ingo. 2013. "Comparative Hypothesis Testing Via Process Tracing." *Sociological Methods & Research*. Online First DOI: 10.1177/0049124113503142.
Rohlfing, Ingo and Carsten Q. Schneider. 2013. "Improving Research on Necessary Conditions: Formalized Case Selection for Process Tracing after QCA." *Political Research Quarterly* 66 (1): 220–230.
Sartori, Giovanni. 1970. "Concept Misformation in Comparative Politics." *American Political Science Review* 64 (4): 1033–1053.
Sartori, Giovanni. 2014. "Logic and Set Theory: A Note of Dissent." *Qualitative & Multi-Method Research* 12 (1): 14–15.
Schneider, Carsten Q. and Ingo Rohlfing. 2013. "Combining QCA and Process Tracing in Set-Theoretic Multi-Method Research." *Sociological Methods & Research* 42 (4): 559–597.
Schneider, Carsten Q. and Claudius Wagemann. 2013. "Fuzzy Sets Are Sets—A Reply to Goertz and Mahoney." *Qualitative & Multi-Method Research* 11 (1): 20–23.
Schneider, Carsten Q. and Claudius Wagemann. 2012. *Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis*. Cambridge: Cambridge University Press.
Seawright, Jason. 2005. "Assumptions, Causal Inference, and the Goals of QCA." *Studies in Comparative International Development* 40 (1): 39–42.
Seawright, Jason. 2007. "Democracy and Growth: A Case Study in Failed Causal Inference." In *Regimes and Democracy in Latin America: Theories and Methods*, ed. Gerardo L. Munck. Oxford: Oxford University Press, 179–198.
Seawright, Jason. 2013. "Warrantable and Unwarranted Methods: The Case of QCA." Presented at the Annual Meeting of the American Political Science Association, Chicago.
Seawright, Jason. 2014. "Limited Diversity and the Unreliability of QCA." *Sociological Methodology* 44: Forthcoming.
Steiger, James H. 2001. "Driving Fast in Reverse: The Relationship between Software Development, Theory, and Education in Structural Equation Modeling." *Journal of the American Statistical Association* 96 (453): 331–338.
Tanner, Sean. 2014. "Evaluating QCA: A Poor Match for Public Policy Research." *Qualitative & Multi-Method Research* 12 (1): 15–25.
Thiem, Alrik. 2013. "Membership Function Sensitivity of Descriptive Statistics in Fuzzy-Set Relations." *International Journal of Social Research Methodology*. Online First DOI: 10.1080/13645579.2013.806118.
Thiem, Alrik. 2014. "Data and Measurement Error in Qualitative Comparative Analysis: An Extended Comment on Hug (2013)." *Qualitative and Multi-Method Research* 12 (2): Forthcoming.
Van Evera, Stephen. 1997. *Guide to Methods for Students of Political Science*. Ithaca, NY: Cornell University Press.
Wedeen, Lisa. 2014. "Letter from the Section President." *Qualitative and Multi-Method Research* 12 (1): 1–2.

Set Theory and Fuzzy Sets: Their Relationship to Natural Language

An Interview with George Lakoff[1]
Interview Conducted by Roxanna Ramzipoor
University of California, Berkeley
roxanna.ramzipoor@berkeley.edu

Background: A recurring argument of scholars who advocate set theory and fuzzy sets for social science is that this framework is valuable and appropriate in part because it reflects the structure of meaning in natural language.[2] George Lakoff has written extensively on these topics and is cited by these scholars as an authority. In this interview, Lakoff synthesizes a large body of research in linguistics and cognitive science, which contends that natural language is not set-theoretic in structure. He also explores Lotfi Zadeh's fuzzy logic, emphasizing both its creative applications in engineering and its poor fit with most features of natural language. Finally, Lakoff discusses the basic contrast between Zadeh's fuzzy logic and Charles Ragin's fuzzy-set scoring. Lakoff emphasizes that he is not in a position to judge the substantive contribution of Ragin's method. However, it does not rely on an empirically adequate account of natural language; and because the scoring is based on fixed numerical values, rather than fuzzy distributions, Ragin's scoring does not qualify as a fuzzy method.

Q: Is natural language set-theoretic?

A: Standard set theory—I will discuss fuzzy sets later on—does not capture the structure of natural language. Categorization is one of the primary means by which humans use natural language to understand the world. The set-theoretic view is based on what we call the classical theory of categorization. This theory posits that we categorize objects or experiences in terms of inherent properties that are necessary and/or sufficient for category membership. In standard set theory, objects and experiences are understood as either inside or outside a specific category. Anything that has a given combination of inherent properties is inside the category, and anything that does not have these properties is outside the category. In the classical theory, there are no degrees of category membership: It's *in* or *out*.

However, this set theoretic concept of categorization does not correspond to the way people categorize objects and experiences using natural language. As Rosch (1975, 1977) has found, we instead categorize in terms of prototypes and family resemblances. Unlike set theory, the theory of prototypical categorization, as extended in my book *Women, Fire, and Dangerous Things* (Lakoff 1987; hereafter *WFDT*), is sufficiently flexible to capture the category structure of natural language. For example, the prototypical chair has a back, seat, four legs,

[1] This interview was conducted in December 2013. Lakoff subsequently revised and amended the text and provided the bibliography.

[2] For references, see footnote 10 in David Collier's Introduction to this symposium.

and usually armrests; it is used for sitting. But we also have nonprototypical chairs—wheelchairs, hanging chairs, and dentists' chairs. We understand the nonprototypical cases as chairs by virtue of their relationship to the prototype. In this way, category structures defined by such prototypes map directly onto the way we conceptualize and describe objects using natural language.

The idea of family resemblances becomes crucial here. We understand nonprototypical chairs as being chairs because they bear a *family resemblance* to the prototype. Family resemblances are not linearly ordered; one thing can bear a resemblance to another in various ways along various dimensions. The theory of radial categories in *WFDT* provides an account of what constitutes a family resemblance. Real radial categories are more complex, incorporating metaphoric and metonymic relations. The radial category structure defines not only what is "sufficiently close," but also the nature of the difference between prototype and object.

Unlike standard set-theoretic categorization, which does not allow us to readily categorize objects or ideas that stretch the limits of a set, prototypes and family resemblances can be systematically extended to define relationships between categories. Modifiers, which I have described as *hedges* (Lakoff 1973), include expressions such as "strictly speaking," "loosely speaking," and "par excellence." The hedges change the category boundaries in ways dependent on context and reflect the structure of prototypical and nonprototypical members. *Strictly speaking* picks out the central examples. *Loosely speaking* somewhat extends the prototype boundaries and eliminates the prototypical examples. *Par excellence* again redraws the category boundaries to include only the best prototypical examples. For example, a robin is a bird *par excellence*, while chickens and ostriches are not.

There are also types of prototypes with different properties: social stereotypes, typical cases, paragons, nightmare cases, salient exemplars, generators, and so on. Standard set theory is too rigid to capture the relationships between categories and families of categories.

We have learned a lot about the structure of family resemblances—specifically, how we pick out things that are similar in certain respects, and different in others. The theory of conceptual metaphor—which is now grounded in new work on the neural theory of thought and language and experimental research on embodied cognition—has been a major advance in understanding real cognitive structure. And of course, metaphor is not just a poetic or literary device, but a basic feature of largely unconscious everyday thought and language. Conceptual metaphors are frame-to-frame mappings that allow a source frame to project content onto a target frame, thus greatly enriching our means of conceptualization. Conceptual metaphors either have a direct bodily grounding or are decomposable into more primitive metaphors that have a bodily grounding. The system of embodied conceptual metaphor is the broad super-structure of our system of concepts. Conceptual frames and metaphors form networks called "cascades," which are used in characterizing the content of categories.

Set theory has none of this real cognitive apparatus.

Q: Say more about the contribution of Rosch.

A: Rosch was a pioneer in breaking with classical categorization. Her experiments in the mid-1970s strongly support the idea that human categorization is organized around prototypes and family resemblances. As an undergraduate at Reed College, she wrote her honors thesis on Wittgenstein and engaged with his concept of family resemblances—the idea that objects in a given category do not necessarily have common properties, but resemble each other like family members who have different combinations of shared features. While this famous idea is only briefly articulated in Wittgenstein's writings, it became crucial for Rosch's research.

Later, as a graduate student at Harvard, Rosch worked closely with Roger Brown, the author of "How Shall a Thing Be Called?" This led to her groundbreaking work on basic-level categories. It was previously thought that categories were simply hierarchical, and that lower-level categories were just special (less general) cases of higher categories. Thus, in this general-to-specific hierarchy, sports cars were seen as special cases of cars, which were seen as special cases of vehicles, while rocking chairs were seen as special cases of chairs and chairs as special cases of furniture.

By contrast, Rosch showed that these categories in the middle—cars and chairs—have special properties. They are defined by a confluence of motor programs, mental images, and gestalt perception. They also tend to be learned first and often have the shortest names.

We now have a neural explanation for this confluence of properties. Mirror neuron system research shows common circuitry linking motor programs and gestalt perception, and Martha Farah's (1989) research demonstrates that mental images use the same circuitry as the visual system. That explains why motor programs, gestalt perception, and mental images fit together in defining basic-level categories. There is nothing in set theory that can deal with those phenomena. Most important, Rosch showed that basic-level categorization is embodied. Set theory, of course, is disembodied. The Brown-Rosch research was confirmed in the work of Berkeley linguistic anthropologist Brent Berlin, who showed that the level of the genus in biology has the properties of the basic level.

In short, research on both prototypes and basic-level categories shows that the real capacities of natural language do not have the structure of set theory and go far beyond what classical set theory can do.

Let me be more specific about Rosch's contribution. Her remarkable work revolutionized the empirical study of categorization. She conducted path-breaking experiments on the Dani people of Papua New Guinea in the early 1970s, and performed further experiments at Berkeley. Rosch's New Guinea experiment involved teaching the Dani a series of made-up focal and nonfocal color terms. She found that they remembered focal color terms—which represent more basic colors, such as red or green—far more easily than nonfocal color terms, which represent complex colors, such as red-orange or pink. In later work she explored the conceptual structure of categorization, showing that people more readily identify prototypical cases as members of a category and have a quicker response time to

questions about prototypical versus non-prototypical cases.

In her Berkeley experiments, Rosch asked subjects to rate a series of terms according to how well they exemplify a certain category. For the category of weapon, for example, she derived from the responses a scale that ranked sixty objects in terms of their centrality. Gun ranked at the top. Bayonet, arrow, fists, and words were successively further from the prototype, with dozens of other objects ranked in between. Based on these and similar results for many other categories, Rosch found that respondents recognized a spectrum of similarity, with an ordered sequence of representativeness in relation to the prototype. Using these innovative methods, she established that people categorize a given object or experience by comparing it with the object or experience—the prototype—they think best represents a given category. As I will explain later, these initial discoveries were a key step—though only an initial step—in developing prototype theory.

In sum, Rosch's work is indeed fundamental to the empirical research on which our understanding of categorization rests. In my 1980 book with Johnson, *Metaphors We Live By*, she is a central point of reference for the argument that we do not categorize in set theoretic terms. In my 1987 book, *WFDT*, I survey her work in much more depth to support the argument that set theory does not reflect categorization in natural language, and my book on mathematics (Lakoff and Núñez 2000) likewise underscores these themes.

Stepping back from Rosch's work, we can say that different set theories place distinct constraints on what they can say about any given domain. We find technical subjects for which set theory is useful; certain types of computer databases were developed to fit classical set theory. Some programming languages, such as HTML, required new and very different set theories—developed in part at the International Computer Science Institute (ICSI) at Berkeley. But if the topic of concern is natural language and human conceptual systems, all set theories are going to fail.

Q: Let us now focus on fuzzy sets and fuzzy logic, which are central to these discussions of set theory in social science. Would you give us your views on Lotfi Zadeh and fuzzy-set theory?

A: I would first point out that Zadeh (1965, 1972) initially developed what he called fuzzy-set theory. I added to this theory by introducing my idea of hedges and of different fuzzy logics. Zadeh built on the work on hedges and created what we now think of as his "fuzzy logic."[3] I will use that term to refer to his contribution.

For me, Zadeh is an admirable scholar. The application of his ideas in engineering is remarkable. Zadeh's fuzzy logic was developed into algorithms and chips used in engineering contexts like rice cookers, vacuum cleaners, washing machines, refrigerators, and especially anti-lock brakes (in the brakes of my car). Zadeh and others have developed fuzzy logic control systems, on which there is a large technical literature.[4] Such systems are useful in devices with ongoing *multiple* linear inputs that require smoothly functioning, *single* linear outputs.

The important contribution of fuzzy logic becomes clearer with an example. When someone applies pressure to the brakes in their car, there is an infinite array of values that the amount of pressure can take. Yet the amount of pressure we apply does not vary in continuous gradations, but rather is closer to a step function. These values or steps can be operationalized using hedges—*moderate* pressure, *strong* pressure, and so on—and each hedge can be graphed with what I have called a "Zadeh function."

Fuzzy logic is more useful than, say, linear scales for capturing this process of braking. Because the functions are anchored in hedge terms, they have clearly defined substantive meanings. Using Zadeh's theory, engineers can thus translate the amount of pressure drivers apply to the brakes into functions that can be visualized as the pressure transitions across a spectrum *light* to *moderate* to *strong*.

Fuzzy logic allows engineers to work with increased precision, and it represented an impressive leap for engineering. Zadeh deserves all the acclaim he has received in the engineering world, especially in Japan.

The question for social scientists is whether any real social or political phenomena work like rice cookers or washing machines, and whether fuzzy logic distorts reality and fails in domains that do not work this way.

Q. You and Zadeh had a dialogue over the relationship between fuzzy logic and linguistic hedges. Would you describe that?

A. We started exchanging ideas in the early 1970s. I had previously made an extensive list of linguistic hedges that serve to modify categories. Most of them were complex natural language cases which did not fit Zadeh's fuzzy logic. A small number, however, fit ordered linear scales—for example, *extremely*, *very*, *pretty*, *sort of*, *not very*, *not at all*. Yet these still did not fit Zadeh's original version of fuzzy logic for a simple reason: the original version placed the values for set membership not just on an ordered scale, but on an infinitely-valued, continuous scale between zero and one. That does not correspond to natural language.

Zadeh understood the problem when I described it to him, and he suggested an ingenious solution in his 1972 article, "A Set-Theoretic Interpretation of Linguistic Hedges." Here he developed a version of fuzzy logic that drew on my hedges paper (later published as Lakoff 1973). He defined a group of mathematical functions taking the real numbers from his original fuzzy logic as input. Each was a Gaussian curve peaking at values that approximated ideas like those expressed by linear hedges such as *very*, *sort of*, *not very*, and so on. These curves incorporated the idea of imprecise, fuzzy gradations around each hedge. The output of these functions defined a new kind of fuzzy logic with a small number of linearly ordered values instead of a continuous spectrum of values. In my 1973 paper, I called these "Zadeh functions." Zadeh (1972) called the resulting set-theoretic logic a "hedge logic," a term that continues to be used (van der Waart van Gulik 2009).

[3] See, for example, Zadeh 1995: 271; also Zadeh 1994.

[4] See the bibliography provided in even as ordinary a source as the Wikipedia entry on Fuzzy Control System.

As his new theory of fuzzy logic evolved for engineering applications, Zadeh simplified the curves to linear triangular and trapezoidal functions. In the triangular functions, the peaks of curves are replaced by points, and the curves leading up to the peaks are replaced by the straight sides of the triangle. In the trapezoidal functions, the peaks of the curves are flattened to encompass a range of values, and the curves going up to the peaks are represented by the straight sides of the trapezoid. That makes engineering computations easier.

I should reiterate that the linear hedges used in Zadeh's hedge logic are a minority of the hedges in English. Hedges like *basically*, *essentially*, *regular*, *technically*, *so-called* and many others cannot be handled by fuzzy logic, and Zadeh has never claimed they could be. Moreover, many modifiers are nonlinear and their compositions with nouns cannot be handled by the compositional functions of fuzzy logic. Well-known cases cited in *WFDT* that require frame semantics include *electrical engineer, social scientist*, *mere child*, *fake gun*, *happy coincidence*, *past president*, and many more.

Q: You are suggesting that fuzzy logic does not reflect the structure of meaning in natural language. Would you spell this out?

A: Fuzzy logic does not characterize most of the human conceptual system as it is found in natural language. It cannot characterize frame semantics, conceptual metaphors, conceptual metonymies, conceptual blends, modalities, basic level concepts, radial categories, most hedges, most conceptual composition, and so on. It especially cannot handle the broad range of contested concepts, especially important ones like freedom and democracy that depend on conceptual metaphor, morally-based frames, and radial categories. It cannot account for the experimental results in embodied cognition research.

My 1973 hedges article is sometimes cited as if it were an endorsement of fuzzy logic, but it in fact discusses many limitations. Let me spell out what I said then—and the context was of course my admiration for Zadeh and my collaboration with him.

I noted in the 1973 hedges piece that fuzzy concepts have had a bad press among logicians, and that these concepts merited serious formal study. I tried to suggest how this formal study should be focused.

It is exciting to think back to 1973, when this article was published. What can be called the Berkeley Revolution in Cognitive Science had only begun. Rosch had just started her path-breaking empirical work, and I refer to that in my article. I had not yet developed the idea of radial categories, which later drew together her work and the emerging literature on frame semantics. But elements of these ideas were present.

I identified different types of hedges, and some are amenable to the linear treatment provided by fuzzy logic. Fuzzy logic is linear in the sense that elements are consistently ordered along a line. Many other hedges definitely are not, and the more I developed these ideas, the more I realized that most hedges modify the central category in diverse ways that are definitely non-linear. Zadeh ingeniously identified a few hedges that were very successfully modelled in his engineering applications. I applaud this. But as an overall characterization of natural language, fuzzy logic fails.

Q: Zadeh's 1982 article "A Note on Prototype Theory and Fuzzy Sets" sought to show that fuzzy logic can accommodate the idea of prototypes. Did he succeed?

A: Zadeh's article fails to make the case. First of all, Zadeh only considers the initial version of prototype theory, in which Rosch shows that within categories, there can be a finite hierarchy from examples that are best, good, less good, and so on. Zadeh says this hierarchy shows fuzzy logic is compatible with natural language. Yet even for this initial version of prototype theory, fuzzy logic is inadequate. The initial version is centered on the idea of closeness based on properties related by family resemblances; by contrast, fuzzy logic takes into account neither properties nor family resemblances and is based on a continuous, infinitely-valued linear scale.

Moreover, the fully developed theory of prototypes is more complex than the linear conceptualization suggested by the initial version. It encompasses the use of prototypes to stand for the category as a whole (i.e., metonymy) with respect to some form of reasoning. For instance: (a) Best example prototypes function as defaults where only the category is mentioned. Thus, if you say "There's a bird on the porch," you will most likely have in mind a small songbird, not a duck that could have flown in from a nearby lake, nor an ostrich from an ostrich farm, and definitely not a pelican that might have strayed from the ocean. (b) Typical case prototypes are used for drawing inferences. (c) Reference point prototypes are used to provide a standard in reasoning. (d) Salient example prototypes are used for judging probability. (e) Ideal prototypes are used for making value judgments. Fuzzy set theory does not have any of these properties.

Another element in prototype theory that is not accommodated by fuzzy logic is the idea of radial categories, which capture how cases branch out in many directions from the central members. For example, there are cluster categories defined by a cluster of frames, with modifiers that only pick out one of the frames. The category *mother* is defined by four frames—for birth, genetics, nurturance, and marriage. But stepmother eliminates marriage, and birth mother picks out birth but not marriage, and genetic mother picks out genetics, but not necessarily birth, and so on.

The linear ordering of fuzzy logic certainly does not reflect this pattern. In this and many other ways, by the early 1980s studies of human categorization had left fuzzy logic far behind.

In sum, it is valuable that Zadeh recognized the importance of prototype theory. But he failed to connect it with fuzzy logic.

Q: Do the inadequacies of fuzzy logic for natural language lead to inadequacies for applications to political and social analysis?

A: Yes, definitely. Good examples would be concepts of freedom and democracy. In my 2007 book *Whose Freedom?* I analyzed this contested concept by extending and refining W. B. Gallie's (1956) theory of contested concepts, an outstanding

example of work by a political theorist that captured important facets of human conceptual structure. Inspired by George W. Bush's second inaugural address, where Bush used the words *freedom*, *free*, and *liberty* 49 times in 20 minutes, I undertook to characterize (1) the shared conception of freedom used by both progressives and conservatives, and then (2) the contested extensions of this shared conception, which differ widely between progressives and conservatives. The differences are huge, and the book covers how they apply to a wide range of social and political contexts, from economic markets, to education, religion, foreign policy, human rights, and gender issues.

