Hug, Simon. 2010. "The Effect of Misclassifications in Probit Models. Monte Carlo Simulations and Applications." *Political Analysis* 18(1): 78–102.

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

Krogslund, Chris, Donghyun Danny Choi, and Mathias Poertner. Forthcoming. "Fuzzy Sets on Shaky Ground: Parametric and Specification Sensitivity in fsQCA." *Political Analysis*.

Krogslund, Chris and Katherine E. Michel. 2014a. "Can QCA Do Causal Inference? An Assessment and Proposed Alternative." Paper prepared for presentation at the Annual Meeting of the American Political Science Association, Washington DC.

Krogslund, Chris and Katherin E. Michel. 2014b. "A Larger-N, Fewer Variables Problem? The Counterintuitive Sensitivity of QCA." *Qualitative & Multi-Method Research* 12(1): 25–33.

Forrester, Jay Wright. 1969. *Principles of Systems*. Cambridge, MA: Wright-Allen Press.

Lieberson, Stanley. 1991. "Small N's and Big Conclusions: An Examination of the Reasoning in Comparative Studies Based on a Small Number of Cases." *Social Forces* 70(2): 307–320.

Lieberson, Stanley. 2004. "Comments on the Use and Utility of QCA." *Qualitative Methods* 2(2): 13–14.

Lucas, Samuel R. and Alisa Szatrowski. 2014. "Qualitative Comparative Analysis in Critical Perspective." *Sociological Methodology* 44(1): 1–79.

Mahoney, James. 2004. "Reflections on Fuzzy-Set/QCA." *Qualitative Methods* 2(2): 17–21.

Michel, Katherine E. and Chris Krogslund. 2014. "Testing the Ability of Set-Theoretic Comparative Methods to Recover Data Generating Processes." Paper prepared for presentation at the Annual Meeting of the Midwest Political Science Association, Chicago.

Mill, John Stuart. 1843/1973. *Of the Four Methods of Experimental Inquiry. In A System of Logic, Ratiocinative and Inductive. Collected Works of John Stuart Mill* (Volume VII). John Stuart Mill, ed. (Toronto: University of Toronto Press), 388–406.

Paine, Jack. Forthcoming. "Set-Theoretic Comparative Methods (STCM): Less Distinctive than Claimed." *Comparative Political Studies*.

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

Ragin, Charles. 2000. *Fuzzy-Set Social Science.* Berkeley: University of California Press.

Ragin, Charles C. 2005. "Core Versus Tangential Assumptions in Comparative Research." *Studies in Comparative International Development* 40: 33–38.

Ragin, Charles C. 2014. "Comment: Lucas and Szatrowski in Critical Perspective." *Sociological Methodology* 44: 80–94.

Ragin, Charles C. and Benoît Rihoux. 2004. "Replies to Commentators: Reassurances and Rebuttals." *Qualitative Methods* 2(2): 22–24.

Rihoux, Benoît. 2003. "Bridging the Gap Between the Qualitative and Quantitative Worlds? A Retrospective and Prospective View on Qualitative Comparative Analysis." *Field Methods* 15(4): 351–365.

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.

Seawright, Jason. 2005. "Qualitative Comparative Analysis vis-a-vis Regression." *Studies in Comparative International Development* 40(1): 3–26.

Seawright, Jason. 2013. "Warrantable and Unwarrrantable Methods: The Case of QCA." Prepared for presentation at the Annual Meeting of the American Political Science Association, Chicago, August 2013.

Seawright, Jason. 2014. "Limited Diversity and the Unreliability of QCA." *Sociological Methodology* 44(1): 118–121.

Tanner, Sean. 2014. "QCA and Causal Inference: A Poor Match for Public Policy Research." *Qualitative & Multi-Method Research* 12(1): 15–25.