These vital distinctions for our politics, and the politics of many countries, cannot be approached in any linear fashion in relation to a general concept of freedom, which is what fuzzy logic would require. The same is true of democracy, as Elisabeth Wehling and I pointed out in *The Little Blue Book* (Lakoff and Wehling 2012). Wehling (2013) subsequently—in collaboration with social psychologists—pursued this line of inquiry further using survey and experimental methods. She confirmed the contested conservative versus progressive extensions of the shared core I found in *Moral Politics* (2002) and extended in *The Political Mind* (2008). None of this research fits fuzzy logic.

Q. How would you compare Zadeh's fuzzy logic with Charles Ragin's (2000, 2008) method of scoring fuzzy sets? To avoid confusion, we can refer to these as "fuzzy logic" and "fuzzy-set scoring."

A. They are very different. Zadeh arrayed complex functions on a linear scale to approximate the fuzziness of hedges like *very*, *pretty much*, *sort of*, and *not much*. Each hedge is represented by a complex function. The overall scale is indeed linear, in that the hedges have a well-defined linear order. The input to the functions is the set of real numbers from zero to one. However, the core idea for Zadeh is that the meaning of each specific hedge is fuzzy.

By contrast, with Ragin's method of fuzzy-set scoring, the entire approach is linear. Based on a completely different, non-fuzzy approach, full membership in the overall category is represented by a fixed numerical value, and each hedge also has a fixed numerical value. It is not fuzzy.

Let's set aside for now my argument that most hedges cannot in fact be arrayed on a linear scale. Zadeh's hedge logic is nonetheless a worthy attempt to capture the linear ordering of some fuzzy hedges, and he thereby did something important. We noted the example of pressure: *light pressure*, *moderate pressure*, *strong pressure*, *intense pressure*. In his system, the overall ordering is indeed linear, and the use of fuzzy logic to represent these hedges is interesting, subtle, and valuable in engineering.

By contrast, I am skeptical that Ragin's fuzzy-set scoring can tell us much about the conceptual understanding of the real world that is contained in natural language. I do not have any serious knowledge of the substantive contribution to social science, so I will only comment on the conceptual part.

The examples I have seen of fuzzy-set scoring in social science are, to reiterate, quite different from Zadeh. The subtlety of fuzzy logic is gone. The subtlety of hedge logic, which captures the fuzziness of each step (hedge) on the linear scale, is gone. It seems to have nothing to do with the complex meaning of hedges, or with anything else in natural language.

Instead, in fuzzy-set scoring the analyst constructs a scale with—for example—three, or sometimes five, values. Sometimes there are more values. The values are evenly spaced, fixed numbers that are arrayed linearly: 1.0, 0.5, 0.0; or if there are five values, 1.0, 0.75, 0.5, 0.25, 0.0. This is completely different from what Zadeh did with fuzzy logic. With fuzzy-set scoring, full membership in the category is scored as 1.0, non-membership is scored as 0.0, and the crossover (tipping) point as 0.5. Leading examples of the overall categories analyzed (Ragin 2000, 2008) include *rich countries*, *Protestants*, *major urban areas*, and *developed countries*—obviously important topics for social scientists.

In fuzzy-set scoring, analysts assign values based on their own interpretations, often combined with a mapping from standard linear measures. For example (Ragin 2000:158), they take the measure of GNP per capita as the basis for assigning membership in the category of rich countries. $18,000 to $30,000 per capita is assigned to *clearly rich* (score=1.0), $8,001 to $17,999 to *more or less rich* (0.75), $8,000 is *in between* (0.5), $2,000–$7,999 is *more or less not rich* (0.25), and $100 to $1,999 is *clearly not rich* (0.0). These are hedges, but they are represented with these fixed numerical values, rather than fuzzy functions. Except for the guidance from conventional linear indicators, the principle behind choosing these values is unclear, and I find the discussions of external anchors for this assignment unconvincing. My goal (1973) in analyzing hedges in natural language was to explore their meaning and fuzziness, and Zadeh attempted to capture the idea that their meaning is fuzzy. I don't find that in fuzzy-set scoring.

I have my own misgivings about economic indicators, and I worry about what they hide. But I would prefer to know the GNP per capita of a country, rather than be told that it has a fuzzy-set score of 0.5. This score indicates that it is exactly halfway between being a full member and a full non-member in the set of rich countries, but I don't know what it means conceptually or empirically to be in this set.

In sum, fuzzy-set scoring seems to rely on a rigid, fixed threshold for full membership and for the intermediate values. With Zadeh, these thresholds are, by contrast, fuzzy. In terms of capturing meaning in natural language, with fuzzy-set scoring I don't see the gain over conventional indicators—whatever their limitations. And to reiterate, I do not view fuzzy-set scoring as actually being a fuzzy method.

Q: Do you have concluding comments about these applications of set theory and fuzzy logic?

A: To reiterate, a common justification offered by texts on set theory in social science is that fuzzy set theory and fuzzy logic capture meaning in natural language.

That is simply wrong. It is not supported by the empirical literature on conceptual systems in natural language. Given this mistaken justification, it is hard for me to understand how any social scientist can take set theory and fuzzy logic seriously.

The goal of the fuzzy logic approach seems to be to represent complexity, but for most phenomena, this approach cannot and does not succeed. While fuzzy logic does have impressive applications in engineering, it fails to address the complexity of data routinely examined in the social sciences. Real political and social phenomena do not fit the constraints of fuzzy logic control systems.

One might be tempted to dismiss the application of Zadeh's fuzzy logic to social and political science as misguided. That would be a great mistake and would fail to honor Zadeh's contribution. The important point is that the technical tools of fuzzy logic *define* the data it can fit. The danger is that the technology can distort what should count as real social science data. This danger is certainly also present in many "big data" statistical methods, which define the relevant data as what the technology can do.

The real issues here are empirical: (1) Does the model fit reality? (2) Does it fit the way we conceptualize reality? Personally, I doubt that Ragin's fuzzy-set theory will work in either case. Again, as a cognitive scientist and linguist, I can only judge how set-theoretical tools fit human thought and language. I am not in a position to judge the empirical utility of Ragin's model from other scientific perspectives.

However, I believe that social scientists would do well to look for alterative tools, ones that reflect human conceptual systems and are appropriate to the phenomena that need to be studied.

References

Brown, Roger. 1958. "How Shall a Thing be Called?" *Psychological Review* 65 (1): 14–20.

Farah, Martha. 1989. "The Neural Basis of Mental Imagery." *Trends in Neurosciences* 12 (10): 395–399.

Gallie, W. B. 1956. "Essentially Contested Concepts." *Proceedings of the Aristotelian Society* 56: 167–198.

Lakoff, George. 1973. "Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts." *Journal of Philosophical Logic* 2 (4): 458–508.

Lakoff, George. 1982. *Categories and Cognitive Models*. Cognitive Science Report, no. 2, Institute for Cognitive Studies, University of California, Berkeley.

Lakoff, George. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: University of Chicago Press.

Lakoff, George. 2002. *Moral Politics: How Liberals and Conservatives Think*. 2nd ed. Chicago: University of Chicago Press.

Lakoff, George. 2007. *Whose Freedom? The Battle over America's Most Important Idea*. New York: Picador.

Lakoff, George. 2008. *The Political Mind: A Cognitive Scientist's Guide to Your Brain and Its Politics*. New York: Penguin Books.

Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Lakoff, George and Rafael Núñez. 2000. *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics Into Being*. New York: Basic Books.

Lakoff, George and Elisabeth Wehling. 2012. *The Little Blue Book: The Essential Guide to Thinking and Talking Democratic*. New York: Free Press.

Ragin, Charles C. 2000. *Fuzzy-Set Social Science*. Chicago: University of Chicago Press.

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

Rosch, Eleanor. 1975. "Cognitive Representation of Semantic Categories." *Journal of Experimental Psychology* 104 (3): 192–233.

Rosch, Eleanor. 1977. "Human Categorization." In *Advances in Cross-Cultural Psychology*. Neil Warren, ed. (London: Academic Press), 1–72.

Van der Waart van Gulik, Stephan. 2009. "Adaptive Fuzzy Logics for Contextual Hedge Interpretation." *Journal of Logic, Language and Information* 18 (3): 333–356.

Wehling, Elisabeth E. 2013. "A Nation under Joint Custody: How Conflicting Family Models Divide U.S. Politics." Doctoral Dissertation, Department of Linguistics, University of California, Berkeley.

Zadeh, Lotfi A. 1965. "Fuzzy Sets." *Information and Control* 8 (3): 338–353.

Zadeh, Lotfi. 1972. "A Fuzzy-Set-Theoretic Interpretation of Linguistic Hedges." *Journal of Cybernetics* 2 (3): 4–34.

Zadeh, Lotfi A. 1982. "A Note on Prototype Theory and Fuzzy Sets." *Cognition* 12 (3): 291–297.

Zadeh, Lotfi A. 1994. "Soft Computing and Fuzzy Logic." *Institute of Electrical and Electronics Engineers (IEEE) Software* 11 (6): 48–56.

Zadeh, Lotfi A. 1995. "Discussion: Probability Theory and Fuzzy Logic Are Complementary Rather Than Competitive." *Technometrics* 37 (3): 271–276.

Logic and Set Theory: A Note of Dissent

Giovanni Sartori
Columbia University

> "My underlying complaint is that political scientists eminently lack—a training in logic—indeed in elementary logic." Sartori (1970: 1033)

Logic is an essential foundation for political analysis. It serves to evaluate "the validity of inferences," i.e., the "relationship between premises and conclusion."[1] In numerous publications, I contend that logic is indispensable for good research. However, I also advise caution in choosing tools for political research, arguing in favor of logic as a broad foundation for *methods*, and against excessive reliance on narrow *techniques* (1970: 1033).

I must therefore dissent from Goertz and Mahoney's (2012) *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. A central claim in this noteworthy book is that qualitative research is and should be based on set theory. In particular, they advocate techniques derived from set theory as the basis for qualitative work. They equate logic

[1] The definition in the Glossary of my *Social Science Concepts* (Sartori 1984: 78) is as follows. Logic is "the study of the validity of inferences (see: Validity). Thus logic deals with the relationship between premises and conclusion, not with the truth of the premises." Vulgarly: logic applies to the form, not to the substance of arguments. Validity (1984: 85) is defined as follows: "In logic an argument is valid when its conclusion correctly follows (inferentially) from its premise. A measurement is valid (empirically) if it measures what it purports to measure."

and set theory,[2] evoking my commitment to logic as an apparent endorsement of their approach.[3] Yet I do not endorse it.

To frame my argument, a key point of agreement should be noted. I have long recommended a semantic approach to concepts, which they adopt.

However, the book's advocacy of set theory as the basis of qualitative research takes us in the wrong direction. They endorse fuzzy-set *techniques* that are far too confining. It is indeed essential to push ourselves—as fuzzy sets do, to ask the basic, logical questions: What is an instance of a concept? What *is not* an instance? Yet the intricate fuzzy set procedures cantilever out from these questions, posing dangers of technique that concern me. In some domains of social science we now see growing skepticism about complex statistical techniques—and a turn to simpler tools. The elaborate procedures of fuzzy sets merit the same skepticism.

In applying logic I strive for parsimony, combined with adequacy to the task at hand. Consider my "ladder of abstraction," which organizes concepts to address the traveling problem in comparative research—the challenge of achieving conceptual traveling without conceptual stretching (1970: passim). Narrower concepts lower down the ladder are indeed subsets of broader concepts further up. However, as I formulated the ladder I kept the argument as simple as possible. I relied on Cohen and Nagel's (1936: 33) classic text on logic, noting their idea of inverse variation.[4] This pattern captured precisely the framing I wanted—no more, and no less. This simple formulation stands at a great distance from Goertz and Mahoney's elaborate techniques of set theory.

Hence, I must dissent from their recommendation to apply set theory as a central technique in qualitative research.

References

Cohen, M. R. and Ernst Nagel. 1936. *An Introduction to Logic and Scientific Method*. London: Routledge and Kegan Paul.

Goertz, Gary and James Mahoney. 2012. *A Tale of Two Cultures: Qualitative and Quantitative Research in the Social Sciences*. Princeton: Princeton University Press.

Sartori, Giovanni. 1970. "Concept Misformation in Comparative Politics." *American Political Science Review* 64 (4): 1033–1053.

Sartori, Giovanni. 1984. "Guidelines for Concept Analysis." In *Social Science Concepts: A Systematic Analysis*. Giovanni Sartori, ed. Beverly Hills, CA: Sage Publications.

[2] Goertz and Mahoney (2012: 16, n.1).

[3] My epigraph (above) about training in logic also serves as the epigraph that leads their Chapter 2, which introduces their argument about logic and set theory. This chapter is entitled "Mathematical Prelude: A Selective Introduction to Logic and Set Theory for Social Scientists."

[4] Sartori (1984: 68, n.40).

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.

Rihoux, Benoît and Heike Grimm, eds. 2010. *Innovative Comparative Methods for Policy Analysis: Beyond the Quantitative-Qualitative Divide*. New York: Springer.

Rihoux, Benoît, Ilona Rezsöhazy, and Damien Bol. 2011. "Qualitative Comparative Analysis (QCA) in Public Policy Analysis: An Extensive Review." *German Policy Studies* 7 (3): 9-82

Rothstein, Richard. 2004. *Class and Schools: Using Social, Economic, and Educational Reform to Close the Black-White Achievement Gap*. Washington: Economic Policy Institute.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes." *Journal of the American Statistical Association* 100 (469): 322–331.

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

Sen, Bisaka. 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, Jon, Jonathan Wistow, and Clare Bambra. 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, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. Cambridge: MIT Press.

Part 2. Where Do We Go from Here?

A Larger-N, Fewer Variables Problem? The Counterintuitive Sensitivity of QCA

Chris Krogslund
University of California, Berkeley
ckrogslund@berkeley.edu

Katherine Michel
University of California, Berkeley
katherine_michel@berkeley.edu

> "…let us turn to a discussion of specific ways and means of minimizing the 'many variables, small-N' problem of the comparative method."
>
> Arend Lijphart (1971: 686)

Studies employing Qualitative Comparative Analysis (QCA) often analyze a relatively small number of cases to assess the impact of a substantial number of variables on a given outcome. As emphasized in the quotation above, in the tradition of writing on the comparative method and multi-method research, this ratio of cases-to-variables is viewed as an analytic problem. In this exploratory research note, we raise questions about the implications of this ratio for the stability of findings in QCA. One common method of assessing result stability is the "drop-one" sensitivity test, which repeatedly reruns a particular analysis, each time dropping a single case. We find that, for the number of cases (*n*) to which analysts most routinely apply QCA, this type of sensitivity analysis produces paradoxical results.

We refer to this cases-to-variables relationship as the *n/k* ratio, where *n* is the number of cases and *k* is the number of explanatory variables.[1] According to standard expectations, more robust findings emerge with a *higher n/k* ratio.

Directly contrary to this standard expectation, we encounter the paradoxical result that in drop-one sensitivity tests, QCA findings based on a *lower n/k* ratio prove to be more stable. This result calls into question the validity of the drop-one test as a sensitivity metric for QCA and forces us to consider why QCA results behave in such a manner. This research note explores these issues.

Before the discussion proceeds, we must offer two caveats concerning simulations and case knowledge. Regarding simulations, it is crucial for the credibility of QCA that researchers test the method's reliability and robustness, and one way to do this is with simulations. Further, the present discussion of potential problems with the drop-one sensitivity test should not be taken as reflecting skepticism about the overall value of sensitivity tests as a means of evaluating QCA. Sensitivity tests reveal major problems with the method—the point of concern here is simply to identify the most appropriate tests. Our goal is to carefully adapt simulations to appropriately evaluate the drop-one test as a sensitivity metric for QCA.

Second, a QCA scholar might argue that our counterintuitive finding about the *n/k* ratio is in fact not surprising, given that the method relies heavily on the close knowledge of relatively few cases for making inferences. Such a scholar will view a small *n* as an advantage, not a disadvantage. We return to this issue in the conclusion, and offer just one comment here. Looking over many articles based on QCA, we see little evidence that close knowledge of cases is crucial to the method. Instead, findings appear to be strongly driven by the application of QCA's basic algorithms. Hence, though case knowledge is crucial in the original design of the method, it is not clear that it is crucial in practice. We therefore conclude that a distinctive role of case knowledge does not explain our paradoxical finding about the *n/k* ratio.

Sensitivity Analysis and the *n/k* Problem

Sensitivity analysis is a fundamental research tool in social science methodology. In the domain of conventional quantita-

The authors thank Alisan Varney for her excellent and timely research assistance. We also benefited from valuable comments from David Collier, Jack Paine, and Sean Tanner.

[1] QCA scholars use the term "condition" to refer to both single explanatory variables (e.g., variables "A," "B," "C," and "D") and combinations of explanatory variables (say, "AB" and "cD"). We utilize the term "variable" when referring to single explanatory factors. We consider a "condition" to be a combination of causal "variables." A "solution," alternatively, is a combination of "conditions." A solution refers to a QCA result of the form AB+cD, meaning the outcome occurs in the presence of causal variables A and B (condition one), or the absence of variable C and the presence of variable D (condition two). As a general rule, capitalized conditions indicate presence, while lowercase conditions indicate absence.

tive analysis, there is a proliferation of methods for evaluating the sensitivity of findings, including a vast amount of work utilizing simulations to assess the robustness of results to case selection and measurement error, among other factors.[2] This work on sensitivity analysis, developed and validated by statisticians and social scientists alike, spans several decades, and has typically been applied to quantitative inferential tools.

Scholars are only now beginning to develop techniques for sensitivity analysis of QCA.[3] Just as in quantitative work, developers of sensitivity tests for QCA must make choices about which aspect of the research design is of greatest concern for the stability of findings. For example, the drop-one test assesses the stability of results through the iterated elimination of specific cases. Other types of tests focus on the effects of measurement error in the independent or dependent variables, on choices in setting calibration parameters, and on model specification.

Given that researchers frequently utilize QCA in situations with a serious *n/k* problem, it is essential to establish whether the drop-one test is an appropriate form of sensitivity analysis. In a first step toward this end, we can attempt to identify what it is about higher *n/k* ratios that might threaten the usefulness of sensitivity tests.

As applied to QCA, we can break down the *n/k* problem into two separate issues. The first is the number of cases per "causal path," or the number of cases that share a given combination of explanatory variables that lead to an outcome. In the QCA "truth table," all potential causal paths are represented as the rows in the table.

When a QCA study reports the alternative causal paths to the outcome, the number of cases per path is often small—sometimes only one or two cases—which raises the question of whether identification of causal paths with few cases is reliable. Scholars such as Schneider and Wagemann (2012: 285–295) underscore the importance of carefully considering the degree to which sensitivity analysis reveals instability in the causal paths. If, for example, a causal path corresponds to only one case and the sensitivity analysis drops that case, the number of observed causal paths will change; how frequently such changes occur should influence our confidence in a QCA study's conclusions.

The second issue is that, depending on the total number of causal variables considered, adding or removing a single variable will differentially change the percentage of potential causal paths that are unobserved, or "empty." This is because the number of potential causal paths grows exponentially with the number of variables. This non-constant increase in the number of empty paths reflects what QCA scholars call "limited diversity," meaning the existence of potential causal paths with no corresponding empirical cases. Such empty paths are termed "logical remainders."[4]

Consider the following two scenarios.

Scenario One: From *k=3* to *k=4*. A *k* of three yields eight (2^3) potential causal paths. Hence, if *n* is equal to seven, at a minimum, one potential causal path will remain empty (8–7=1). If we hold n constant and increase k (to reiterate, the number of explanatory variables) to four, the number of potential causal paths increases to 16 (2^4). This means that, at a minimum, nine potential causal paths will inevitably remain empty (16–7=9). This jump from one to nine, with the addition of a single explanatory variable, produces a minimum of eight additional causal paths that will remain empty. Another way to think of this is in terms of percentages. With *n* equal to seven and *k* increasing from three to four, the minimum percentage of empty causal paths jumps from 13 to 56 percent.

Scenario Two: From k=4 to k=5. If we again hold *n* constant at seven and increase *k* to five, the number of potential causal paths now increases to 32 (2^5), with a minimum of 25 empty potential causal paths (32–7=25). Here, the addition of one explanatory variable produces a minimum of 16 additional empty paths. With *n* equal to seven and *k* increasing from four to five, the minimum percentage of empty causal paths jumps from 56 to 78 percent.

The key things to notice from these two scenarios are, first, that the percentage of empty causal paths becomes large quickly, with a relatively small number of explanatory variables; and, second, that the jump in the minimum percentage of empty causal paths in scenario two is half as large as that in scenario one. This pattern holds if we subsequently increase the number of explanatory variables from five to six, from six to seven, and so on. Because the *n/k* ratio directly reflects this exponential nature of the limited diversity problem, how sensitive results are to adding or subtracting variables (while holding *n* constant) should influence our confidence in a QCA study's results.

Below, we demonstrate that these two issues are of great consequence for whether the drop-one test should be applied to QCA. We find that this type of sensitivity test fails to capture the crucial problem of concern here: For a standard range of *n*, QCA results can appear *more* robust when the *n/k* problem worsens. Put another way, if we increase the number of explanatory variables relative to the number of cases—a move that typically weakens result validity—QCA results appear relatively more robust.

We first illustrate this counterintuitive finding by comparing two examples of QCA studies. We then analyze a larger set of 52 examples, showing that QCA studies tend to focus on a range of *n*—roughly between five and 35—within which increasing the n/k ratio will, paradoxically, heighten the sensitiv-

[2] For overviews of sensitivity analysis see, among many others, Gelman et al. (2004), Gelman and Hill (2007), Morgan and Winship (2007), and Rosenbaum (2002).

[3] See, for example, Schneider and Wagemann (2012), Hug (2013), Krogslund et al. (2014), and Thiem (2013).

[4] See, for example, Ragin (2000; 2008), Ragin and Sonnett (2004), Rihoux (2006), and Rihoux and Ragin (2009). There are multiple types of the limited diversity problem. Schneider and Wagemann (2012: 153–157), building on their earlier work (2006; 2010), identify three: (1) arithmetic remainders, wherein the number of rows is greater than the number of cases, (2) clustered remainders, wherein some causal paths do not exist in social reality, and (3) impossible remainders, wherein some causal paths can never exist.

Table 1: Summary of Krook (2010) and Kim and Lee (2008)

	Cases (n)	Variables (k)	n/k ratio	Average cases per causal path	Potential causal paths (2^k)	Paths with at least one case
Krook (2010)	22	5	4.4	2.6	32	14
Kim and Lee (2008)	16	6	2.7	1.6	64	14

ity of the results. Finally, we conclude by highlighting untapped research areas that should be central components of further investigation.

An Initial Illustration

Consider the following two studies, both of which use crisp-set QCA. We briefly describe each study and then compare how they perform in drop-one sensitivity tests.

Krook (2010) seeks to explain cross-national differences in the percentage of female members of parliament. She analyzes 22 cases, with binary observations on the dependent variable (scored one if the percentage of female members of parliament is above 30 percent) and five causal variables.[5] The n/k ratio is 22/5, or 4.4. After performing the QCA minimization, she identifies five causal paths that lead to a value of one on her dependent variable.[6] These paths all contain either two or three cases, with an average of 2.6. For the scholar concerned with noise in the data or a potential random element in the causal process, paths with few cases might especially raise concerns. Prior to minimization, of the 32 potential causal paths, she empirically observes just 14, meaning that the observed paths represent 44 percent of the potential causal paths to the outcome.

Kim and Lee (2008) seek to account for variations in types of welfare state policies regarding pensions and employment. Their dataset of 16 countries incorporates six causal variables, yielding a n/k ratio of 2.7.[7] They analyze four dependent variables that reflect different pension and employment security policies and run the analysis separately for each outcome. In these four iterations, the average number of cases per path is 1.6. The observed data corresponds to only 14 of the 64 potential causal paths to the outcome, or 22 percent.

Were one to guess which of these QCA articles produces more stable results, an informed choice might be Krook's study of women's representation, due to its higher n/k ratio. Under the best of circumstances, teasing out the connection between any given causal variable and the outcome using cross-case comparison requires at least one case per variable.[8] As the n/k ratio decreases, one would expect that the omission of a case increases the probability of a significant departure from the original result. This is due to the fact that, as the n/k ratio decreases, the probability increases that a score on any given case may influence the outcome.

Surprisingly, this expectation is incorrect: With a drop-one sensitivity test, the results for the welfare state study of Kim and Lee, with a n/k ratio of 2.7, are less sensitive than those for Krook's study of women's representation, with a n/k ratio of 4.4. The number of Kim and Lee's solutions expands from an original finding of one to an average of 5.5 (across four dependent variables), while the average number of conditions—the combinations of explanatory variables that make up the solutions—grows from 2.5 to 4.5. By contrast, under the same drop-one test, the number of Krook's solutions grows dramatically from an original finding of one to seven, while the number of conditions across these solutions increases from five to ten.

To summarize, in this particular QCA comparison, a common sensitivity metric counterintuitively rates the study with a greater n/k problem as yielding more robust findings.

Sensitivity Analysis and QCA: Using Simulations to Explore the n/k Problem

On closer inspection, this counterintuitive finding turns out to

[5] Specifically, the dependent variable is the percentage of women in the lower house of the national parliament, and the independent variables are indicators for (a) proportional electoral system, (b) gender quota, (c) social democratic welfare state, (d) autonomy of the women's movement, and (e) strong left parties. The country set includes Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom, and United States. Krook also conducts her analysis on 26 sub-Saharan African countries (with one altered independent variable), but for illustrative purposes, we focus here on her analysis of developed countries.

[6] Note that these results refer to her conclusions without logical remainders incorporated.

[7] Specifically, the independent variables are binary indicators for (a) high per capita GDP, (b) high age-dependency ratio, (c) high pension maturity level, (d) strong trade unions, (e) decentralized constitutional structure, and (f) high decommodification. The country set includes Australia, Belgium, Canada, Denmark, Finland, France, Germany, Ireland, Japan, Netherlands, New Zealand, Norway, Sweden, Switzerland, United Kingdom, and United States.

[8] By "best-case scenario," we mean that the observations are perfectly orthogonal, there is no multiple or conjunctural causation, and there is no measurement or model error.

[9] Note that, while multi-value and fuzzy set QCA can technically have an infinite number of paths due to their use of certain fuzzy set score calibration and logical reduction parameters, these variants still ultimately require dichotomization for inference.

be unremarkable. To reiterate the finding, for a certain range of n—and, indeed, the most common n found in QCA work—studies with a more severe *n/k* problem will actually appear more robust in the drop-one sensitivity test than studies with a less problematic *n/k* ratio.

To see why this is the case, it is important to first remember what drives changes in QCA results: the proportion of potential causal variable combinations that have at least one case. In contrast to regression, for instance, QCA and other set-theoretic methods have a discrete number of possible solutions, which is a function of the total number of causal variables evaluated.[9] This is the source of the familiar 2^k number of potential paths for *k* causal variables.

As with any sensitivity analysis, the greater the percentage of total potential paths that we expect will be left empty with the removal of a case, the more sensitive QCA results will be to dropping cases. But because the number of total potential paths is a nonlinear function of the number of causal variables, the relationship between the *n/k* ratio and the sensitivity of QCA results is complex.

Consider Figure 1, which uses simulations to show the relationship between the number of cases (*n*), the number of explanatory variables (*k*), and the expected percentage—out of the total potential paths—that are observed, i.e., that have at least one case. Each of the curved lines ranging in shading from black to grey represents a different *k*—and, therefore, a different number of potential causal paths. For each of these curves with a given *k*, the vertical axis gives the expected proportion of total potential paths observed with *n* cases. We calculated these probabilities at each point by (1) creating 10,000 randomly generated datasets of size *n*, (2) observing, out of the total number of potential paths for a given *k*, how many paths actually appeared in the simulations, and (3) calculating the average percentage of potential paths with at least one case observed across all draws. Note that the dashed grey lines running from roughly top-left to bottom-right in the figure connect points that have identical *n/k* ratios. As these dashed lines move toward the top right of the figure, the *n/k* ratio is increasing.

In order to see in Figure 1 how the typical *n/k* ratio prescription of "higher is better" breaks down in QCA, consider a situation in which n=10 and k=3. The slope of the curve corresponding to this situation is steeper than the slope of the curve representing n=10 and k=6. Adding or removing a case when n=10 and k=3 will therefore, in expectation, change the total number of potential paths with at least one case more than if n=10 and k=6. This means that QCA results with n=10 and k=3 are likely to be more sensitive than QCA results with n=10 and k=6.

This flies in the face of standard expectations regarding the *n/k* ratio. Paradoxically, increasing the number of explanatory variables in a QCA model with a small *n* can produce results that appear to be relatively *more* stable. Put another way, making the traditional small-*n*, many variables problem more severe can, in fact, yield more stable results—at least as measured by a common sensitivity test. This is why we point instead to the "larger-*n*, fewer variables problem" in the title of this research note.

We show this same result in Table 2, which builds on Figure 1. The cell representing the intersection of each row and column contains an inequality relationship between "A" and "B." This indicates which of the two (*n*, *k*) situations produces more stable results according to the drop-one sensitivity test: either the configuration of n and k found in the row ("A") or the configuration in the column ("B"). We highlight the result that runs directly contrary to the standard idea of the superiority of a higher *n/k* ratio by marking it with an asterisk, bolding, and underlining it.

Sensitivity Analysis and QCA: Using 52 Examples to Explore the *n/k* Problem

To what extent are these simulation findings relevant to everyday applications of QCA? To answer this question, we turn to the Comparative Methods for Systematic Cross-Case Analysis (COMPASSS) website, an outstanding repository for published articles using QCA.[10] For some of these articles, the repository additionally includes the dataset.

We collected all available, fully-calibrated datasets hosted on the COMPASSS server, covering applications of csQCA, mvQCA, and fsQCA.[11] We did not include uncalibrated datasets, as the calibration process itself is time-intensive, idiosyncratic, and often poorly documented. For studies with multiple dependent variables, we split the original dataset into multiple datasets, with each including only one of the dependent variables.[12] This process left us with 52 datasets.

Table 3 presents descriptive statistics for these 52 QCA datasets. Roughly one-half use csQCA, while the remaining studies are evenly split between mvQCA and fsQCA. The median number of cases (*n*) is roughly 15, the median number of variables (*k*) is five, and there is an average of 0.5 cases per causal path.

Figures 2 and 3 overlay the distribution of QCA datasets on a figure similar to Figure 1. However, the new figures include a larger range of explanatory variables and show the expected sensitivity of QCA results. Across all figures, the sensitivity metric is the percentage of extra solutions produced by sequentially dropping each case in a given dataset. Figure 2 displays the two-dimensional density of QCA studies, whereas Figure 3 differentiates the distribution according to the cs, mv, and fs versions of QCA, as indicated by the different shapes.

The density plots in Figures 2 and 3 confirm what we suspected on the basis of the descriptive statistics in Table 3: The vast majority of QCA studies fall in the *n/k* ratio range where the drop-one sensitivity test yields the paradoxical result of low *n/k* studies as more robust than higher *n/k* ratio studies. For the most commonly used *n* in QCA studies, re-

[10] We scraped the COMPASSS website on January 7, 2014 (http://www.compasss.org).

[11] Note that "cs" stands for crisp set (binary) membership scores, "mv" stands for multi-value (non-binary) memberships scores, and "fs" stands for fuzzy set (non-binary, bounded between zero and one) membership scores.

[12] When we apply QCA, we include all the explanatory variables contained in each dataset in the causal model.

Figure 1: Simulations of relationship between the *n*, the *k*, and the expected percent of paths with at least one case

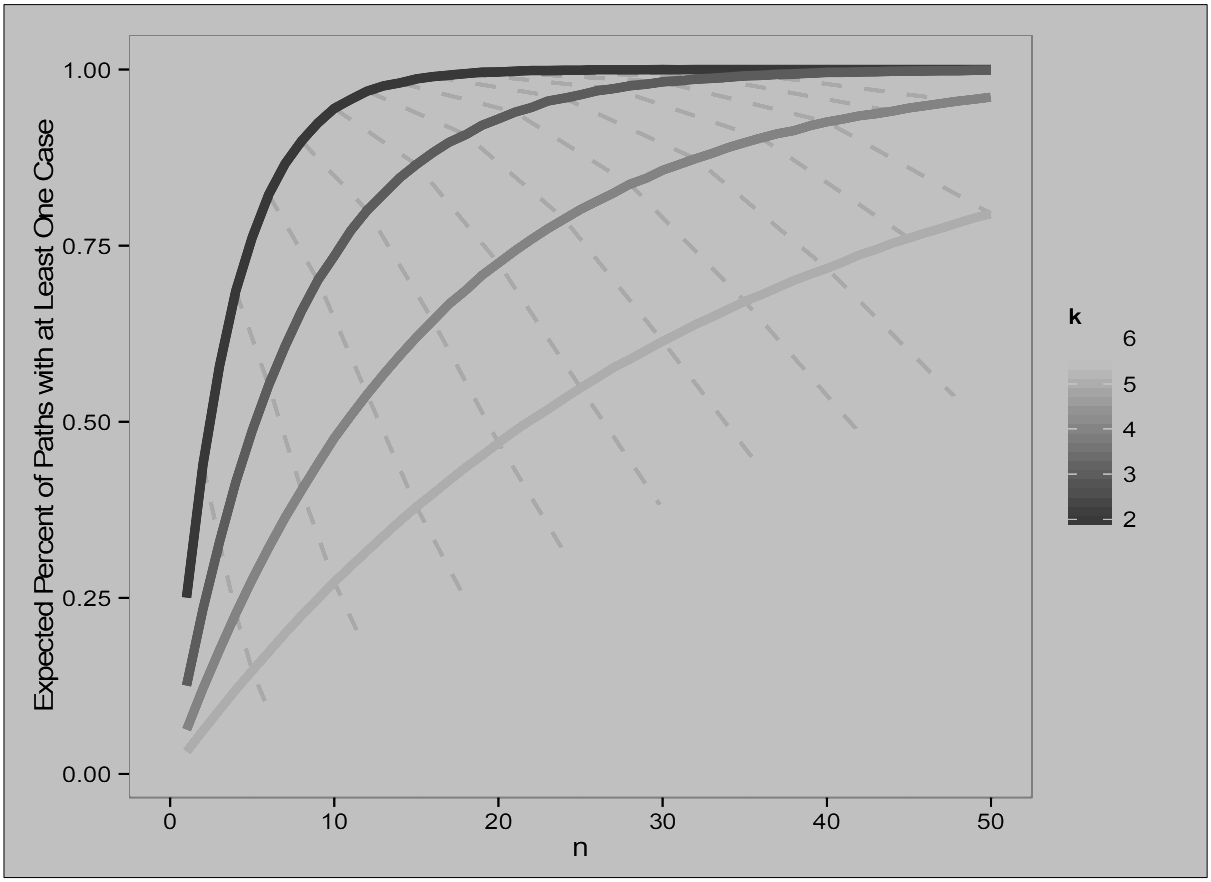


Table 2: Stability of Simulation Results: Paired Comparisons of Alternative *n/k* Ratios

		B			
		Low *n*, Low *k* *n/k* = 1	High *n*, Low *k* *n/k* > 1	Low *n*, High *k* *n/k* < 1	High *n*, High *k* *n/k* = 1
A	Low *n*, Low *k* *n/k* = 1	-	B > A	**B > A***	B > A
	High *n*, Low *k* *n/k* > 1	-	-	A > B	A > B
	Low *n,* High *k* *n/k* < 1	-	-	-	B > A
	High *n*, High *k* *n/k* = 1	-	-	-	-

Notes:
1. The lower-left triangle of cells in this table simply mirrors the upper-right triangle. We therefore do not fill in the cells in the lower-left triangle.
2. The inequality in each cell reflects which configuration is more stable for a given pairing.
3. The cell that is contrary to standard expectations about the stability of alternative *n/k* ratios are underlined in bold with an asterisk.

Table 3: Descriptive Statistics for 52 COMPASSS Datasets

	csQCA	fsQCA	mvQCA	All
Count	27	13	12	52
Median *n*	16	15	18	15.5
Median *k*	6	5	5	5
Median 2^k	64	32	32	32
Median *n/k*	2.67	3.00	3.05	2.71
Median $n/2^k$	0.25	0.47	0.63	0.47

Figure 2: Expected sensitivity of results for 52 QCA datasets (Relationship between the *n*, the *k*, and the expected percent of paths with at least one case)

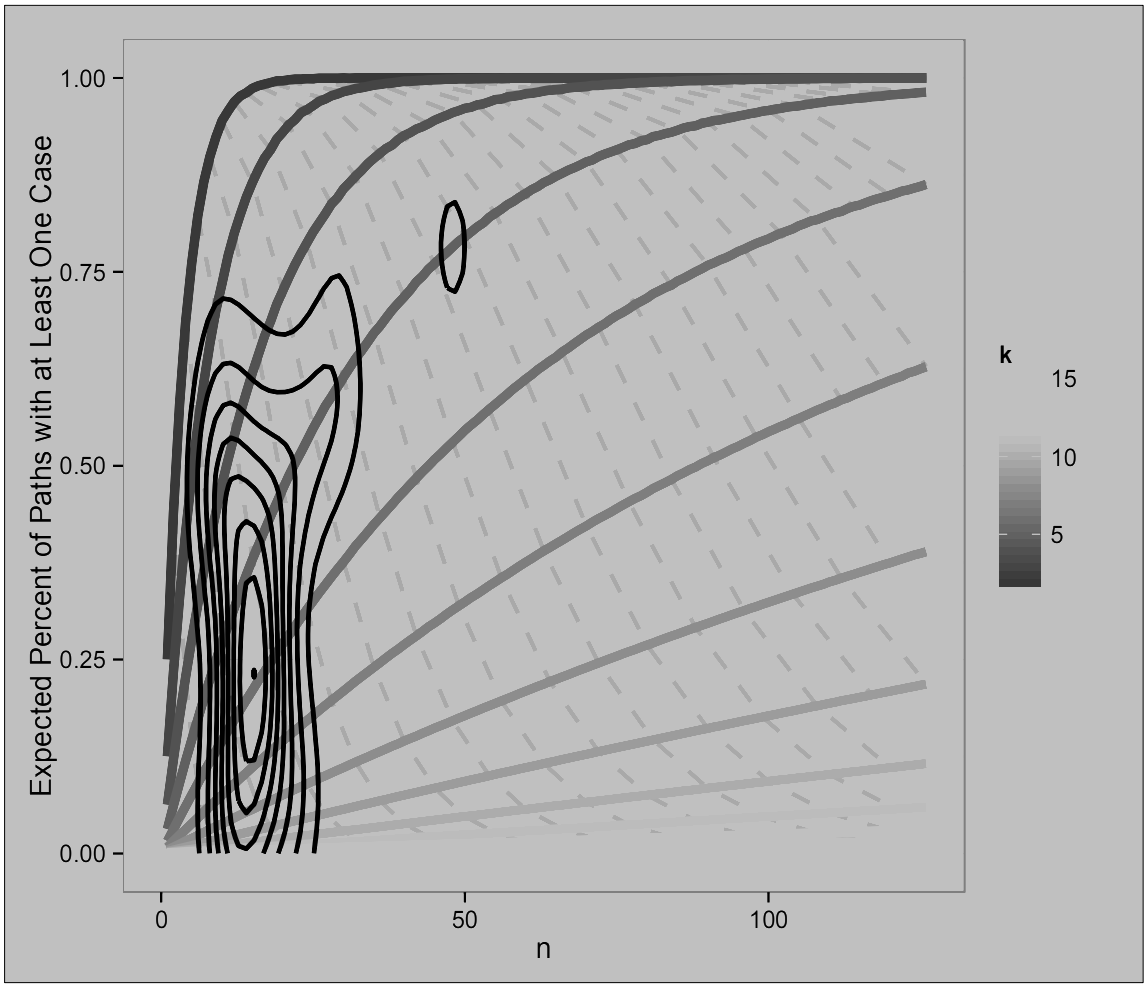
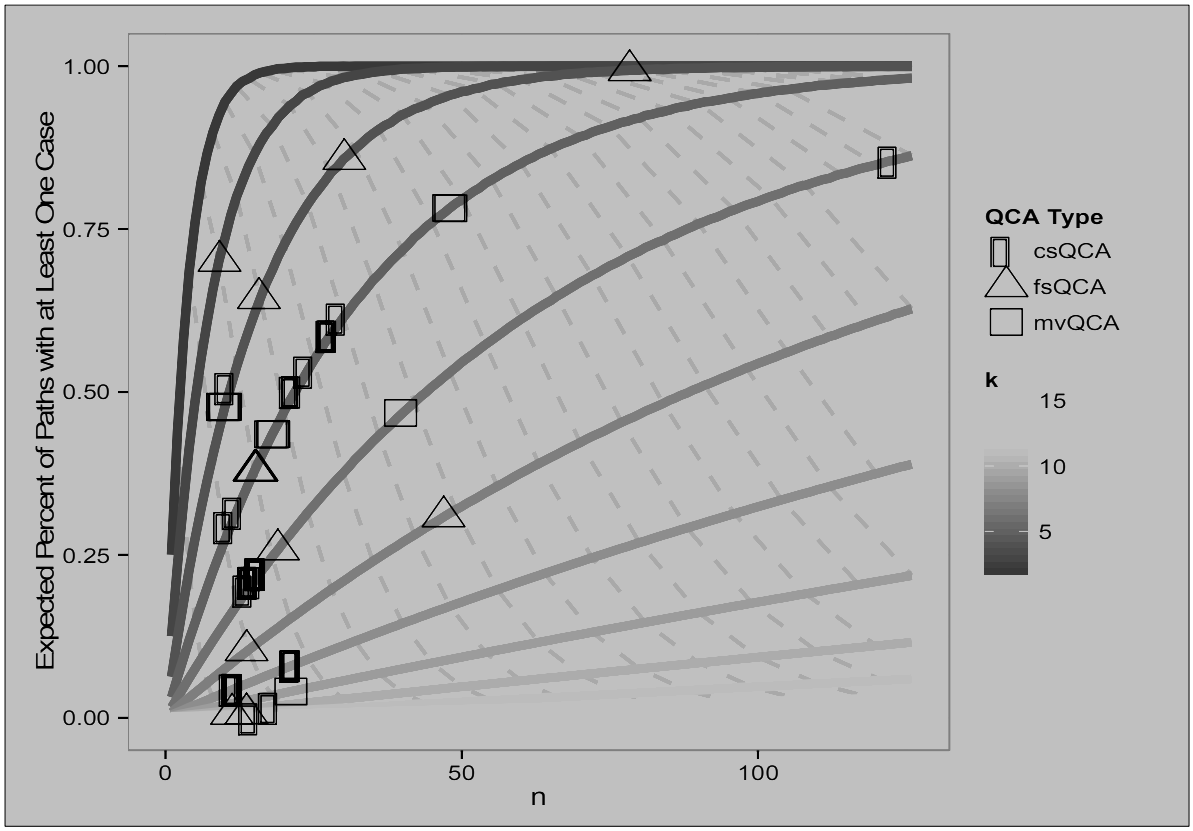


Figure 3: Expected sensitivity of results for 52 QCA datasets, by QCA type (Relationship between the *n*, the *k*, and the expected percent of paths with at least one case)



searchers could actually mask instability with complexity, by simply adding causal variables.