Thiem, Alrik. 2014. "Mill's Methods, Induction and Case Sensitivity in Qualitative Comparative Analysis: A Comment on Hug (2013)." *Qualitative & Multi-Method Research* 12(2): 19–24.

Clarifying Misunderstandings, Moving Forward: Towards Standards and Tools for Set-Theoretic Methods

https://doi.org/10.5281/zenodo.894398

Ingo Rohlfing
University of Bremen, Germany
rohlfing@bigsss.uni-bremen.de

Carsten Q. Schneider
Central European University, Hungary
schneiderc@ceu.hu

Critical debates can push boundaries if they are open-ended in nature. In this spirit, we appreciate David Collier's efforts in putting together a set of articles about set theory and QCA in a symposium published in the previous issue of this newsletter. Many important issues about the principles and current practices of set-theoretic methods are raised and at least some contributors seem open to the possibility that set theory and QCA are worthy of being pursued and improved. We are grateful to the editor of the newsletter, Robert Adcock, for inviting us to respond to these arguments—an invitation we accept based on the expectation that the goal of everyone involved in this exchange is to strengthen set-theoretic methods rather than prematurely dismissing them based on what we find to be shaky arguments. We further believe that this can only be the start of a larger, open debate aimed at resolving misunderstandings and enhancing set-theoretic research.[1]

Since we cannot address every aspect of the contributions to the previous issue (henceforth only 'contributions'), we focus on what we perceive to be the most salient topics for the current assessment and future development of set-theoretic methods. We first outline points of agreement and briefly indicate what we see as the major points of disagreement. The disagreements are then discussed in more detail in the following sections. Our major response is that several contributions identify important topics and problems, but they take their point too far or even in the wrong direction. Some contrib-

[1] *Editor's Note*. This exchange will continue in the next issue of the QMMR newsletter, when David Collier will respond to Rohlfing and Schneider's response to the newsletter symposium he organized. The authors would like to thank Benoît Rihoux and Alrik Thiem for comments on an earlier version of this text.

utions misrepresent how QCA works and overstate the problem, or they incorrectly infer flaws in the principles of QCA from problems in its current practice in empirical research. What we propose, instead, is to take the current pitfalls as challenges and develop tools for handling them.

Points of Agreement

Getting fundamentals of causal analysis right

Several contributions criticize the capacity of QCA for drawing causal inference.[2] We agree that an intelligible application of methods and the generation of causal inferences require an explicit statement as to what is meant by 'causal inference.' In the symposium, Bennett, Tanner, and Braumoeller discuss this issue from two different angles. Bennett makes a plea for *Bayesian inference* and argues that it is superior to the necessary and sufficient criteria for hypothesis tests he himself proposed just a few years ago (2010). We do not dispute the idea of using new evidence for Bayesian updating of the confidence in a hypothesis, which is a central topic in the current process tracing literature (Beach and Pedersen 2013; Bennett 2008; Rohlfing 2012: chap. 8). We think it is important to note, however, that Bayesian inference and set-theoretic research are not contradictory and can be fruitfully *combined*.

Tanner highlights the importance of *identification* in causal analysis and presents fine pieces of non-QCA work from policy analysis that aim at identifying the causal effect of a treatment. We follow Tanner's distinction between identification and estimation and agree that the QCA literature so far has mostly been concerned with estimation (i.e., the algorithm, handling of truth table rows without cases etc.), but hardly ever with identification. Yet neither the current neglect of identification issues nor Tanner's critical discussion of five QCA studies from policy analysis provide reasonable grounds to dismiss QCA as an "unsuitable method" (2014: 24).

Along similar lines, we appreciate the astute discussion of *different forms of interactions* by Braumoeller. His contribution accepts set-theoretic intersections as a meaningful and viable form of "interaction." Moreover, it also argues forcefully against those who claim that there is no difference between a set-theoretic intersection and an interaction term in regression models (e.g., Clark, Gilligan and Golder 2006; Brady 2013). Because of this, we agree that intersections cannot be easily modeled with existing quantitative tools, such as regression models with interaction terms (Goertz and Mahoney 2013).