We confirm this finding by running the drop-one sensitivity test on each of the 52 datasets. Figure 4 again overlays the positions of each study with respect to the *n* and the *k*, with the size of each point corresponding to the sensitivity of the findings. As expected, many of the studies we would typically consider to have more problematic *n/k* ratios in fact appear more robust, as compared to those with less problematic *n/k* ratios. Thus, the paradoxical finding that motivated this analysis emerges again with the data from these 52 studies.

Conclusion: Moving Forward

To summarize, our results suggest that a canonical tool for sensitivity analysis employed in social science may not, in practice, reliably assess the robustness of QCA findings. When we apply the drop-one sensitivity test to QCA results, a low *n/k* ratio no longer yields greater instability. For the small *n* common in applied QCA work, datasets with relatively more explanatory variables appear to produce more stable results than datasets with fewer such variables. This stands contrary to one of the basic tenets of cross-case causal inference. For QCA, it appears that results become more unstable with "*many* cases, *few* variables," rather than the reverse. This finding stands Lijphart's famous dictum on its head.

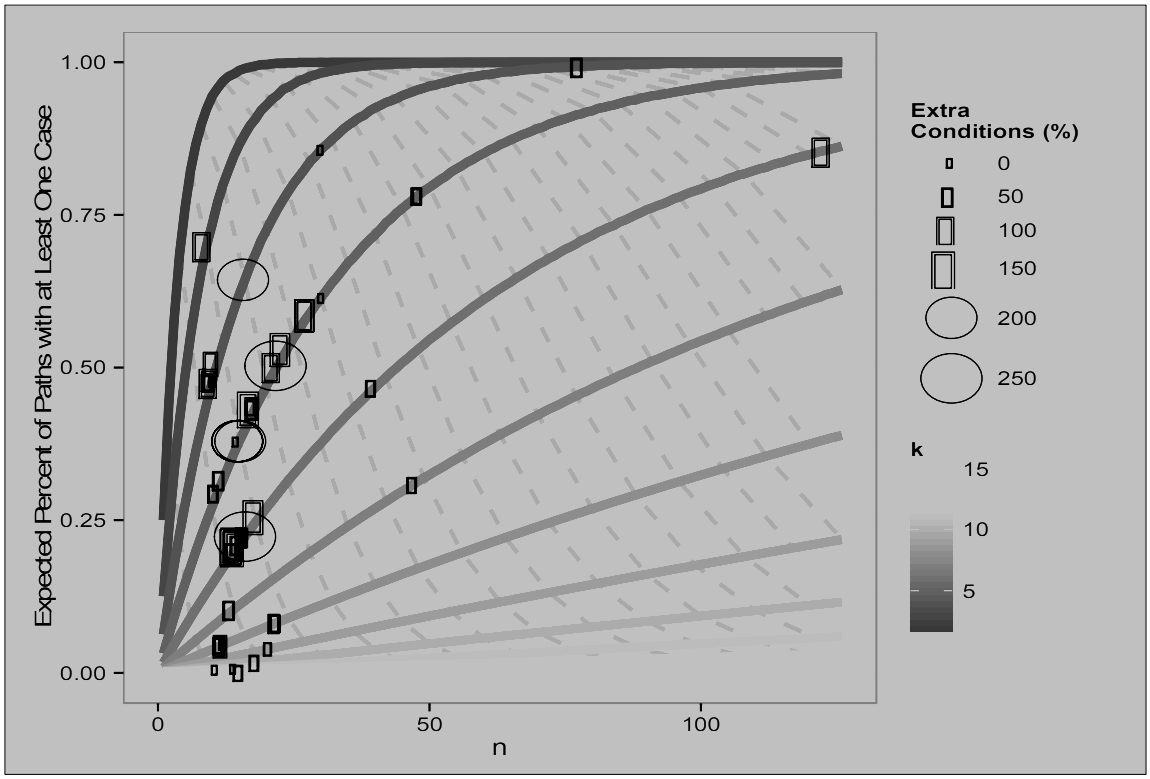
To reiterate a key point from the introduction, this concern about the drop-one test does not reflect skepticism about the contribution of sensitivity tests to evaluating QCA. Sensitivity tests show that the method has major vulnerabilities. The concern here is simply to identify the most appropriate tests.

The counterintuitive finding presented in this exploratory research note leads us to conclude by identifying three untapped research areas that we believe should be central in the future.

1. Distinctive focus of QCA. We noted in the introduction that QCA practitioners may take a very different view of these *n/k* issues. They may consider the close examination of a small number of cases, in conjunction with many variables, as distinctively well-suited for the analysis of multiple and conjunctural causation. We noted above that many existing QCA articles show little evidence that close knowledge of cases plays a strong role. Nonetheless, we must take this argument seriously and consider (a) the contributions that might be made by case knowledge—for example, reducing measurement error, improving model specification, and providing an alternative basis for inference; and (b) how and in what ways such gains from case knowledge might be reflected in sensitivity tests.

2. *Alternative simulation tests*. Might other, QCA-specific sensitivity tests return results consistent with the standard no-

Figure 4: Sensitivity of results for 52 QCA datasets according to the drop-one sensitivity test (Relationship between the *n*, the *k*, and the expected percent of paths with at least one case)



tion that a higher *n/k* ratio indicates more robustness, not less? A key goal moving forward must be to create a QCA-specific sensitivity test that adequately incorporates the complex relationship between the number of cases and the number of explanatory variables.

3. *Choices for dealing with logical remainders*. Relating directly to the problem of limited diversity, might user choices about dealing with empty paths have consequences for robustness? Before running a QCA analysis, users are able to specify how the algorithm will deal with logical remainders. The algorithm will produce one of three solution types: (a) a "complex solution," the default option, which does not incorporate logical remainders in the minimization process; (b) a "parsimonious solution," which incorporates logical remainders; or (c) an "intermediate solution," which incorporates logical remainders, but filters them according to the analyst's directional expectations (Thiem and Dusa 2013). How might this user choice affect result stability?

These three issues clearly merit further attention in ongoing research on the "larger-*n*, fewer variables problem" in QCA.

References

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis*, 2nd ed. Boca Raton: Chapman and Hall/CRC.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Hug, Simon. 2013. "Qualitative Comparative Analysis: How Inductive Use and Measurement Error Lead to Problematic Inference." *Political Analysis* 21 (2): 252–265.

Kim, Kyo-Seong and Yeonjung Lee. 2008. "A Qualitative Comparative Analysis of Strategies for an Ageing Society, with Special Reference to Pension and Employment Policies." *International Journal of Social Welfare* 17 (3): 225–235.

Krogslund, Chris, Donghyun Danny Choi, and Mathias Poertner. 2014. "Fuzzy Sets on Shaky Ground: Parametric and Specification Sensitivity in fsQCA." Revised version of a paper presented at the 2013 Annual Meeting of the American Political Science Association, Chicago.

Krogslund, Chris and Katherine E. Michel. 2014. "Testing the Ability of Set-theoretic Methods to Recover Data Generating Processes." Presented at the Annual Meeting of the Midwest Political Science Association, Chicago.

Krook, Mona Lena. 2010. "Women's Representation in Parliament: A Qualitative Comparative Analysis." *Political Studies* 58 (5): 886–908.

Lijphart, Arend. 1971. "Comparative Politics and the Comparative Method." *American Political Science Review* 65 (3): 682–693.

Morgan, Stephen L. and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

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

Ragin, Charles C. 2000. *Fuzzy Set Social Science*. Chicago: University of Chicago Press.

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

Ragin, Charles C. and John Sonnett. 2004. "Between Complexity and

Parsimony: Limited Diversity, Counterfactual Cases and Comparative Analysis." In *Vergleichen in der Politikwissenschaft*, eds. Sabine Kropp and Michael Minkenberg. Wiesbaden: Verlag für Sozialwissenschaften, 180–197.

Rihoux, Benoît. 2006. "Qualitative Comparative Analysis (QCA) and Related Systematic Comparative Methods: Recent Advances and Remaining Challenges for Social Science Research." *International Sociology* 21 (5): 679–706.

Rihoux, Benoît and Charles C. Ragin. 2009. *Configurational Comparative Methods*. Thousand Oaks: Sage Publications.

Rosenbaum, Paul R. 2002. *Observational Studies*, 2nd edn. New York: Springer-Verlag.

Schneider, Carsten Q. and Claudius Wagemann. 2006. "Reducing Complexity in Qualitative Comparative Analysis (QCA): Remote and Proximate Factors and the Consolidation of Democracy." *European Journal of Political Research* 45 (5): 751–786.

Schneider, Carsten Q. and Claudius Wagemann. 2010. "Standards of Good Practice in Qualitative Comparative Analysis (QCA) and Fuzzy-Sets." *Comparative Sociology* 9 (3): 397–418.

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

Thiem, Alrik. 2013. "Clearly Crisp, and Not Fuzzy: A Reassessment of the (Putative) Pitfalls of Multi-Value QVA." *Field Methods* 25 (2): 197–207.

Thiem, Alrik and Adrian Dusa. 2013. "QCA: A Package for Qualitative Comparative Analysis." *The R Journal* 5 (1): 87–97.

Measuring Partial Membership in Categories: Alternative Tools

Zachary Elkins
University of Texas, Austin
zelkins@austin.utexas.edu

Almost any attempt at classification runs into a boundary problem. Some cases fit neatly into one category, some fit one category only partially, and some fit multiple categories. This is a well-understood issue among both cognitive psychologists, who have documented how the brain's hard-wiring classifies stimuli, and taxonomists,[1] who seek to "soft-wire" additional sorting schemes. My focus here is mostly on the soft wiring. How, exactly, *should* researchers build classification systems—referred to here as taxonomies—that account for partial membership in categories, if at all? An important reference point is fuzzy sets, an intriguing concept that has gained some traction in sociology and political science. I explore a set of measurement strategies for assigning partial membership scores in the context of executive-legislative relations, a research domain overdue for innovation in conceptualization and measurement.

Measuring Partial Membership in Categories?

I define a "partial membership score" as a measure of the extent to which a particular case belongs to a specific category. While scholars disagree on the necessity and utility of calculating such a score, one common rationale for doing so is that this calculation fits nicely with how the human mind works. Are we, perhaps, hard-wired to classify information according to a partial membership process? Some important insights from cognitive psychology on how we process and classify phenomena lead us to think so.

In the "classical" view of categorization (Murphy 2004), categories are defined by necessary and jointly sufficient conditions for membership. For example, parliamentary democracies may be defined by "assembly confidence," wherein the executive is both selected and removed by the legislature. This view of concepts admits no borderline cases and treats each member of the category as a full instance of the concept, with no significant distinctions among members.

The modern view, associated closely with Ludwig Wittgenstein and Eleanor Rosch, shifted toward a more graded view of concepts, thereby challenging the idea of well-defined membership and non-membership. Wittgenstein's (1953) concept of "family resemblance" undermines the idea that there is any common (much less necessary) attribute of category members. In Wittgenstein's view, parliamentary systems might be a family of systems whose members share—in varying combinations—a substantial number of characteristics, such as executive decree, minimal legislative oversight of the executive, and a figure-head for head of state. Rosch's (e.g, 1975) large body of experimental work advances the idea that people differentiate with respect to the *degree* of belonging to a prototype. For instance, Rosch showed that—in the framework of protoypes—a chair is a highly typical instance of furniture, a bookcase less typical, and a piano even less so. This focus on degree of belonging shifted the understanding of classification processes away from the idea of sharply defined category membership based on the conception of necessary and sufficient conditions.

In political science, David Collier's (Collier and Mahon 1993; Collier and Levitsky 1997) work on classical versus radial subtyping highlighted the necessity of using graded approaches to categorization, particularly with central yet contested concepts like "democracy." Collier's work left political scientists with a stronger appreciation for partial membership in categories, though he stopped short of recommending particular measurement instruments with which to assign scores.

I should note a parallel set of studies in cognitive psychology that reveal a certain "categoriness" to the mind. That is, for some concepts at least, we tend to lump phenomena into classes and to minimize the conceptual distance between co-classified items and exaggerate the distance between cross-classified items. This phenomenon, *categorical perception*, is evident with phenomena such as color, sounds, and—I suspect—a fair number of learned categories such as those in social science (Harnad 1990).

In sum, there appears to be a strong basis in cognitive psychology for the idea that partial membership is central to our neurological hard-wiring and "natural" categorization. It also seems likely that the continuum underlying many of these

Thanks to David Collier and Pam Paxton for helpful comments on earlier drafts.

[1] In this context, taxonomist refers broadly to scholars concerned with classification, and not narrowly to specialists in biological taxonomy.

classification schemes is subject to perceptual discontinuities that lead to a natural clustering of items. Thus, membership, but membership by degree. The question remains, however, of how an analyst can adapt existing measurement practices to reflect these ideas, in particular with respect to the learned categories of social science.

A Point of Departure: Fuzzy Sets

These insights explain the appeal of fuzzy sets, which extend the logic of set theory to graded membership (Zadeh 1965, Smithson and Verkuilen 2006, Ragin 2008). Charles Ragin has taken the lead in introducing fuzzy sets to the social sciences, in connection with an analytical method known as Qualitative Comparative Analysis (QCA—the fuzzy set version is fsQCA). For our purposes, it is important to separate QCA, the inferential method, from the concept of fuzzy sets. While the utility of QCA is the subject of debate (including contributions to this newsletter), fuzzy sets, at least as a descriptive device, are considerably less contested. Still, much of the measurement technology for fuzzy sets (at least for social science) has developed in the context of fs/QCA, and so it makes sense to start there.

The appeal of fuzzy sets, as summarized by its proponents, is clear:

> With fuzzy sets, it is possible to have the best of both worlds, namely the precision that is prized by quantitative researchers and the use of substantive knowledge to calibrate measures that is central to qualitative research. (Ragin 2008: 82)

Assigning fuzzy set scores to cases is challenging, no matter how one does it. Charles Ragin has offered a transparent approach, which seems to have some currency among fs/QCA scholars. Ragin identifies two related methods, which he labels "direct" and "indirect" (Ragin 2008). Both, at least in his examples, build on continuous measures of an underlying concept. In his classification of countries into the set of "developed" countries, for instance, he uses a continuous base measure of GDP/capita.

For the direct method, researchers "calibrate" the measurement by identifying three "anchor points" in the base continuous measure: the points at which a case reaches (1) full membership, (2) full non-membership, and (3) the crossover point between membership and non-membership. Researchers use these values to sort cases into one set or another and compute scores between 0 and 1 by transforming deviations in GDP/capita from the cross-over point with a log-odds function.[2]

The indirect method is similar to the direct method, except that the analyst codes each case with one of the following six membership scores, all of which reflect the level of membership in a target set—for example, developed countries (Ragin 2008: 84). The six scores would be (1) full membership, (2) mostly in, but not fully, (3) more in than out, (4) more out than in, (5) mostly out, but not fully, or (6) full non-membership. The analyst assigns each of these categories an equally spaced number between zero and one (1.0, 0.8, 0.6,...) and then regresses the scores on the base measure (GDP per capita) using a fractional logit model. The predicted scores thereby become the fuzzy set scores. In some fsQCA applications, the number of scores may be greater or less than six.

As is probably clear, both the direct and indirect approaches to creating fuzzy-set membership require some strong theoretical assumptions regarding the location of the calibration points. (All measurement approaches, of course, lean on theory to some degree in order to build the ship at sea, as it were). It seems likely that the location of these calibration points will vary significantly across researchers (descriptive heterogeneity) and, relatedly, will vary with respect to the relationship of the measure with other constructs (causal heterogeneity). fsQCA seeks to take context into account, but the relevant features of context can readily be well beyond the reach of any standard approach to contextualization.

Consider an everyday example of descriptive heterogeneity. For any individual, there is some noticeable and abrupt cross-over point between cold and hot. But this cross-over point will depend on whether one is a Texan or a Minnesotan, young or old, playing soccer or watching it from the stands, and a never-ending list of other factors. If asked, each observer would identify a different cross-over point based on their own perception of temperature. In such a case, is it helpful to have a fuzzy-set score that indicates to which category a certain temperature belongs? Does it make sense to say that the temperature has a fuzzy-set membership of 0.43? Perhaps, but it will depend crucially on an inter-subjective and inter-contextual agreement about the location of crossover points.

It seems implausible that scholars can agree on crossover points for a great many political variables of importance, such as democracy or economic development. Ragin is quite candid on this point:

> The collective knowledge base of social scientists should provide a basis for the specification of precise calibrations. For example, armed with an adequate knowledge of development, social scientists should be able to specify the per capita income level that signals full membership in the set of developed countries. However, the social sciences are still in their infancy and this knowledge base does not exist. (Ragin 2008: 86)

This acknowledgment of the difficult theoretical exercise of assigning calibration points for economic development and GDP per capita is telling. These two make for a well-known concept/indicator pair and uncertainty in this domain suggests that these decisions will be even more fraught in other domains.

With respect to *causal* heterogeneity, it is quite possible that base measures like GDP/capita are related to outcomes along different functional forms. Imagine, for example, that two outcomes interest a researcher: democracy and happiness. To the extent that there are discontinuities in either of the two GDP/capita-outcome relationships, it is likely that democracy and happiness "kick in" at different levels of GDP/capita.

[2] See Ragin (2008: Chapter 5) for more detail.

If this is the case, the problem of causal heterogeneity highlights the risky rescaling process inherent in both the direct and indirect approaches to fuzzy-set membership. If one agrees with Ragin that full membership is hard to establish, even for familiar, widely-studied phenomena like development and per capita income, then the complex gradations of full membership and non-membership for many other phenomena may be illusive indeed. Fuzzy-set measures may well add a layer of complexity to the continuous measure, without a corresponding gain in meaning. Without a useful calibration point, the fuzzy-set measure rescales the base measure into units that are no longer directly observable or meaningful. Compare a GDP/capita of $4500 to a fuzzy-set membership in the set of developed countries equal to 0.43. While 0.43 is some function of GDP/capita, it is no longer directly observable, not particularly meaningful, and unclear whether two researchers with different outcomes of interest will interpret the measure in the same manner.

To be fair, fuzzy set scores are not alone in their intangibility. Many measurement strategies involve rescaling observable scores, either by constraining their quantities between endpoints or constraining their distributional parameters (e.g., normalizing, with a mean of zero and a standard deviation of one). Researchers must therefore ask two questions. First, how much does the rescaling procedure reduce interpretability? Second, does the gain of rescaling outweigh the cost regarding interpretation?

In fs/QCA, the benefit is presumably that the anchor points have real meaning, indicating full membership, full non-membership, and the point between the two. And, in fact, Ragin sees this calibration procedure as comparable to the creation of a Celsius scale, in which zero and 100 degrees mean something real with respect to the effects of temperature on water. To the extent that these calibration points *do* have real intersubjective meaning, then perhaps moving beyond concrete units is a large benefit of rescaling. But, as discussed above, the tenuous nature of the assignment process, even for merely establishing full membership and non-membership, makes it difficult to believe that the calibration points are *actually* interpretable in any consistent manner.

Fuzzy set measurements also seem to have some difficulty, at least as they are specified, in fully representing the meaning of a systematized concept. Let's think of some examples of categories and cases that exhibit "boundary" problems: Olives (fruit), Poker (sport), and Duckbill Platypi (mammals). These cases induce categorical head-scratching because they share attributes with both co-classified cases and cross-classified cases. In each case, they have been categorized as such because researchers have preferred to privilege one dimension (respectively, seeds, competition, and mammary glands) over another. However, other secondary characteristics are associated with cases in each category (again, respectively: sweetness, athleticism, and internal gestation).

So, variance within categories derives in part from multiple, semi-related dimensions of the concept. A satisfying measurement strategy would be one that could represent and test the dimensionality of the category using multiple measures. Combining multiple indicators both to represent the concept more fully and to improve reliability is a virtue of most measurement models, and something on which fuzzy sets—as conventionally measured—fail to capitalize. Conventional fuzzy-set measures typically identify, and measure membership for, each set/dimension separately.

A Framework for Evaluation

The limitations of current fuzzy-set measurement practices cast something of a shadow on the use of such measures in analysis. As Ragin (2008: 71) himself notes, "the key to useful fuzzy-set analysis is well-constructed fuzzy sets." Unfortunately, the measurement challenges leave the method—by its own criteria—limited in its applications to the social sciences. One wonders whether more attractive solutions are available for measuring partial membership scores. But more attractive in what way? Here, the foregoing examination of fuzzy-set measurement practices can be helpful. Not only do these practices serve as a focused reference point for comparison, but the comparison suggests some useful criteria for evaluating such measures.

Of course, researchers will have different analytic and descriptive uses for partial membership scores. However, some basic concerns seem relevant to anyone who builds or uses partial membership—concerns that I express here in terms of three points of inquiry. Whatever else they do, helpful measurement strategies should be able to shed some light on one or more of these points.

1. **Homogeneity within Categories**. How much diversity is evident in the categories in question? That is, to what degree can an analyst make the claim that categories are sufficiently uniform?

2. **Conceptual Architecture of Categories**. Which attributes are responsible for potential heterogeneity within a category? That is, can we identify the multiple dimensions, or components, that structure a category and produce its diversity?

3. **Degree of Membership in Categories**. To what degree (with what probability) does a particular case "belong" to a given category? That is, can we assign membership meaningfully and generate useful partial membership scores?

Alternative Measurement Approaches and an Application

This exploration of partial membership methods is not an abstract exercise for me. I have devoted many years to the Comparative Constitutions Project (CCP), with the goal of describing the world's written constitutions, historic and contemporary, and testing theories regarding the origins and consequences of constitutional choices (Elkins, Ginsburg, Melton 2013). By reading and re-reading 800 constitutions, I have become acutely aware that the constitutional landscape of the world's states is not particularly well-conceptualized. Specifically, we lack a well-developed sense of how constitutions express rights, duties, powers, and principles and how these properties co-vary. Moreover, we do not seem to have provided those who are in the business of *writing* constitutions

with a helpful conceptual framework for understanding their choices.

Executive-Legislative Relations

One of—if not *the*—most important decisions that constitutional designers must make is how to structure the roles of the executive and the legislature. For well over a century, political scientists (and constitutional drafters) have conceptualized this choice predominantly as one between two basic types: presidentialism and parliamentarism. At least since de Gaulle, a mixed type—sometimes called semi-presidentialism—has entered the academic and political dialogue. Though other scholars have suggested additional intermediate categories (e.g., Shugart and Carey 1992), for the purpose of this analysis we will consider these three basic categories.

Perhaps *because* of its familiarity, this typology is highly relevant to our purposes. The definitional criteria are rarely in dispute. Presidential and parliamentary systems are usually distinguished by two related attributes—the procedures for the *election* and *survival* of the head of government. In a parliamentary system, the assembly selects the head of government, who serves at their pleasure; in a presidential system, citizens select the president, who serves for a fixed term.

In conjunction with these defining attributes, strong stereotypes have developed about what these types look like across a set of other secondary properties, which are often denoted by three prototypic systems: the U.S., the British, and the French. Mounting evidence, however, suggests that these three types mask great diversity in executive-legislative systems (Shugart and Carey 1992, Tsebelis 2002, Cheibub et al. 2013); many presidential systems take on parliamentary attributes and many parliamentary systems take on presidential attributes. How can we take these partial memberships into account?

Strategy One: A Continuous Scale Instead of Classification

A first potential strategy is to sidestep classification entirely—thereby deliberately setting aside these three questions. After all, the baseline conditions for measuring partial membership require that concepts exhibit some combination of "categoriness" and partial membership in such categories. Many concepts, it seems, exhibit one but not the other. The potential objection to measuring partial membership, therefore, is that a researcher either does not recognize any discontinuity in a variable or does not know where to draw the boundary lines.

In such a case, the dominant strategy is to build continuous scales that tap gradations on a dimension (or dimensions) of the particular concept. It is a familiar task for many scholars. Indeed, the goal in most measurement models is not to classify units, but rather to assign a score to the units across a continuous measure. Ideally, one would assemble a set of multiple measurement items that both represent the systematized concept adequately and improve the reliability of the overall score. A long tradition of measurement strategies, based principally on covariance structure modeling, allow for the construction and testing of such measures. But even simpler scaling techniques can deliver measures with graded scores that might satisfy researchers ostensibly interested in partial membership.

Thus, in the case of executive-legislative relations, a researcher may be tempted to eschew classification—even a classification system as familiar as the dichotomous presidentialism versus parliamentarism—and instead opt for a continuous scale that taps a *primary* dimension of the distinction. Some very accomplished scholars have built exactly these sorts of scales, which are central to the executive-legislature relationship (e.g., Shugart and Carey 1992, Fish and Kroenig 2009, Tsebelis 2002, Lijphart 2012). To unpack some of the issues involved in this sort of scaling, consider *another* scale: scope of executive authority (power) as encoded in written constitutions (Elkins, Ginsburg, and Melton 2013).

The data for this scale include measures of a comprehensive range of executive and legislative powers. However, constructing a valid scale from these characteristics requires special attention to aggregation and weighting. Because powers are substitutable, it is not especially meaningful to simply add powers, such as the executive veto or the executive's initiation of legislation. While executives may win in one arena (say, the budget), this may come only after threatening the legislature in another arena (say, military action); in general, power in one domain will likely provide clout in another domain.

The implication of substitutability for measurement is that these various powers will not necessarily correlate, which limits their utility in standard measurement models.[3] More precisely, these powers are not typical "reflective" indicators (manifestations, or reflections, of the latent construct); rather, they are "formative" indicators, in that they are causes of (or routes to) to a latent variable.[4] The set-up for formative indicators is, importantly, quite different from standard reflective models. The weighting for a given item in the scale is derived not from its intercorrelation with other items, but from its prediction of variables that are manifestations of the latent construct. So, in order to identify the measurement model, one needs to specify at least one reflective indicator (i.e., the outcome) of the latent variable along with the host of formative indicators (i.e., the potential causes).

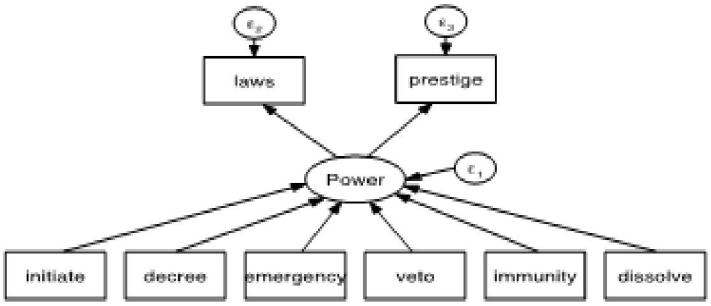
Figure 1 depicts the relationships we theorize in a measurement model of constitutional executive power. In measurement modeling terms, this structure is known as a MIMIC (Multiple indicator, multiple cause) model. The outcome variables are (1) the executive's success rate in passing bills that originate from the executive and (2) the relative prestige of the office (labeled as *laws* and *prestige* in the figure). The substitutable power variables (which, again, are treated as causal) all relate to a constitution's degree of executive power, shown in the six attributes in the figure.

[3] Standard measurement models develop weights based on an item's intercorrelation with other items.

[4] The idea of a latent variable is most useful if not over-interpreted—in the reified sense of an underlying phenomenon in the real world; but rather is treated as a metaphor for the conceptual understanding of how and why these items (i.e., characteristics) may be interrelated.

Figure 1: MIMIC (Multiple Indicators Multiple Causes

Measurement Model of Executive Power



Legend
laws = success rate of executive-initiated bills
prestige = expert coding of executive importance
initiate = executive empowered to initiate legislation
decree = executive empowered to enact legislative decrees
emergency = executive empowered to declare state of emergency
veto = executive empowered to veto legislation
immunity = executive granted immunity from prosecution
dissolve = executive empowered to dissolve the legislature

Clearly, this sort of measurement scheme sidesteps classification and is therefore not helpful in answering the three evaluative questions I raise above. However, it may also be possible to approximate, or at least to explore, partial-set membership using a continuous measure. One approach would be to test, iteratively, the effect of any discontinuities by employing a combination of continuous and dichotomous variables in a regression and identifying, diagnostically, any discontinuity in the association between the outcome and the continuum under consideration.

It may also be possible to combine the continuous measure with a related classification system in order to assess internal diversity and, even, approximate degrees of membership. So, in our example, one could pair a continuous measure of executive power together with an existing classification of parliamentary and presidential systems. Assuming that executive power constitutes a dominant dimension of presidentialism and parliamentarism, one may be able to describe degrees of category membership.

Of course, this approach is probably not too dissimilar from traditional fuzzy set practices, as described above, and carries with it some of the aforementioned limitations. For example, one must make the strong assumption that the continuous measure is a dominant dimension of presidentialism and parliamentarism. As we suggest above, it seems more likely that categories would be defined by a mix of traits and that a single continuum would describe only one of multiple dimensions of the concept. Still, it may be instructive to calculate the variation with respect to the one dimension (executive power) within categories (presidentialism and parliamentarism) and even to use the interaction of the two in analytic models.

One would want to be clear that these are not, strictly speaking, partial membership scores. Still the joint effect of membership and variation on a primary dimension of membership would amount to something close to partial membership. Nevertheless, one might more satisfactorily measure partial membership with a more multidimensional approach.

Strategy Two: Similarity-Based Measures of Family Resemblance

An approach more consistent with Wittgenstein's idea of family resemblance and Rosch's prototype analysis is to group (or at least measure similarities among) cases based on a set of relevant characteristics. As in various algorithms used in cluster analysis, the idea here is to calculate quantities of similarity or distance among cases in light of their scores on a set of presumably multidimensional characteristics. Further specifying a prototypical case allows researchers to calculate an explicit measure of degree of membership: the distance between each case and the prototype.

These kinds of methods are often employed in a more exploratory fashion. They are useful for identifying units that flock together, exploring alternative classification strategies, and discovering different types or "species." However, as I suggest above, it is is possible to calculate partial membership scores in a straightforward manner with these methods.

A first step might be to classify cases based on one or more definitional attributes. In the case of presidentialism and parliamentarism, such clasifications abound: here we use an authoritative coding by Cheibub, who classifies cases based on the selection and survival properties of the executive, as is conventional. The next step is to identify a set of secondary characteristics that are associated with membership in one or more of the classes. It is then possible to build partial membership scores for each case's "resemblance" to identifiable "families," based on their values on these secondary characteristics.

For example, Table 1 identifies seven attributes typically associated with presidentialism, semi-presidentialism, and parliamentrism. From these I calculate a simple measure of similarity (the Pearson correlation *between* cases and *across* the seven attributes) for each dyad in the data. The similarity to a prototypical case constitutes a measure of partial membership.

For the 108 constitutions included in this analysis,[5] Fig-

[5] The sample includes all independent states for which the constitution specifies executive-legislative relationships to a sufficient de-

Table 1: Characteristics of Executive-Legislative Systems

	System		
	Presidential	Semi-Presidential	Parliamentary
Assembly Confidence	No	For head of govt	Yes
Executive decree	No	Depends	Yes
Emergency powers	Strong	Strong	Weak
Initiation of legislation	Legislature	Depends	Executive
Legislative oversight	Yes	Depends	No
Executive veto	Yes	Depends	No
Cabinet appointment	Executive	Depends	Legislature

ure 2 depicts the distribution of the measure of similarity to the United States, grouping cases according to whether they are categorized as presidential, semi-presidential, parliamentary based on the defining attributes. A score of 1.0 would represent perfect similarity (with the same values on all component attributes); 0.0 would reflect the absence of any shared attributes.

One way to interpret these scores is as a measure of the degree to which cases belong in the "presidential" category, defined by the U.S. prototype. The results are startling, though they confirm something that institutional researchers (e.g., Shugart and Carey 1992) have suspected for years: there is enormous heterogeneity within the classic categories.

Specifically, and counter-intuitively, the mean similarity scores vis-à-vis the United States differ relatively little across the three system types. Indeed, Parliamentary systems are, on average, more similar to the U.S. prototype than are presidential systems—a rather shocking result. One could also analyze the variance, by category, of each of the secondary characteristics to determine which ones are more or less responsible for the lack of family resemblance. It turns out that all of these characteristics vary substantially (though to varying degrees) within the three categories (Cheibub, Elkins, Ginsburg 2013).

All of this to say that this measure appears to provide satisfactory answers to the three evaluative questions identified above. The scores allow us to assess the degree of heterogeneity and the sources of diversity within categories, and to measure degree of membership. One could even imagine further analyses in which one employed these family resemblance scores in statistical models that use the classic typology, perhaps by substituting the "degree of presidentialism" for a binary variable of presidentialism in a regression analysis.

So what do you do when you identify a family that does not appear to exhibit much resemblance among its members? Do you stop speaking of those families altogether? Do you reject the findings and seek other, overlooked, characteristics that are more related to the family line? Or do you use these new measures of familial distance to speak more accurately of the family's members—say, of siblings versus second, third, and fourth cousins? In the case of executive-legislative relations, the second and third responses seem worth pursuing. The family still matters: presidentialism and parliamentarism still connote remarkably important differences between systems.

Of course, it is possible that I tested family resemblances with characteristics that are not connected to the family's "DNA." To evaluate this potential problem, it is useful to revisit our expectations regarding the secondary characteristics and test the diversity question with *other* known concomitants of presidentialism and parliamentarism. In fact, it really does make sense to speak of close and distant relatives, given that in further testing we do not find new evidence of homogenous families. Also, we are convinced that the secondary characteristics originally tested really *are* attributes closely associated with the family. It is therefore plausible, as in Figure 2, to suggest that Liberia (LBR) is a close Presidential member, while Nicaragua (NIC) is something of a distant cousin.

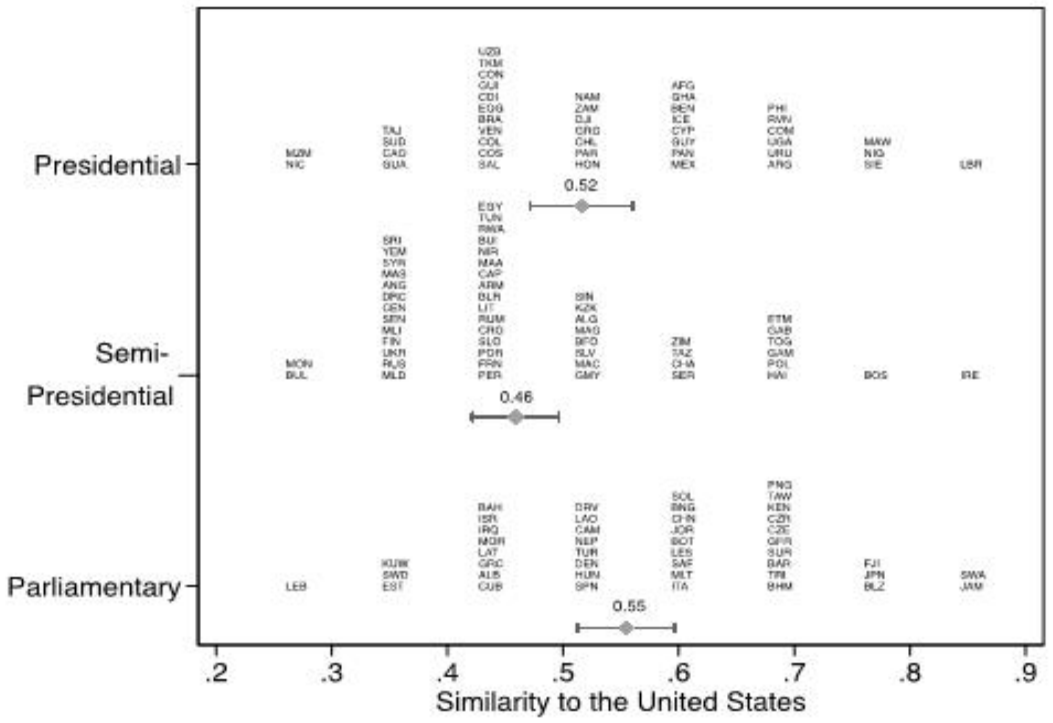
One will note that these families are only plausible if we are able to identify defining attributes with which to make such designations. That makes sense in the case of the presidential and parliamentary families. But what can we do if we are in a purely Wittgenstein/Rosch-like world (as opposed to a Linnean taxonomical world), in which we do not have the luxury of necessary and sufficient conditions to define categories?

Strategy Three: Latent Class Analysis

The cluster-analytic techniques of Strategy Two are illuminating, but they are for the most part exploratory. They lack the statistical properties that would justify more precise statements about the degree to which particular units belong to categories. Latent Class Analysis (LCA), by contrast, provides somewhat more precise answers in this regard. It is a productive third strategy, given that we are concerned with categorical distinctions. LCA is a version of cluster analysis in which one analyzes the attributes that characterize each category, and uses estimates of the clustering of categories to sort cases

gree as to allow classification with respect to the three systems.

Figure 2: Surprising Similarity to the United States of Presidential, Semi-Presidential, and Parliamentary Systems



Note: N=137 national constitutions as of June 2013. Table includes independent states with constitutions that specify procedures for executive selection and survival. A score of 1.0 signifies perfect similarity to the United States. Average similarity scores: Presidential 0.52; Semi-Presidential 0.46; and Parliamentary 0.55.

into appropriate groupings. This analysis can estimate the probability that cases belong to a particular category, as well as the association of the various items with each category.

In order to understand its uses, think of an everyday concept, *autism*, the diagnosis of which many modern parents puzzle over at some point. Like most of the learned concepts under consideration here, autism is highly multidimensional, and characterized by a variety of social, cognitive, and emotional symptoms. It is also regularly treated as a category with *partial* memberships within this multidimensional space (cases are said to be somewhere on the "spectrum"), but memberships seem to matter nonetheless. "Having it" triggers certain treatments, certain accommodations, and certain sympathies. But how to assign membership given these multiple continuous dimensions? LCA offers one approach, which not only allows for the estimation of membership in a single category (e.g., Autism) but also the estimation of membership in subcategories (e.g., Aspergers).

The parallel to executive-legislative relations is striking. Presidentialism, parliamentarism, and semi-presidentialism are as multidimensional and graded as any other category: however, the categories themselves matter. What can LCA tell us about these classes and their members? An initial answer is suggested by building on the characteristics listed in Table 1 and performing an LCA analysis. Table 2 reports a critical set of quantities from this analysis: the probability of membership in the three broad categories, for a selected set of countries. These memberships, then, are *graded* and *multiple*. At the same time, the probabilities suggest that cases may be assigned to one of the categories, based on the highest probability in each row (shown in bold in Table 2). The number of categories analyzed was fixed at three but that number can be permitted to vary and, like the number of dimensions in confirmatory factor analysis, should be subjected to close scrutiny.