On the subject of estimation, we concur with Krogslund and Michel, and Collier that the results produced by any method should be subject to *robustness tests*. Because we usually confront a number of design and modeling decisions that are equally plausible, we need to know whether our results are sensitive to making a decision in one direction or the other. We agree that the QCA literature has started to provide some guidance in making these choices (Schneider and Wagemann 2012: chap. 11.2; Skaaning 2011), but robustness tests still remain the exception rather than the rule in applied research. Yet, this again (see Tanner) is a problem of practice and not of principles. We are confident this practice will change in the future because of an increasing focus on this topic and the availability of easy-to-handle robustness routines in *R* (Ambuehl, Baumgartner, Epple, Kauffmann and Thiem 2014; Medzihorsky and Quaranta 2014; Thiem and Duşa 2012). Yet, contrary to Krogslund and Michel, we are not at all convinced that the evidence provided so far (both by them and others) allows for a robust conclusion that QCA "[…] has major vulnerabilities" (2014: 31).

Linguistics is not our area of expertise, but we are inclined to follow Lakoff's claim that set theory does not correspond with *human cognition and natural language*. Lakoff argues that natural language is often best described by family resemblance concepts in which objects are ranked according to their similarity. In contrast, classic concepts order cases according to their degree of membership in a set (or type) as, for example, discussed by Sartori (1970) and Collier and Mahon (1993). While we do not challenge these insights on natural language, we do question the relevance of this point to the debate on the merits of set-theoretic methods. None of them depends on being in accord with natural language. The key issue is whether *social science concepts* and theories can be framed in terms of sets and their subset relations. We believe that some can while others probably cannot. For those concepts and theories that are set-theoretic in their nature, set-theoretic methods are of course appropriate.

We welcome Sartori's insistence on logic being an important foundation for social science methods and research. We also share his concern that "excessive reliance" on specific methods is usually a vice and not a virtue for a discipline. However, we think everyone agrees that this argument is not confined to the excessive use of logic but extends to the excessive use of any kind of formalized system. Furthermore, we part ways with Sartori in his reading of Goertz and Mahoney's *A Tale of Two Cultures* which, in our view, is about common practice of methods application, not best practices (2012: chap. 1).[3]

Putting case knowledge upfront

One of the core themes highlighted in David Collier's introduction (although barely so in the contributions) is that set-theoretic research takes its strength from the *close engagement with cases*. We entirely agree that case knowledge is an important asset in empirical research *regardless* of the method that is used (Freedman 1991). Nobody seriously questions that, in QCA, it is fundamental to engage with cases.[4] However, we are

[2] The symposium introduces the term set-theoretic comparative methods (STCM). We do not use this label because we find it misleading. For one, when the symposium explicitly criticizes a method, it is only about QCA. All other criticism is directed at set theory as the foundation of QCA. We therefore distinguish between QCA as a method and approach on the one hand, and set theory as the basis of it, on the other.

[3] Goertz and Mahoney might not maintain this distinction on all pages, but this is the explicitly stated aim of the book.

[4] The foundational book by Ragin (1987) leaves no doubt about the key role of case knowledge. In order to make clear that QCA is not just the application of algorithms to data, several authors emphasize

less skeptical with regard to the use of QCA for data analysis without deep case knowledge; it is "just" that a QCA based on thorough case knowledge leads to more credible inferences than the same QCA that invokes only superficial case insights. In any case, the neglect of case knowledge in many applied QCA studies is no indication for a foundational flaw of this method. As holds for all methods, QCA as a technique cannot be assessed by how it is implemented.

Improving calibration

Set calibration is a core ingredient of any good application of set-theoretic methods. Tanner considers calibration a weakness because the raw data, which allows the estimation of marginal effects in quantitative research, is allegedly hidden behind less meaningful set membership scores. On the contrary, we argue that the requirement of calibration can be an important asset. Without doubt, calibration is a challenging task, which is why we appreciate Elkins' contribution to improving calibration strategies. But for reasons that are not obvious to us, Elkins' chapter is framed as specifying "alternative tools" (2014: 33) that stand in contrast to fuzzy sets. We rather see his three scaling techniques as providing useful ideas on how to *improve* fuzzy set calibration when good and numerous variables are at hand.

Points of Disagreement and Clarifications

Set-theoretic research can be Bayesian

Bennett's contribution recants a typology of necessary and sufficient criteria for hypothesis tests in process tracing that he himself pioneered in one of his earlier writings (2010) published in the second edition of *Rethinking Social Inquiry* (Brady and Collier 2010).[5] In his book section, Bennett states that he develops *necessary and sufficient criteria for hypothesis tests* and that he is not concerned with necessary and sufficient *conditions* in causal analysis (2010). Our teaching experience tells us that the difference between test criteria and causal conditions is frequently overlooked and thus the central message of Bennett's book section misunderstood. The distinction is crucial in the context of the symposium, though replacing the test criteria with a Bayesian framework is *unrelated* to the question of whether set-theoretic research with its in-built search for necessary and sufficient conditions is useful for causal analysis.

In fact, Bayesianism is *compatible* with a set-theoretic perspective on causal relationships (just as the two test criteria are compatible with it). Bayes' theorem tells us how we should change our confidence in a hypothesis in light of new evidence. For the use of Bayes' theorem, it is irrelevant whether our hypotheses and evidence are about set relations, correlations, or average treatment. Bennett's brief discussion of Bayesianism, set theory (in relation with Mahoney 2012), and directed acyclic graphs (in relation with Waldner 2014) might give a different impression to the reader because he presents them as alternatives instead of complementary accounts.[6]

Principles of identification versus research practice

Tanner introduces the idea of identification to the debate on QCA. In quantitative research, the causal relationship between two variables is identified when we are able to estimate a causal effect that is uniquely in accord with the data at hand. The identification *strategy* refers to the way in which we exploit variation in the treatment for causal inference. In lab or field experiments, the identification strategy is to randomly assign units to the treatment and control groups. Other identification strategies that have recently received attention are natural experiments, quasi-experiments (interrupted time-series, regression discontinuity, differences-in-differences), instrumental variables, and matching (Morgan and Winship 2007).

We believe that identification strategies are essential for set-theoretic research and that, so far, reflection on identification is rare, if existent at all in this field. So far, causal terminology, such as causal heterogeneity (i.e., equifinality), has been used to highlight the special features of set-theoretic research. These are important concepts, but problems such as the common-cause problem or endogeneity are also a threat to causal inference in set-theoretic research and should receive more attention in a broader discussion about identification strategies.

If this was Tanner's message, we would be in full agreement and stop our discussion here. He, however, compares selected quantitative and QCA studies from the field of policy analysis and comes to the conclusion that QCA is an "unsuitable method" (2014: 24). We neither share this conclusion nor do we find convincing the basis from which it is derived. Most importantly, Tanner conflates principles with practice in a way that biases the comparison to the disadvantage of QCA. Tanner selects six quantitative studies that follow different state-of-the-art identification strategies and are examples of best practice in quantitative research.[7] It comes as no surprise that the comparison with five QCA studies is in favor of the quantitative studies because, so far, identification has practically been a non-issue in the set-theoretic domain. While it is plausible to argue that the selected QCA studies could have done more for strengthening causal inference, it does not follow that, as a method, QCA is incompatible with any meaningful identification strategy.