The labeling of each category requires interpretation. In Table 2, I have assigned labels to the categories, based on (a) the clustering of cases; and (b) case scores on the defining attributes of the categories.[6] The results in Table 2 suggest some intriguing answers with respect to partial membership. In general, membership scores seem to corroborate those calculated in the cluster analysis above. The difference, now, is a much more precise sense of how and why they do and do not fit well. So, is Liberia presidential? Yes, unequivocally so. It belongs to that category with a probability of 0.91 and to the others at less than 0.30. Brazil, however, might just as easily be categorized as semi-presidential ($p = 0.56$) as presidential ($p = 0.60$). And so on.

With respect to the criteria identified, we might think of LCA as something like a more precise version of the family

[6] The conditional probability of the items for each category is not reported here. This probability essentially maps the relationship of the items with the categories.

Table 2: Latent Class Analysis: Probability of Membership in Three Derived Categories for Selected Cases

	Categories, with Interpretation Shown in Brackets		
	1 [Presidential]	**2** [Semi-Presidential]	**3** [Parliamentary]
Guatemala	**0.46**	0.32	0.22
Brazil	**0.60**	0.56	0.33
Peru	**0.87**	0.34	0.17
Liberia	**0.91**	0.28	0.12
Belarus	**0.68**	0.54	0.34
Ukraine	**0.56**	0.45	0.23
Russia	0.38	**0.65**	0.13
Denmark	0.21	0.39	**0.69**
Spain	0.15	0.34	**0.95**

Note: Numbers in bold indicate the highest category in each row. Substantive interpretation of categories is based on visual inspection of: (a) clustering of cases; and (b) case scores on the defining attributes of the categories.

resemblance measures in Strategy 2. That is, LCA allows us to describe the diversity within categories and to assign partial membership scores to individual cases. It also facilitates an investigation of the architecture of the various categories through an analysis of the correspondence between the various attributes and category membership (though that analysis is not shown here). The advantage over more informal clustering methods is a nuanced one. The LCA results, like those generated in the cluster analysis, have a tight connection to the idea of partial membership. The difference is that the LCA results have a stronger, or at least more widely understood, grounding in statistical and measurement theory.

Further Observations on Partial Membership in Categories

A final consideration that emerges from our evaluation of partial membership strategies concerns the interpretation of scores. Partial membership is not directly observable and the scores generated from any of the methods under discussion will be scientific constructs. Some of these constructs, however, are simply more meaningful than others. Typical fuzzy set methods rely upon a calibration approach that depends upon some rather aggressive assumptions about the location of set-membership boundaries, which then serve as reference points. Ultimately, it is not entirely clear what the scores surrounding these boundaries mean, exactly. 0.4 may mean that a case is slightly more out than in (if the set cutoff is 0.5), but that relative judgment is not itself particularly easy to grasp, or to convey to others.

In other methods, however, 0.4 may well have a more comprehensible, or at least more *established*, meaning. In a measure of family resemblance, that score—depending upon how similarity is measured—may mean that a case shares 40 percent of some group of characteristics with the category's prototype (if similarity is measured as percent matching) or that it correlates at 0.4 with the set's prototype (if similarity is measured as the correlation between two cases across their characteristics). 0.4 in a LCA model suggests that a case belongs to a particular category with a probability of 0.4. Any of these interpretations are just as unobserved as are those in the fuzzy set context. The difference is that these units are constructed as mathematical concepts (probabilities, correlations, percentage) that need no introduction and have well understood properties. Ultimately, that sort of resonance will be important, at least in a descriptive endeavor.

Conclusion

The idea of building taxonomies with partial membership is compelling. The idea makes even more sense once we understand insights from cognitive psychology about how our minds process stimuli. But how to operationalize the idea of partial membership? The concept of fuzzy sets is helpful, but the measurement tools associated with that approach in the social sciences, at least, are quite underdeveloped. Still, identifying the shortcomings of extant fuzzy-set measurement practices focuses our attention on some desirable properties of partial-membership measures—and establishes a basis for evaluation. Helpful measurement properties are found in some alternatives to fuzzy sets. In particular, clustering and latent-class analytic methods generate family resemblance scores that seem to deliver the punch that we expect from partial membership. The illustrations in the domain of executive-legislative relations help us describe and diagnose the bounded nature of some well-established categories, presidentialism and parliamentarism.

The illustration raises a set of questions about what, exactly, to do with partial membership scores. Description is a worthy end in itself. However, once acquainted with a set of valid partial membership scores, one will also be inclined to put partial membership scores to work in explanatory models. This essay does not tackle that problem. However, it is easy to imagine statistical tests in which a binary measure of, say, presidentialism is replaced with a family resemblance measure of that class. Indeed, the evaluation of partial membership in explanatory models would be a productive next step.

References

Cheibub, Jose Antonio. 2007. *Presidentialism, Parliamentarism, and Democracy*. Cambridge: Cambridge University Press.

Cheibub, Jose Antonio, Zachary Elkins, and Tom Ginsburg. 2013. "Beyond Presidentialism and Parliamentarism." *British Journal of Political Science*. Published first online, November 14, 2013. http://dx.doi.org/10.1017/S000712341300032X.

Collier, David and James E. Mahon, Jr. 1993. "Conceptual 'Stretching' Revisited: Adapting Categories in Comparative Analysis." *American Political Science Review* 87 (4): 845–855.

Collier, David and Steven Levitsky. "Democracy with Adjectives." *World Politics* 49 (3): 430-451.

Diamond, Jared. 1997. *Guns, Germs, and Steel: The Fates of Human Societies*. New York: Norton.

Elkins, Zachary. 2000. "Gradations of Democracy? Empirical Tests of Alternative Conceptualizations." *American Journal of Political Science* 44 (2): 293–300.

Elkins, Zachary, Tom Ginsburg, and James Melton. 2013. The Comparative Constitutions Project. Datasource. http://comparativeconstitutionsproject.org/ (Accessed April, 22, 2014).

Fish, M. Steven and Mathew Kroenig. 2011. *Handbook of National Legislatures*. Cambridge: Cambridge University Press.

Harnad, Stevan R., ed. 1990. *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.

Lijphart, Arendt. 2012. *Patterns of Democracy: Government Forms and Performance in Thirty-Six Countries*. New Haven: Yale University Press.

Murphy, Gregory L. 2004. *The Big Book of Concepts*. Cambridge: MIT Press.

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

Rosch, Eleanor. 1975. "Cognitive Representations of Semantic Categories." *Journal of Experimental Psychology: General* 104 (3): 192–233.

Seawright, Jason. 2013. "Warrantability." Presented at the Annual Meeting of the American Political Science Association, Chicago.

Shugart, Matthew and John Carey. 1992. *Presidents and Assemblies: Constitutional Design and Electoral Dynamics*. Cambridge: Cambridge University Press.

Tsebelis, George. 2002. *Veto Players: How Institutions Work*. Princeton: Princeton University Press.

Wittgenstein, Ludwig, 1953 [2001]. *Philosophical Investigations*. Oxford: Blackwell Publishing.

Zadeh, Lotfi. 1965. "Fuzzy Sets." *Information and Control* 8 (3): 338–353.

Analyzing Interactions: Four Alternative Models

Bear F. Braumoeller
Ohio State University
braumoeller.1@osu.edu

A Car-Buyer's Guide

To a consumer of methods in political science, the act of choosing an appropriate model can resemble the process of buying a new car. The journal article that introduces the method generally plays up its strengths while giving short shrift to its potential weaknesses, and other users often have little incentive to dwell on its potential shortcomings. "Check out this year's new model!" the author seems to say. "It lets you make asymptotically unbiased estimates with fewer observations than your existing model—which," and here the voice drops to a whisper—"can provide *really* terrible answers in circumstances like these. And nothing could be simpler to use! Just download this Stata package and add a single line of code to your batch file."

Practitioners are generally looking for a tool to solve a particular problem, not an in-depth discussion of the pros and cons of a particular method or set of methods. They often don't stop to take a close look under the hood or to ask the hard questions. "Asymptotically unbiased, you say... but what about the precision?" "Ah, you say you want *precision*? Perhaps you'd like to take a look at this model over here...."

As a result, many scholars doing substantive research tend to hop from one flashy new model to another without fully exploring the capabilities and limitations of each. When next year's model comes along, they jump on that, every bit as disdainful of last year's methods as they were of those that came the year before. Should they come across a crosstab or a chi-squared test in a published paper, they shake their heads sadly at the author's methodological naïveté. It rarely occurs to them that the chi-squared test has been chugging along reliably for more than a century, while newer, flashier models have ended up in the ditch.

Rectifying this situation mainly involves more, and better, methods training for practitioners. In the short run, however, we can offer some straightforward advice to applied researchers to help get to the heart of the issue. Perhaps the most important of these is this: You rarely get something for nothing. More inferential oomph generally comes at a cost, and it is important to know what that cost is before adopting the model.

To illustrate this point, I will discuss four different ways to model interactions: fs/QCA, multiplicative interaction terms, a stochastic frontier model, and Boolean logit. They are located, roughly, on a spectrum between assumption-intensive (fs/QCA) and information-intensive (Boolean logit). Each has some advantages vis-à-vis the others, but in every case those advan-

Acknowledgement: I am grateful to David Collier, Kimberly Twist, Katherine Michel and Alisan Varney for comments on earlier drafts.

tages come at a cost. In some cases, the cost is reliance on a set of assumptions that may or may not hold; in others, the cost can only be paid by obtaining lots and lots of data. The general lesson—assumptions take the place of data, and it is usually hard to relax the former without more of the latter—is a worthwhile heuristic for scholars trying to choose among them.

What Are Interactions?

Before we discuss different ways to model interactions, it is important to understand what interactions are, especially given the variety of terms that have been used to describe different facets of the same thing.

An interaction effect among independent variables (or, in the QCA tradition, "conditions") occurs when a change in the value of one independent variable or condition (call it *X1*) alters the impact of another independent variable or condition (*X2*) on the dependent variable or outcome (*Y*). To take a simple example, changes in the amount of sunlight a plant is exposed to make little difference if the plant does not receive any water but make a substantial difference if it does: water (or its absence) moderates the impact of sunlight on plant growth. The converse is true as well, of course: the amount of sunlight that a plant receives moderates the impact of water on plant growth.

It is worth noting that this is generally the case: when two variables or conditions interact, each moderates the impact of the other. It may sometimes seem that one is doing the causal heavy lifting and the other does nothing but moderate its effects—for example, a healthy diet and exercise increase life expectancy, but a hefty dose of arsenic can dramatically attenuate their impact. At first blush it hardly seems that the converse could be true, but that is largely because of the poison's efficacy: at smaller doses, arsenic *does* have less of an effect on healthy people than it does on the infirm.

The arsenic example illustrates a useful limiting case of interactions—namely, threshold effects, or necessary and sufficient conditions. Ingestion of a large quantity of arsenic is sufficient to cause death in humans, more or less regardless of the other conditions present at the time.[1] Another, more awkward way of saying the same thing is that non-ingestion of that same quantity of arsenic is a necessary condition for life. The relationship described by these two statements has two important characteristics: it is *universally* interactive (arsenic moderates the impact of everything else on longevity) and it allows few if any counterexamples.

Under such circumstances, the quantity of interest to researchers is often not the average value of *Y* that is associated with a given value of *X1* but rather the *threshold*—the *X1*-*Y* line or curve below which (or above which) no observations are found. To revert to the plant growth example for a moment, we could imagine that no plants of a certain type could grow taller than 1" with only a teaspoon of water per week, regardless of the amount of sun they get; none could grow taller than 2" with only two teaspoons of water a week; and so on. If we plot inches of height vs. teaspoons of water, we should see a threshold at *Y=X1* above which no observations will be found.

It is worth noting that, although data thresholds and necessary/sufficient relationships are fungible—necessary and sufficient conditions define a threshold above or below which few if any observations should be found, and any upper or lower threshold could be interpreted as a boundary implied by a necessary or sufficient condition—the latter terminology seems to evoke a more rock-ribbed response from social scientists. The qualities of "necessity" and "sufficiency," if taken literally (and how else do we take things?), seem to demand a complete absence of counterexamples, while a threshold that bounds a data region could easily be designed to accommodate a bit of measurement error or probabilism. This is why we find that, despite the apparently deterministic nature of necessity and sufficiency, Dion (1998) and Braumoeller and Goertz (2000) explore the question of how many counterexamples are required to reject a hypothesis of necessity, while Clark, Gilligan and Golder (2006) subsume both interactions and necessary/sufficient conditions under the general heading of "asymmetric causation" and argue that interaction terms are sufficient to capture them.[2]

Semantics aside, the key point to convey here is that "interaction" can mean interactions among independent variables or conditions in the model (as rain modifies the impact of sunlight on plant growth, and vice versa), or it can mean a threshold effect that modifies the impact of variables *outside* the model (as arsenic modifies the impact of everything else on life expectancy).

Under the Hood in an Interaction Model

The next issue, before we discuss the pros and cons of different estimators, is the question of what is to be estimated in an interaction model. Obviously, we need to start with *measures* of the phenomena that are interacting. Not so obviously, those measures may or may not have to be estimated from other data. They may or may not have obvious units. The scale of those units may or may not be cardinal. In a perfect world of realized observations measured on a tidy, cardinal scale, measurement requires little thought. In practice, we don't often get to ignore all of these issues.

We also need to know just how the data interact. What combinations of independent variables, in what configuration, are associated with increases in the dependent variable? Are we modeling an instance of *conjunctural* causation (*X1* and *X2* produce *Y*)? An instance of *substitutability*, or *disjunctural*

[1] For the sake of the illustration, I ignore last-minute interventions—first aid; a bullet; and so on—that could render the impact of the arsenic moot. It is difficult to find a necessary or sufficient condition short of the sun going supernova that could not theoretically be counteracted by prompt action of some sort, but to quibble over whether that makes them *truly* necessary or sufficient is to miss the point.

[2] As someone with an (admittedly aging) dog in this fight, I respectfully disagree on this point. Necessary and sufficient conditions are interactive, but interactive relationships are not necessarily ones of necessity or sufficiency. Interaction terms do very well at capturing interactive relationships but are not designed to answer the question of whether they are regular enough to constitute relationships or sufficiency—indeed, this is precisely the critique of interaction terms that Ragin (2013) offers and which I discuss below.

causation (*X1* or *X2* produces *Y*)? Some combination of the two, as in the case of INUS causation ([*X1* and *X2*] or [*X3* and *X4*] produces *Y*)? How do we know?

Moreover, we need some measure of the extent to which changes in independent variables, individually or jointly, are associated with changes in the dependent variable. In large-N models, these quantities are generally referred to as *coefficients*: if $Y=\beta 0 + \beta 1X1 + e$, for example, $\beta 1$ is the coefficient that translates changes in *X1* into changes in *Y*.[3] Coefficients typically describe *average effects*—that is, the impact of a unit change in *X1* on the value of *Y*, on average—but in some models they describe a *threshold*—a boundary between possible and impossible or near-impossible outcomes.

Finally, when any of these quantities is estimated, we ideally want a *measure of uncertainty* associated with the estimate so that we can know, for example, how likely it is that we would have seen an effect of a given magnitude or larger by chance.

These are all quantities that interaction models have in common. They are generally either estimated or assumed, depending on how much data can be brought to bear. Assumptions may or may not be accurate; we generally make them when we do not have the data necessary for estimation, so often the best we can do is explore how robust our answers are to other reasonable assumptions. Estimation may require a little data or a lot; in the worst cases, it may require far more data than we actually possess. Accordingly, we can learn quite a bit about a model's strengths and weaknesses by asking five simple questions of our estimators:

What does the estimator assume?
What happens if the assumptions are wrong?
What has to be estimated?
How much information do we need to estimate it?
Taking all this into account, what are its weaknesses?

Four Models of Interaction

To make this discussion more concrete, and to provide useful guidance for the car-buying practitioner who motivated this essay, I will explore four of the different models for estimating interaction effects among variables.[4]

The first, and by far the most common, is the simple interaction term, the multiplicative relationship known to every second-semester econometrics student and expressed in the equation $Y=\beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X1X2 + e$. A simple example might be a test of an argument that deterrence (*Y*) is a function of both capabilities (*X1*) and resolve (*X2*)—the idea being that deterrence will very often fail in the absence of either one.

The second model, which has gained a significant foothold especially in European academic circles, is fuzzy-set qualitative comparative analysis (fs/QCA), which uses Boolean minimization to identify combinations of variables that comprise thresholds. To continue the example, an fs/QCA analysis of deterrence would examine four kinds of cases, corresponding to the four possible combinations of independent variables, and would assess which combination(s) produce(s) deterrence success with very high probability.

The third model, born of production frontier modeling in the economics literature, is the stochastic frontier model. A straightforward version of this model starts with the Cobb-Douglas production function, $Y=AL^{\beta 1}K^{\beta 2}$, which models the relationship between production (*Y*), labor (*L*), capital (*K*), and factor productivity (*A*), and adds an error term with two components: a Gaussian one to capture the uncertainty around the estimated frontier and a skewed nonpositive one to allow for production inefficiency. The intuition is more straightforward than the math might suggest: the model is designed for situations in which the combination of labor and capital defines the upper threshold for production, our estimate of that threshold is a bit noisy and individual firms may fall short of that threshold due to inefficiency. In this case the functional form of our deterrence model would look like $Y=AX1^{\beta 1}X2^{\beta 2}$, where *Y* is a continuous variable that equals zero if either *X1* or *X2* equals zero and a variety of exogenous factors could cause states to fall short of their potential to deter.

The fourth model, Boolean logit, is a large-N, binary dependent variable technique designed to capture highly complex interactions (Braumoeller 2003). It is an extension of bivariate logit with partial observability, a technique for modeling conjunctural interaction between two latent (unobserved) variables: "partial observability" refers to the fact that we can observe the product of the two dependent variables in the model but not the values of the variables themselves. While the estimation of these quantities requires still more information, it is a very valuable asset when no direct measures are available. In the running example of deterrence, for instance, we might want to test the argument that deterrence success is a function of capabilities and resolve, but we may only have measures of the determinants of capabilities and resolve,

$$\Pr(\text{deter}=1) = \Pr(\text{capability}=1)^* \times \Pr(\text{resolve}=1)$$
$$\Pr(\text{capability}=1)^* = \Lambda(\beta 01 + \beta 11X1 + \beta 21X2)$$
$$\Pr(\text{resolve}=1)^* = \Lambda(\beta 02 + \beta 12X3 + \beta 22X4),$$

where $\Lambda(X\beta)$ represents the logit function, $1/(1+e^{-X\beta})$, *X1–X2* represent the determinants of capabilities, and *X3–X4* represent the determinants of resolve. The asterisk (*) denotes the fact that the probabilities of having sufficient capabilities to deter (given resolve) and having sufficient resolve to deter (given capabilities) are unobserved and must be estimated based on their determinants and on the assumption that their product is associated with deterrence success.

These four models represent a broad range of methods for analyzing interactions with a wide mix of assumptions and tools for estimation.

Measurement. For measurement, fs/QCA requires the re-

[3] Contrary to popular belief, if $Y=\beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X1X2 + e$, the coefficients $\beta 1$ and $\beta 2$ do *not* represent the unconditional impact of *X1* and *X2* on *Y* simply because in an interaction term *there is no such thing* as an unconditional effect. This point cannot be repeated often enough.

[4] I omit perhaps the most straight forward test of all—crosstabs—because most of the summary statistics for crosstabs capture association rather than interaction. For two rare exceptions to this generalization see Hildebrand, Liang and Rosenthal (1976) and Braumoeller and Goertz (2000).

searcher to provide a theoretically-grounded estimate of a case's degree of membership in a fuzzy set; Russia's membership in the set of democratic states, for example, might be 0.3, while that of the United States might be 0.9. These are theory-laden observations, perhaps even more so than most in that they rely on the theorist to aggregate the various facets of democracy into a single metric of membership. The measurements used in interaction terms and stochastic frontier models vary quite a bit, from dummy variables to ordinal to cardinal measures, but they are generally realized rather than latent. The interactions in a Boolean logit model take place among latent variables—like capabilities and resolve, above—whose values must be estimated.