Finding identification strategies appropriate for QCA should be viewed as an intellectually challenging endeavor rather than a discussion point for drawing definite conclu-

the double-headed nature of QCA as a *technique* or *method* and as an *approach* (e.g., Berg-Schlosser, De Meur, Ragin, and Rihoux 2008; Schneider and Wagemann 2010). We contributed to this debate by spelling out how to combine QCA with process tracing in set-theoretic multi-method research (Rohlfing and Schneider 2013; Schneider and Rohlfing 2013, 2014).

[5] We agree with Bennett that Bayesianism is more flexible than the 2x2 typology if it is used for Bayesian process tracing, but this is of secondary importance here.

[6] We find it surprising that Dion's (1998) seminal discussion of Bayesian inference in necessary-condition analysis—which has been referenced by Bennett elsewhere (Bennett and Elman 2006)—is not mentioned here.

[7] Tanner labels these "standard tools," but a glance at common practice in quantitative policy analysis suggests otherwise.

sions about QCA's inadequacy. A meaningful judgment about the quality of QCA can only be made after there is clarity as to what identification means in the context of set-theoretic research and which identification strategies work under which circumstances.[8]

Another argument of Tanner as to why QCA is unsuitable for policy research rests on the claim that "[…] policy research should be able to reveal modest effects at the margins" (2014: 24). QCA is close to useless for unravelling marginal effects because it focuses on multiple conjunctions and distinguishes between necessary and sufficient conditions as opposed to marginal effects. We are not policy researchers ourselves, but would find it remarkable if policy research, by definition, should never be interested in joint (conjunctive) effects of policy measures and always in marginal effects.[9] Tanner's claim that insights into necessary and sufficient conditions are per se undesired in the policy community seem far-fetched to us.

Social science concepts are not natural language...and don't need to be.

Lakoff takes issue with the claim that classic sets (crisp and fuzzy sets as used in QCA) directly correspond to natural language and questions their value for empirical analyses. Lakoff argues that natural language is best captured by family resemblance concepts because humans order objects according to their degree of similarity. Moreover, humans use hedges such as "strictly speaking," which can change the very boundaries of categories.

Our response to Lakoff's contribution and Collier's discussion of it is two-fold. First, we disagree that it is a "recurring" (Collier 2014: 3) claim in the set-theoretic literature that classic sets mirror natural language. The five cited set-theoretic publications by a single author and two pairs of authors are not strong enough evidence to call this a "common justification" (Lakoff 2014: 13). Furthermore, our reading of the referenced pages is that the authors do not claim that set theory reflects natural language. Schneider and Wagemann (2012: 7), for example, refer to the linguistics of concept formation in the social sciences and do not make any claim about the fit between logic and set theory on the one hand and natural language on the other hand. Similarly, Goertz and Mahoney (2012: 11–12, 16–18) say that qualitative researchers "naturally use the language of logic." They also have one section (perhaps misleadingly) titled "natural language and logic," but they do not claim that natural language is best captured by classic categorization. Like Schneider and Wagemann (2012), they discuss concept formation in the social sciences and suggest that for this purpose qualitative scholars draw on classic categorization.

Our second, related, response is that Lakoff's contribution is in our eyes largely irrelevant to the enterprise of set-theoretic research. Set-theoretic concept formation and methods do not aspire to be in line with natural language. The question therefore is less whether social scientists must take into account research in the domain of cognitive science[10]; we do not have any principled concerns against the marriage of cognitive and social science. Rather, the question is whether social science theories postulate covariation between variables or necessary and sufficient relations among sets. A look at any social science research field shows that both types of theories exist: some postulate covariation and others invoke set relations. Given the different formulations of theories, what matters most is that we align ontology with methodology and methods (Hall 2003).

A glance at existing research suggests that classic categorization can be of help in explaining the political and social world and has its place in empirical social science. For example, conceptualizations of democracy that neither invoke hedges nor rely on a family resemblance concept seem to work quite well in measuring the concept and contribute to our understanding of its causes and effects (Elkins 2000). Last but not least, if the misfit between classic categorization and natural language was a problem, then it would be a problem beyond the confines of set-theoretic research because the concepts used in other domains are usually also based on classic categorization.