Coefficients. Each of the three statistical procedures estimates coefficients, as one might expect. fs/QCA does not have coefficients in the standard sense of the word, but if we think about a coefficient more generally as measuring the impact of the independent variables on the dependent variable, the implicit coefficient is either 0 or 1. That is, if sunlight (S) and water (W) are individually necessary and jointly sufficient for plant growth (PG), all of which are measured as either absent or present, the resulting equation would be

$$PG = WS$$

which, in a nutshell, means that the coefficients on W and S are 0 and the coefficient on WS is 1. They are not free to be anything else, given fs/QCA's deterministic framework. Similarly, if "some" water and "some" sunlight are necessary and sufficient for "some" growth, the assumed coefficient is 1: there really isn't room for "a little" water to produce "a lot of" growth. If such an observation were found, it would be above the $X=Y$ threshold line on the scatterplot.

Moreover, both fs/QCA and Boolean logit extend the logic of interaction to capture arguments with a large number of latent variables and any combination of conjunctural and disjunctural interactions.[5] For example, if deterrence success depended on capabilities *and* resolve, but resolve could arise from either situational or dispositional sources—that is, deter = capability AND (situational resolve OR dispositional resolve)—, a useful Boolean-logit variant of the above equation might look like

$$Pr(deter=1) = Pr(capability=1)^* \times 1-([1-Pr(resolve_{sit}=1)]^* \times [1-Pr(resolve_{disp}=1)])$$
$$Pr(capability=1)^* = \Lambda(\beta 01 + \beta 11X1 + \beta 21X2)$$
$$Pr(resolve_{sit}=1)^* = \Lambda(\beta 02 + \beta 12X3 + \beta 22X4)$$
$$Pr(resolve_{disp}=1)^* = \Lambda(\beta 03 + \beta 13X5 + \beta 23X6 + \beta 33X7)$$

Thresholds. While interaction terms are not focused on measuring thresholds,[6] fs/QCA and stochastic frontier models are, and thresholds can often be derived from Boolean logit models. They differ significantly in terms of how they go about doing so, however. fs/QCA assumes by default that its variables are commensurate, so that (for example) if X is necessary for Y, it is also true that a one-third fuzzy membership in X is necessary for a one-third fuzzy membership in Y. If observations are found close to the line, Ragin allows for a "fuzzy adjustment" to capture the slippage—only fair, given that the original measurements are likely to be quite approximate. Stochastic frontier models, by contrast, estimate the threshold surface rather than assuming it—a more data- intensive process, to be sure, but one that is also more nuanced. Boolean logit models may take a functional form that imply threshold effects, or they may not; when they do, calculating the threshold is straightforward (see Braumoeller and Carson [2011] for an example).

Form of Interaction. In all three of the statistical models the form of the interaction is specified *a priori* based on theory rather than estimated. Different specifications can be compared to allow the data to influence which interactive model is chosen, but the form of interaction must be specified prior to estimation. In fs/QCA, by contrast, the relevant combinations of conditions are derived inductively from the data. That is not to say that no assumptions are involved: indeed, a key assumption in the minimization procedure is that the minimum of the two values represents the observation's joint membership in the combined set. Concretely, that means that if the combination of $X1$ and $X2$ produces Y, we should see no observations above the line $\min(X1, X2)=Y$. Interactions are estimated, therefore, by leveraging strong assumptions about interaction.

Critiques. Finally, it can be useful to explore critiques of each technique.[7] Hug (2013) and Seawright (2013) focus on the restrictive assumptions of QCA and fs/QCA with regard to measurement, thresholds, and so on, and they highlight the sensitivity of the results to violations of those assumptions. This is a very real concern, especially when a single observation can make or break a hypothesis. The fundamental problem, of course, is that the information necessary to relax these assumptions—a large enough N to estimate latent variables or coefficients or thresholds—is very hard to come by, which is why the assumptions are made in the first place. The best that can be done is to engage in sensitivity tests that vary the assumptions and assess the robustness of the conclusions. This is a worthwhile exercise, though the sheer number of assumptions makes it a daunting prospect.

[5] Jack Paine (personal communication) suggests that interaction terms might be capable of capturing more complex interactions—for example, a three-way interaction term would capture the argument that $X1$ and ($X2$ or $X3$) produces Y if the coefficients on $X1X2$ and $X1X3$ are positive but the coefficient on $X1$ is zero. While the point about the malleability of the functional form is correct and, to my knowledge, entirely original, the implications for hypothesis- testing are not as straightforward—in particular, $X1=0$ is the null hypothesis, and standard significance tests are biased *toward* failing to reject it.

[6] There is one special case in which this statement is not exactly true. When Boolean logit models capture purely conjunctural causation, the predicted probabilities of the constituent logit equations can be interpreted as upper bounds on the probability that $Y=1$. Similarly, when Boolean logit models capture substitutability, or purely disjunctural causation, the predicted probabilities of the constituent logit equations can be interpreted as lower bounds on the probability that $Y=1$.

[7] For the purposes of this article a few recent examples will suffice; this is far from a complete survey.

Table 1: Characteristics of Four Models of Interaction

	fs/QCA	Interaction Terms	Stochastic Frontier Model	Boolean Logit
Measurement	Latent; theorized	Varies	Varies	Latent; estimated
Coefficients	Assumed	Estimated	Estimated	Estimated
Thresholds	Fixed at *X*=*Y*	Not estimated	Estimated	Sometimes estimated
Form of Interaction	Combinations of conditions	Multiplicative	Multiplicative	Multiplicative
Information needed	Very little	Little	Intermediate	A lot

While few recent articles critique interaction terms,[8] a recent memo written by Charles Ragin (2013) does suggest that the sorts of causal interactions captured by QCA cannot be captured by interaction terms. Ragin offers the following dataset summary as an example:

x1	x2	y=0	y=1	proportion	QCA code
0	0	10	10	0.500	0
0	1	10	20	0.667	0
1	0	10	25	0.714	0
1	1	10	90	0.900	1

According to Ragin, the much higher proportion of *y=1* cases when *x1=x2*=1 is clear evidence of conjunctural causation: "The recipe is clear: when both x1 and x2 are present, outcome y=1 is highly consistent." Yet despite the relatively large number of observations (adding up the third and fourth columns produces *N*=185), a statistical model with interaction terms fails to capture the interaction between the two independent variables.

This critique is thought-provoking because it reflects thoroughly different understandings of interaction. From a QCA perspective, the *x1/x2* combination is the only one that produces *y=1* with very high probability; the combination of the two is therefore close enough to sufficient (in the fs/QCA context) to warrant the conclusion that the relationship is interactive. From a regression perspective, the co-occurrence of *x1* and *x2* actually does not produce *y=1* with that much greater frequency than the sum of their separate occurrences would suggest. Relative to the first row, the second row represents an increase of 0.167 in the proportion of *y=1* cases and the third row represents an increase of 0.214 in the proportion of *y=1* cases. In a purely additive statistical model, we would expect the fourth row to represent an increase of 0.167 + 0.214 = 0.381 in the proportion of *y=1* cases. The increase that it does represent, 0.400, is not all that different from 0.381, and as a result a regression model does not suggest an interactive relationship. Ragin is correct, therefore, in suggesting that interaction terms do not capture the sort of interactions posited by QCA.

The flipside of this point, however, is considerably more surprising: in one significant regard, *QCA is not a methodology designed to capture interactions among variables*. QCA focuses on combinations of conditions that create thresholds. Interactions occur when a change in the value of one condition alters the impact of another condition on the outcome. As Ragin's example demonstrates, it is entirely possible for the combination of two conditions to constitute a threshold when neither condition alters the impact of the other on the outcome. The difference between finding combinations of conditions, on the one hand, and demonstrating that the impact of the combination of those conditions is greater than the impact of the sum of the parts, on the other, is a subtle one, but it is very important: QCA does the former but not the latter. QCA and interaction terms, in other words, do not do the same thing.

While QCA is not designed to tell you whether its conditions interact with one another, it *is* designed to tell you whether combinations of conditions interact with all of the other potential conditions out there—that is, whether conditions combine to form thresholds.

Stochastic frontier models and Boolean models have been applied less often in political science than either of the first two methods, so critiques are harder to come by. The most damning critique of the stochastic frontier model is its rigidity: it serves very well as a means of modeling a production function, but its assumptions—zero production in the absence of any factor, a specific form of interaction, a dual error term with rigid distributional assumptions—may or may not suit other

[8] For a recent exception see Berry, DeMeritt and Esaray (2010), which argues that interaction terms are often unnecessary in logit and probit models because the functional form of the model induces interactivity. This is not really a critique of interaction terms *per se*, however.

applications. As to Boolean logit, my own experience is that its information requirements and the convoluted likelihood functions that it produces are its Achilles' heel: the former can result in inestimable models, while the latter result in frequent violations of the Wald assumptions that underpin estimated standard errors. The remedy in the first case is more data. The remedy in the second is bootstrapped standard errors, which are time-consuming but produce the correct standard error estimates.

All in all, the main critiques of these models, unsurprisingly, revolve around their position on the spectrum from high-assumption to high-information. In the case of fs/QCA, if the many assumptions fit, you can believe the results; in the case of Boolean logit, if you have enough information you can estimate the model. The two models in between, while less ambitious in terms of their ability to model causal complexity, are also less ambitious in their requirements.

Conclusion: Caveat Emptor

Taken as a whole, to reiterate, these models represent points on a spectrum, from the assumption-laden fs/QCA procedure to the data-hungry Boolean logit. The implications of data-intensivity are fairly straightforward: a data-hungry procedure runs the risk of providing null results if too few data are available to estimate all of the necessary quantities. What are the implications of incorrect assumptions?

In the case of the three statistical procedures, the main assumption has to do with the form of the interaction. A bad assumption at this stage will, in a nutshell, produce conclusions that are inaccurate to an unknowable degree—and that is every bit as bad as it sounds.

In fs/QCA, because conclusions depend on a wider range of assumptions, the cumulative implications of violating those assumptions can be even more dire. If fuzzy-set membership is estimated improperly, if the mean rather than the minimum defines joint membership in the conjunction of two sets, if the true threshold between possible and impossible cases is really $Y=X^2$, and if an independent variable's contribution to an outcome is partial (or, worse, unconditional), the results can bear shockingly little resemblance to the reality they are meant to capture.

In all cases it pays to question assumptions and to do so thoroughly. For statistical models, it is at least possible to use model fit to adjudicate among competing assumptions. We may never arrive at the One True Specification, but we can at least know which is the best given the data we have at hand. fs/QCA offers fewer assurances of this nature, but practitioners can at least get a sense of the range of possible conclusions by varying the assumptions at each step and exploring the extent to which the results are robust to those changes.

References

Berry, William D., Jacqueline H. R. DeMeritt, and Justin Esarey. 2010. "Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?" *American Journal of Political Science* 54 (1): 248–266.

Braumoeller, Bear F. 2003. "Causal Complexity and the Study of Politics." *Political Analysis* 11 (3): 209–233.

Braumoeller, Bear F. and Austin Carson. 2011. "Political Irrelevance, Democracy, and the Limits of Militarized Conflict." *Journal of Conflict Resolution* 55 (2): 292–320.

Braumoeller, Bear F. and Gary Goertz. 2000. "The Methodology of Necessary Conditions." *American Journal of Political Science* 44 (4): 844–858.

Clark, William Roberts, Michael J. Gilligan, and Matt Golder. 2006. "A Simple Multivariate Test for Asymmetric Hypotheses." *Political Analysis* 14 (3): 311–331.

Dion, Douglas. 1998. "Evidence and Inference in the Comparative Case Study." *Comparative Politics* 30 (2): 127–145.

Hildebrand, David K., James D. Liang, and Howard Rosenthal. 1976. "Prediction Analysis in Political Research." *American Political Science Review* 70 (2): 509–535.

Hug, Simon. 2013. "Qualitative Comparative Analysis: How Inductive Use and Measurement Error Lead to Problematic Inference." *Political Analysis* 21 (2): 252–265.

Ragin, Charles C. 2013. "QCA versus Statistical Interaction." Department of Sociology, University of California, Irvine.

Seawright, Jason. 2013. "Warrantable and Unwarranted Methods: The Case of QCA." Presented at the Annual Meeting of the American Political Science Association, Chicago.

Process Tracing with Bayes: Moving Beyond the Criteria of Necessity and Sufficiency

Andrew Bennett
Georgetown University
bennetta@georgetown.edu

Social scientists have long recognized the study of evidence from within individual cases as a fundamental tool for causal inference. This evidence helps guard against the inferential errors that can arise from making causal inferences based only on comparisons among cases. Process tracing, the systematic study of evidence from within a single case to assess alternative explanations of that case, is a key method of within-case analysis.

Yet until recently, formal articulation of the underlying logic of process tracing has been incomplete. One line of inquiry has sought to organize the traditional process tracing tests in terms of whether they provide necessary and/or sufficient grounds for inferring that a given piece of evidence confirms a particular hypothesis (Bennett 2010; Collier 2011). Thus, (1) the results of a straw in the wind test may provide suggestive, but far from definitive, support for the hypothesis; (2) the hoop test must be passed for the hypothesis to be seriously

Author's Note: This is an abridged and revised version of "Disciplining our Conjectures: Systematizing Process Tracing with Bayesian Analysis," the technical appendix to Andrew Bennett and Jeffrey Checkel, eds., *Process Tracing: From Metaphor to Analytic Tool* (Cambridge University Press, forthcoming 2014). I would like to thank Derek Beach, Jeff Checkel, David Collier, Colin Elman, Dimitri Gallow, Macartan Humphreys, Alan Jacobs, James Mahoney, Ingo Rohlfing, and David Waldner for their insightful comments on an earlier draft of this paper. Any remaining errors are my own.

entertained, and passing is therefore necessary for sustaining the hypothesis; (3) the smoking gun test affirms the hypothesis and passing is therefore sufficient for sustaining the hypothesis, although it does not exclude other explanations, and (4) a doubly decisive test affirms the hypothesis and excludes other explanations. This framework draws informally on Bayesian logic, but strictly speaking, Bayesianism requires that we never be one hundred percent convinced of the truth or falsity of any explanation, so the terms "necessary" and "sufficient" are too categorical.

Bayesianism is not the only way to understand process tracing. James Mahoney (2012) has demonstrated that set theory can be used to arrive at many of the same insights regarding process tracing, provided that no evidence is considered fully necessary or sufficient to judge explanations as either certain or impossible. Similarly, David Waldner (forthcoming) has argued that directed acyclic graphs are a useful way to think about process tracing.[1] Process tracing has been most fully explicated in terms of Bayesianism, however, and the following discussion continues this approach.[2] It concludes that using Bayesian logic more fully, systematically, and transparently can improve the quality and replicability of process tracing and strengthen causal inferences, including those based on qualitative and quantitative cross-case comparisons.[3]

Fundamentals of Bayesian Analysis

Process tracing analyzes within-case evidence to develop or test explanations of individual cases. Doctors diagnosing patients, detectives investigating crimes, and social scientists developing and testing both general theories and historical explanations of particular cases are all interested in how we should update theories and explanations in the light of evidence from individual cases. One of the most powerful ways of thinking about this challenge is the logic first systematized by Thomas Bayes in the mid-1700s. Bayes focused on the question of how we should update our confidence in an explanation given new, relevant evidence. This updated confidence in the likely truth of a theory is referred to as the posterior, or the likelihood of a theory conditional on the evidence.

In Bayes's approach, we need three key pieces of information, in addition to the evidence itself, to calculate this posterior likelihood. First, we need to start with a "prior" likelihood, which expresses our initial confidence that a theory is true even before looking at the new evidence. For example, let us assume we have an explanation of a case that we think is 40 percent likely to be true, and for simplicity let us further assume that it is mutually exclusive with the alternative explanations—that is, only one could be true—so the likelihood it is false is one minus 40 percent, or 60 percent.[4]

Second, we need information on the likelihood that, if a theory is true in a given case, we will find a particular kind of evidence for that case. This is referred to as the evidence conditional on the theory. We can view the theory as an attempt to capture the underlying "data generating process," and hence as a useful way to understand the claim that the evidence is conditional on the theory. Let us assign this a likelihood of 20 percent to illustrate a "smoking gun" test. This is a test in which confirmatory evidence, if found, strongly increases our confidence in the explanation, but the failure to find that evidence does not strongly undermine our confidence in the explanation.

Third, we need to know the likelihood that we would find the same evidence even if the explanation of interest is false—i.e., a false positive. In our example, to complete the logic of a smoking gun test, let us assign this a probability of 5 percent.[5]

Smoking Gun Test

Analysis of the three estimated probabilities necessary for Bayesian updating of our explanation can be illustrated with the smoking gun test. Using P for the explanation, *pr*(P) for the prior probability that P is true, and k for the evidence, we have:

[1] It is not yet clear whether there are methodologically consequential differences among Bayesianism, set theory, flow graphs, and directed acyclic graphs with regard to process tracing. There are many ways in which these three logics are compatible and translatable; on this point, see Zdislaw Pawlak, "Bayes' Theorem—the Rough Set Perspective," at http://bcpw.bg.pw.edu.pl/Content/1935/btrsp_or.pdf, accessed May 1, 2014, and Abell (2009: 45–58).

[2] Bennett (2008); Abell (2009); Beach and Pedersen (2013a, 2013b); Collier (2011); Humphreys and Jacobs (2013); Mahoney (2012); and Rohlfing (2012, 2013a, 2013b).

[3] This point has often been made with regard to combining statistical analysis and within-case analysis. However, Bayesian analysis can also strengthen qualitative methods of cross-case comparisons, whether typological theory (George and Bennett, 2005) or Qualitative Comparative Analysis (QCA) (Ragin 2008). These qualitative methods are greatly strengthened by combining them with process tracing. On the latter point, see Schneider and Rohlfing (2013).

[4] One complication is that theories or explanations may not be mutually exclusive, but rather complementary. If I sneeze, for example, it may be due to allergies, to having a cold, to sudden exposure to bright lights, or to a combination of any two or all three factors; thus, showing that there was exposure to bright light does not necessarily raise or lower the likelihood that having a cold or allergies contributed to my sneezing. The present discussion, like many pedagogical presentations of Bayesianism, simplifies this point by considering only whether one explanation is true or false, and assuming other theories are mutually exclusive, so the likelihood that the explanation is false is one minus the likelihood that it is true (see also Rohlfing 2012: chap. 8). In social science research, researchers often face the more complex question of hypotheses that, overall, are partly complementary and partly competing; or, alternatively, competing in the context of some cases and complementary in others (on this challenge see Rohlfing 2013a).

[5] Ideally estimates of priors and of the likelihood of finding evidence depending on whether a theory is true or false would be based on studies of many prior cases or well-validated theories or experiments. This is true in the medical research examples common in textbook discussions of Bayesianism. Unfortunately, in the social sciences we often lack such data and must begin with more subjective guesses on these probabilities. The reliance on subjective expectations of probabilities, and differences in individuals' estimates of these probabilities, is an important challenge for Bayesianism, although strongly probative evidence can lead to convergence between observers who start with greatly different assumptions on their priors.

Smoking Gun Test

Prior likelihood P is true, or $pr(\mathrm{P}) = .40$
Likelihood of smoking gun evidence k, if P is true = .20
Likelihood of smoking gun evidence k, if P is false = .05

We can now address the following question: if the evidence supporting the explanation is found, what is the updated likelihood that the explanation is true?

In a common form of Bayes' Theorem, the updated likelihood that a proposition P is true in light of evidence k, or $Pr(\mathrm{P}|\mathrm{k})$, is as follows:

$$Pr(\mathrm{P}|\mathrm{k}) = \frac{pr(\mathrm{P})pr(\mathrm{k}|\mathrm{P})}{pr(\mathrm{P})pr(\mathrm{k}|\mathrm{P}) + pr(\sim\mathrm{P})pr(\mathrm{k}|\sim\mathrm{P})} \qquad (1)$$

Notation:
$Pr(\mathrm{P}|\mathrm{k})$ is the posterior or updated likelihood of P given (i.e., conditional on) evidence k
$pr(\mathrm{P})$ is the prior likelihood that proposition P is true
$pr(\mathrm{k}|\mathrm{P})$ is the likelihood of evidence k if P is true (or conditional on P)
$pr(\sim\mathrm{P})$ is the prior likelihood that proposition P is false
$pr(\mathrm{k}|\sim\mathrm{P})$ is the likelihood of evidence k if proposition P is false (or conditional on ~P)

If we put our illustrative numbers into equation (1), the updated likelihood of the explanation being true is .73:

Likelihood the explanation is True for a Passed Smoking Gun Test

$$\frac{(.4)(.2)}{(.4)(.2) + (.6)(.05)} = \frac{.08}{.08 + .03} = \frac{.08}{.11} = .73 \qquad (2)$$

We can use Bayes' theorem to calculate the posterior likelihood of a failed smoking gun test to be .36. Hence, as the name of the test implies, passing the test raises the theory's likelihood far more (from .4 to .73) than failing it would lower this likelihood (from .4 to .36). This illustrates a key feature of Bayesianism. The extent of updating when a test result is positive is driven by the prior likelihood of the theory and the likelihood ratio, which is the ratio of true positives to false positives (Rohlfing 2013b).[6] Here, the likelihood ratio for positive evidence on the smoking gun test is:

$$\frac{\text{Likelihood of true positive}}{\text{Likelihood of false positive}} = \frac{.2}{.05} = 4 \qquad (3)$$

The higher the likelihood ratio (above a minimum value of 1) the more powerful or discriminating the evidence: finding positive evidence when the likelihood ratio was 4, as in the smoking gun test example, greatly increases the likelihood that the proposition is true.[7] When the likelihood ratio is equal to one, evidence has no discriminatory power: the posterior is the same as the prior.

Straw in the Wind, Hoop and Doubly Decisive Tests

The other three tests also exhibit continuous gradations in their strength. Hoop tests are the converse of smoking gun tests. In a hoop test, the absence of confirming evidence strongly undermines an explanation, but the presence of such evidence does not strongly increase the likelihood that the explanation is true. A straw in the wind test provides only weak evidence for or against an explanation. Finally, a doubly decisive test strongly increases the likelihood of an explanation that passes, and strongly undermines that of an explanation that fails.

Macartan Humphreys and Alan Jacobs (2013) have devised an excellent diagrammatic representation of how the likelihood ratio establishes the strength of these evidentiary tests. Figure 1 (adapted from Humphreys and Jacobs 2013:17), shows how these tests relate to the two measures that comprise the likelihood ratio: the likelihood of observing evidence k when a proposition P is true (labeled q_1 on the y-axis of the figure) and the likelihood of observing evidence k even when the proposition P is false (labeled q_0 on the x-axis of the figure).

The figure brings into sharp focus the mirror-image relations among tests depending on whether evidence k is present or absent. A test that provides smoking gun evidence for P when k is present constitutes hoop test evidence for ~P when k is absent, and vice-versa. Similarly, a hoop test for P is a smoking gun test for ~P. This is because P and ~P are inversely proportional—their probabilities add to one.

Humphreys and Jacobs also introduce a set of figures that further illustrate the properties of different evidentiary tests, again reproduced here as Figures 2 to 5.[8] These figures show how different prior probabilities map onto posterior probabilities for the illustrative likelihood ratio used in each graph. Examples are shown for likelihood ratios representing hoop, smoking gun, doubly decisive, and straw in the wind tests. Because q_0 and q_1 can vary continuously between zero and one, in addition to the examples in Figures 2 to 5, one could draw any number of curves for tests of different discriminatory power within each family of tests.[9]

These graphs nicely illustrate the point that the extent to which we should update our prior depends on the values of both the prior and the likelihood ratio. As Humphreys and Jacobs point out, we will not lose as much confidence in a hypothesis that has achieved a high prior through repeated earlier testing, even in the face of a failed hoop test. In Figure

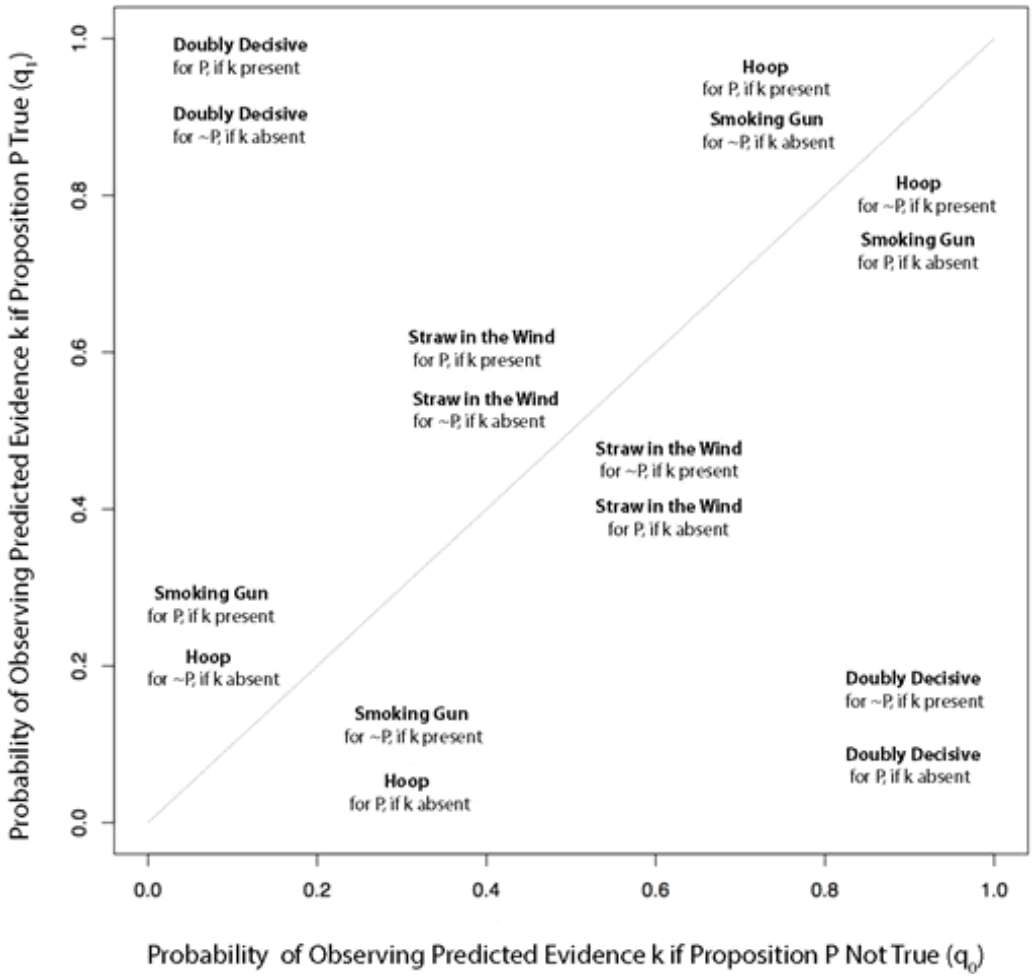
[6] For arguments that the likelihood ratio, or more specifically the log of the likelihood ratio, is the best measure of the evidential or confirmatory support of evidence, see Fitelson (2001) and Eels and Fitelson (2002).

[7] There is also a likelihood ratio with regard to a negative finding. This is the ratio of the likelihood of a false negative divided by that of a true negative. This ratio ranges from zero to one, and the closer it is to zero, the more powerful a negative finding is in undermining the likelihood that an explanation is true. The likelihood ratio for a positive finding is designated as LR+, while that for a negative finding is designated LR-. For present purposes, I use the term "likelihood ratio" to refer to LR+.

[8] Humphreys and Jacobs (2013: 19); see also Rohlfing (2013b: 20–29).

[9] On this point, see Mahoney (2012) and Rohlfing (2013a).

Figure 1: Mapping Process Tracing Tests Based on the Likelihood Ratio



Note: Figure adapted from Humphreys and Jacobs (2013: 17), with permission of the authors. P = Proposition being tested. k = Evidence evaluated to carry out test. q_0 and q_1 = Probability of finding evidence k, according to falsity or truth of proposition P

3 the vertical distance from the 45-degree diagonal to the curved line for the failed hoop test, which shows how much lower the posterior is than the prior, is less when the prior is close to 1.0 than when it is when the prior is between 0.4 and 0.8.

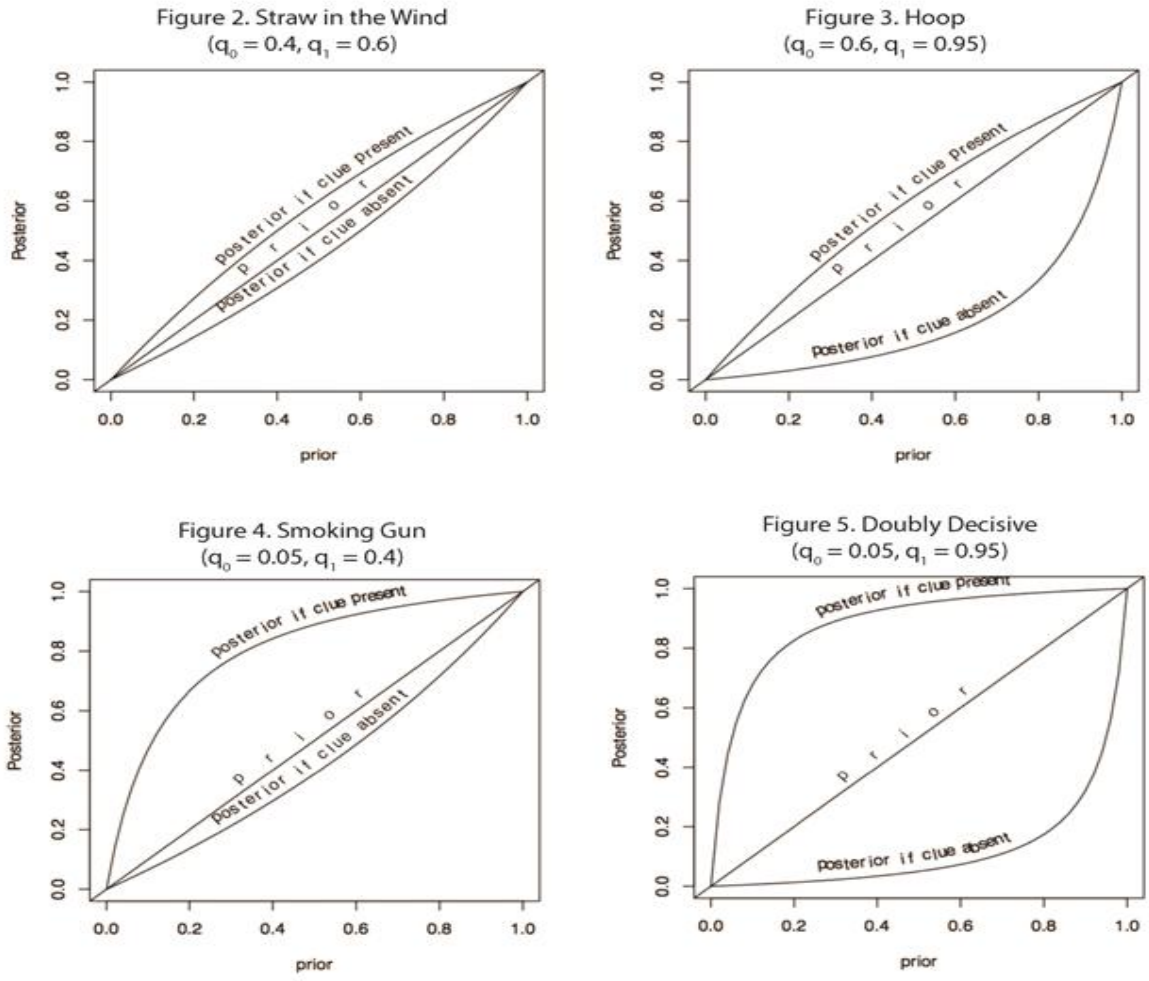
The mathematical relationships among q_0, q_1, the prior, and the posterior allow us to test for consistency. Given any three of these likelihoods, we can determine what the value of the fourth must be if our thinking is consistent. Alternatively, given any two, we can determine what the ratio of the other two should be. For example, given an individual's prior and posterior, we can determine what their likelihood ratio should have been for the evidence they examined. In cases where a scholar has a prior of 40 percent and a posterior of 95 percent, we know that their likelihood ratio for the evidence they examined should have been just over 28. That is, they should have been 28 times as likely to expect the evidence if the theory was true than if it was false, which is an extremely high ratio. If the scholar did not think this likelihood ratio was justified, they might have to lower their estimate of the posterior.

The Implications of Bayesianism for Process Tracing

I explore elsewhere how Bayesian logic reveals a number of implications for process tracing (Bennett 2008; Bennett forthcoming 2014); here, I focus on five.

First, the explication of Baysianism above improves upon my earlier writings on the subject. Earlier, I infelicitously suggested that it was *necessary* for an explanation to pass a hoop test in order to remain viable, whereas passing a smoking gun test was *sufficient* to confirm an explanation (Bennett 2010). This language is misleading, in that Bayesianism reminds us that we can never be 100 percent confident that an explanation is true, or that it is false. There are several reasons for this. First, there may be alternative explanations that fit the evidence better. Second, there is always some evidence that is inaccessible. Third, there may be errors in the measurement of the evidence. More generally, we cannot tell for certain if a theory's failure in an evidentiary test undermines the theory or if it undermines auxiliary hypotheses, explicit or implicit in the theory, about the observation and measurement of evidence.

Figures 2 to 5: Illustrative Examples of the Four Tests
(Adapted from Humphreys and Jacobs, 2013: 19.)



Thus, although some pieces of evidence may be highly probative, we cannot infer with certainty that a theory or explanation is true based on the evidence.

Second, counter-intuitively, evidence consistent with a theory can actually lower its posterior because this same evidence is even more consistent with an alternative theory. Conversely, evidence that does not fit a theory can actually raise its posterior by more severely undermining an alternative explanation. These outcomes happen when the likelihood ratio is less than one.[10] Figures 2 to 5 all have likelihood ratios where q_1 is greater than q_0; that is, they are all drawn from above the 45-degree diagonal in Figure 1. When q_0 is greater than q_1, the likelihood ratio is less than one (as in the area below the 45-degree diagonal of Figure 1), and evidence consistent with P actually reduces the likelihood that P is true.[11]

Third, Bayesianism provides a logical rationale for the methodological prescription that independence and diversity of evidence is important in process tracing. Desirable evidentiary tests are those that are independent of one another, and diverse—i.e. they bear on different alternative hypotheses. Regarding independence, if one piece of evidence is wholly determined by another, it has zero additional power to update prior probabilities. As for diversity of evidence, as we accumulate more and more pieces of evidence that bear on only one alternative explanation, each new piece has less power to update further our confidence in that explanation. This is true, even if the evidentiary tests are independent, because we have already incorporated the information of the earlier, similar evidence.

Fourth, multiple weak tests, if independent from one another, can sometimes cumulate to strongly update priors. Straw in the wind tests, and weak smoking gun and hoop tests, are the kinds of tests that might be called "circumstantial evi-

[10] See also Rohlfing (2013b: 5, 19, 20).

[11] In medical tests, the positive likelihood ratio as discussed in footnote 7 above, LR+, is simply *defined* as the test result that makes it more likely a patient has a particular disease. If a doctor thought a certain test result was likely to be associated with the disease, but found the opposite to be true, she or he would simply flip the interpretation of what reading on the test constituted a "positive" outcome. With the testing of posited social mechanisms, however, social scientists do not necessarily flip the interpretation of what it means to find that the hypothesized evidence of the mechanism was observed.

dence" in a court case. If most of these kinds of tests point in the same direction, this provides strong evidence for the explanation in question. This is analogous to the high likelihood that a coin is biased toward heads if it comes up heads significantly more than 50 percent of the time in a large number of fair coin tosses.

The final implication points to a crucial choice: whether to "fill in the numbers" by explicitly assigning priors and likelihood ratios and using Bayesian mathematics, at least for the few pieces of evidence that a researcher considers the most probative, in an effort to make process tracing more rigorous and transparent. Earlier discussions treated Bayesianism as a useful metaphor for process tracing (McKeown 1999) or a way of clarifying its logic (Bennett 2008), without arguing that Bayesian mathematics should be used explicitly in process tracing. Other researchers also argue that more explicit use of Bayesian mathematics in process tracing is impractical and would convey a false sense of precision (Beach and Pedersen 2013a).[12] More recently, however, a number of scholars (Abell 2009; Humphreys and Jacobs 2013; Rohlfing 2013b) have suggested that researchers should in fact implement Bayesianism more concretely, explicitly identifying their priors and likelihood ratios and using Bayes' theorem to determine posterior probabilities.

A powerful argument for actually filling in the numbers in process tracing is that it asks researchers to make specific and transparent the assumptions that they must in any case make implicitly if process tracing is to have probative value. The process of clearly identifying the likelihood of finding a certain kind of evidence, not only conditional on the truth of a theory but also conditional on the falsity of the theory, can push researchers to clarify their own thinking. It also makes this thinking more transparent to other scholars, eliminating the considerable ambiguity in many verbal formulations used to convey the likelihoods of explanations and evidence.

We have good examples of process tracing in which scholars have been exceptionally careful and explicit in the evidence they used and the type of tests (e.g. hoop tests, smoking gun tests) they applied in making inferences (Fairfield 2013). So far, however, we have no full-fledged examples where scholars have done process tracing with explicit priors and numerical Bayesian updating; this remains an area where the advice of at least some methodologists diverges from the practices of working researchers.[13] Whether one ultimately prefers to use Bayesian logic implicitly or explicitly, understanding this logic unquestionably helps clarify the logic of process tracing.

References

Abell, Peter. 2009. "A Case for Cases: Comparative Narratives in Sociological Research." *Sociological Methods and Research* 38 (1): 38–70.

Beach, Derek and Rasmus Brun Pedersen. 2013a. *Process Tracing Methods: Foundations and Guidelines*. Ann Arbor: University of Michigan Press.

Beach, Derek and Rasmus Brun Pedersen. 2013b. "Turning Observations into Evidence: Using Bayesian Logic to Evaluate What Inferences are Possible Without Evidence." Presented at the Annual Meeting of the American Political Science Association, Chicago.

Bennett, Andrew. 2008. "Process Tracing: A Bayesian Perspective." In *The Oxford Handbook of Political Methodology*, eds. Janet Box-Steffensmeier, Henry E. Brady, and David Collier. (Oxford: Oxford University Press), 702–721.

Bennett, Andrew. 2010. "Process Tracing and Causal Inference." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 2nd ed., eds. Henry E. Brady and David Collier. Lanham, MD: Rowman & Littlefield, 207–219.

Bennett, Andrew and Jeffrey Checkel, eds. Forthcoming 2014. *Process Tracing: From Metaphor to Analytic Tool*. Cambridge: Cambridge University Press.

Collier, David. 2011. "Understanding Process Tracing." *PS: Political Science and Politics* 44 (4): 823–830.

Eels, Ellery, and Branden Fitelson. 2002. "Symmetries and Asymmetries in Evidential Support." *Philosophical Studies* 107 (2): 129–142.

Fairfield, Tasha. 2013. "Going Where the Money Is: Strategies for Taxing Economic Elites in Unequal Democracies." *World Development* 47: 42–57.

Fitelson, Branden. 2001. "A Bayesian Account of Independent Evidence with Applications." *Philosophy of Science* 68 (3): S123–S140.

George, Alexander L. and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge: MIT University Press.

Humphreys, Macartan, and Alan Jacobs. 2013. "Mixing Methods: A Bayesian Unification of Qualitative and Quantitative Approaches." Presented at the Annual Meeting of the American Political Science Association, Chicago.

Mahoney, James. 2012. "The Logic of Process Tracing Tests in the Social Sciences." *Sociological Methods and Research* 41 (4): 570–597.

McKeown, Timothy. 1999. "Case Studies and the Statistical World View." *International Organization* 53 (1): 161–190.

Pawlak, Zdislaw. "Bayes' Theorem—the Rough Set Perspective." http://bcpw.bg.pw.edu.pl/Content/1935/btrsp_or.pdf.

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

Rohlfing, Ingo. 2012. *Case Studies and Causal Inference: An Integrative Framework*. New York: Palgrave Macmillan.

Rohlfing, Ingo. 2013a. "Comparative Hypothesis Testing via Process Tracing." *Sociological Methods and Research*. Advance online publication. doi: 10.1177/0049124113503142

Rohlfing, Ingo. 2013b. "Bayesian Causal Inference in Process Tracing: The Importance of Being Probably Wrong." Presented at the Annual Meeting of the American Political Science Association, Chicago.

Schneider, Carsten and Ingo Rohlfing. 2013. "Combining QCA and Process Tracing in Set-Theoretic Multi-Method Research." *Sociological Methods and Research* 42 (4): 559–597.

Waldner, David. Forthcoming 2014. "What Makes Process Tracing Good? Causal Mechanisms, Causal Inference, and the Completeness Standard in Comparative Politics." In *Process Tracing*, eds. Andrew Bennett and Jeffrey Checkel. (Cambridge: Cambridge University Press).

[12] But see Beach and Pedersen (2013b) which urges more explicit and transparent use of Bayesian logic, if not specific use of mathematical probability estimates.

[13] Abell (2009: 59–61) provides a brief illustrative example of explicit Bayesian updating in process tracing. In this example, he uses a panel of trained researchers, rather than an individual researcher, to estimate likelihood ratios based on shared evidence from the case.