With regard to the hedges that are central to natural language, a screening of social science concepts shows that we generally invoke classic categorization without hedges (as lucidly discussed by Collier in several publications, e.g., 1993). This again might be a mismatch with natural language and our cognitive processes, although there is also cognitive research ascribing a role to Boolean concepts for prototypes and examining the acquisition and processing of Boolean concepts (e.g., Goodwin and Johnson-Laird 2013).

There is one interesting critique raised by Lakoff. He argues that the way fuzzy sets are generated and used in current set-theoretic research is not in line with the way in which they should be used "because the scoring is based on fixed numerical values, rather than fuzzy distributions" (2014: 9). We agree but also hasten to point out that there is a large body of literature on so-called type-2 fuzzy sets, originally introduced by Zadeh (1975), summarized by Mendel (2007), and left unmentioned by Lakoff. Type-2 fuzzy sets contain information about the level of uncertainty of a case's fuzzy-set membership score that stem from uncertainty about the exact meaning of the concept to be captured by the fuzzy set.[11] Due to this property, type-2 fuzzy sets are sometimes called "fuzzy-fuzzy sets" (Mendel 2007: 21). Attempts at designing type-2 fuzzy-set QCA are under way (Korjani and Mendel 2014), which, once completed, will take care of Lakoff's critique of current practice.

[8] Without yet having considered these issues in further detail, we do not see a reason why randomization, either by our intervention, natural, or as-if, or other identification strategies, should not work in a set-theoretic context.

[9] Tanner seems to agree, judging from his appraisal of quantitative approaches capable of incorporating "context and causal heterogeneity" (2014: 22).

[10] While we are not cognitive scientists, our reading of the literature is that cognitive science is divided over some of the issues that Lakoff brings up as criticism against QCA (e.g., Pinker 2008).

[11] This is different from measurement error in fuzzy-set scores.

Developing meaningful robustness tests

Sensitivity analyses and robustness tests should become an integral part of empirical QCA research. We agree with Krogslund and Michel (hereafter KM) that simulations are among the useful tools for assessing the performance of set-theoretic research in general and QCA in particular. This is not undisputed among QCA proponents because simulations necessarily do away with case knowledge (e.g., Ragin 2014). We believe that QCA is much stronger when complemented with case knowledge, but we also think that simulations are helpful to understanding important issues that are at stake, even when researchers do have strong case knowledge.

KM's simulations are sophisticated in terms of their computational implementation, but simulations are only as useful as the *theory* on which they are based. By "theory," we mean a theory of QCA in terms of its inner workings and assumptions that must be fulfilled for the generation of correct QCA solutions. This is where we take issue with KM (and most other simulations on QCA that recently have been published): most of them complete the second step before the first by failing to offer a careful discussion of QCA's theoretical foundation.

KM build on Lijphart's famous diagnosis of a many-variables-few-cases problem (1971) and aim to determine whether the same problem pertains to QCA. The ratio of cases *n* to conditions *k* is at the heart of their analysis.[12] If QCA was suffering from an *n/k* problem, QCA results should be more robust with more cases, fewer conditions, or both. KM find the opposite, as they argue that QCA results tend to be more robust with a lower *n/k* ratio. We deem KM's main message ambiguous because they pursue two lines of reasoning. They state that they want to evaluate the utility of the *n/k* tool for QCA and whether it confirms the expectation. At the same time, they make claims about the (non)robustness of QCA results, i.e., they do evaluate the method. In the following, we focus on KM's claims about the robustness of QCA.[13]

For the discussion of the "drop-one rule," i.e., the stability of QCA results under what is probably more widely known as case-wise deletion, we find that KM misunderstand essential concepts of QCA with important consequences for the validity of their claims. First, they say the number of potential causal paths is equal to the number of truth table rows (Krogslund and Michel 2014: 27). Technically, the size of the truth table is the upper bound for the number of solution terms, but a truth table with all rows linked to the outcome is meaningless. Therefore, by definition, taking the number of truth table rows as the number of potential paths leads to an overestimation of the latter.

Second, and related, it is misleading to say that the number of "paths with at least one case" (Krogslund and Michel 2014: 27, table 1) is equal to the number of truth table rows with at least one case. KM ignore the crucial distinction between truth table rows that are sufficient for the outcome and those that are not. Without such a distinction, an assessment of whether QCA produces robust results is impossible simply because "truth table rows with cases" do *not* constitute the QCA result. What KM treat as the indicator of robustness—the number of truth table rows populated by cases—is not directly related to the more important and meaningful question of whether two separate QCA based on slightly different groups of cases return the same solution.

Our third critique is related to the data-generating process (DGP) in QCA. We deem it plausible to argue that the configuration of conditions that best describes a case embodies the DGP for this case. KM derive the expected share of rows with cases from the number of all truth table rows relative to the number of cases. This means they assume that each DGP is equally likely. In principle, every configuration captured by a truth table row should be able to produce the outcome and include cases. However, the phenomenon of *limited diversity* and social science theory tell us that some DGPs are more likely than others. If we were to use the expected share of rows with cases as a metric, we would need to take this into account and model unequal probabilities across DGPs. Once this is taken into account, it becomes much less straightforward to know in advance whether the expected share of rows with cases increases at a faster or slower rate as the number of conditions increases.[14]

Finally, the *n/k* ratio only captures the *expected* sensitivity of results, as clearly flagged by KM (2014: 28). For Krook (2010) and a second empirical QCA study they present, KM directly assess the robustness of the results to case-wise deletion (although we think it is based on misleading parameters; see our previous points). In contrast, the simulations and assessments of 52 datasets derived from the COMPASSS website (www.compasss.org) do not compare different QCA solutions and thus fail to demonstrate that the results are non-robust. They may or may not—we cannot tell from KM's findings. Even if direct robustness tests were done, we would still lack theory about the conditions under which case-wise deletion leads to non-robust results (see also Thiem in this newsletter issue in his response to Hug).

To conclude, we fully embrace the idea of robustness tests and sensitivity analyses for QCA. Because they are so important, it is crucial that future work on these issues puts the theory underpinning QCA and set theory first. The development of QCA theory is an important step toward the sophistication of this method and simulations can only be as good as the theory on which they rest.

Understanding and improving set calibration

The calibration of sets is an important and challenging component of applied set-theoretic research. Rather than taking base variables at their face value, the calibration process aims at interpreting base variables in light of their meaning vis-à-vis

[12] KM make matters unnecessarily complicated when using the term "variable" for what is called a "condition" in set-theoretic research, while at the same time, calling a condition what is usually termed a "conjunction," "sufficient term," or "path." A conjunction is one term of a QCA solution. We use the established terminology.

[13] Since we question KM's finding that QCA is more robust with a smaller *n/k* ratio, we indirectly question their conclusion as to the validity of the *n/k* sensitivity test.

[14] The rate of increase is an important point in KM's discussion.

the concept of interest. In this regard, Elkins' contribution is much more in line with the spirit of set-theoretic research than its title suggests, which invokes the notion of "alternative tools." Elkins embraces the idea of partial membership in categories and even sees fuzzy sets as a hardly contested descriptive device (2014: 34). Our interpretation of Elkins is that he does not want to argue against fuzzy sets, a framing of his contribution that emerges from Collier's introduction to the symposium. Quite to the contrary, he proposes enhanced calibration procedures for situations in which set calibration is based on more than one indicator and when more than just a dozen of cases are at hand. The three measurement strategies he proposes do not substitute but instead support the calibration of sets.

One point of discussion that we take as central in Elkins' proposal is the following. His strategies place a great deal of trust in the data for the assignment of fuzzy set membership scores. Conventional calibration, as envisaged by the current QCA literature, instead emphasizes the importance of employing criteria that stem from theory and conceptual knowledge and are exogenous to the data. In a sense, then, Elkins' strategies should only work well when a researcher has good reasons to trust the data more than conceptual judgment (Schedler 2012).[15] We believe that there is no principled argument to always prefer more data-driven driven strategies for classifying cases. The fear (or allegation) that the researcher's choice of the qualitative anchors as practiced in the calibration of sets per se renders the results subject to manipulation is not convincing to us. Those who want to cheat do cheat, no matter which methods are used. A much more significant question to be asked is this: Are the results produced by a given QCA robust against equally plausible set calibrations?[16]

Against this backdrop, Tanner makes several critical remarks related to calibration that can easily be resolved. First, summarizing the findings of Befani and Sager's QCA (2010), he states 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" (2014: 19). Fair enough, but if we think it is important to distinguish between 'complete non-compliance' and 'merely one missed deadline,' we are free to calibrate a new set accordingly and perform a separate QCA on this outcome. Tanner is right that the substantive interpretation of QCA findings for the negated outcome can be problematic, but it is wrong to believe that there are no QCA-inherent tools available to overcome this problem.

Second, discussing Lee's QCA (2013), Tanner (2014: 20) remarks that the large difference in fuzzy set membership scores of South Korea (0.95) and Japan (0.58) is perplexing, given that the difference between both case's values on the base variable used for calibration (percentage of temporary work-force) does not seem very big. This perplexity overlooks the fact that the difference in membership score is only a difference in degree between qualitatively identical cases: both are instances of the set under investigation because they hold a membership higher than 0.5. QCA results are more sensitive to the location of the 0.5 calibration anchor than to differences between cases on the same side of this anchor because the location of the 0.5 anchor determines to which truth table row a case belongs.

Third, Tanner holds that fuzzy sets imply a loss of information that researchers would have if only they stuck to the base variables used for calibration. Several remarks need to be made: (a) If the calibration function is reported—as it should be—each case's value on the base variable can be reconstructed if we think this is of substantive importance.[17] (b) The reason why sets and not base variables are used is because the latter do not have a one-to-one relation to the concept of interest. (c) In the social sciences—and we suspect also in policy research—indicators can be void of any easily interpretable scale. Think, for example, of prominent democracy indices and the impossibility of meaningfully interpreting a one-unit increase in scales like Freedom House or Polity.

In sum, Tanner's bold verdict of QCA being a "poor match for public policy" (2014: 15) seems inappropriate on the basis of his arguments and evidence. By this, we of course have not conclusively demonstrated that QCA is a useful tool in policy research. We simply invite scholars to ask more meaningful questions and to search for constructive answers.

The Way Forward: From Diagnosis of Problems to Improved Tools

We agree with the critics that a greater focus on the identification issue would be beneficial; that case knowledge should play a central role in set-theoretic research; and that set calibration is both crucial and improvable. We disagree whenever matters of current QCA practice are confounded with the method's principles, and when statements about QCA's viability and quality are based on misunderstandings about its inner working. We share parts of the diagnoses, but we emphatically disagree with the overall message that set-theoretic methods are inherently problematic and, as some contributors insinuate, are better abandoned by empirical researchers. Our take on the subject matter is different: we should use the diagnosis of problems as the starting point for a debate about what the nature and implications of the problems are and how their resolution can improve set-theoretic research.[18] To us, nothing in

[15] In Elkins' example (2014: 37f.), the US is identified as being equally presidential, parliamentary, and semi-presidential. Without an assessment of the relative trust in one's data and conceptual knowledge, it is difficult to judge whether the results produced by such data-driven calibration strategies do, in fact, reveal a conceptually interesting novelty or are just artifacts of the underlying data and/or specific method used.

[16] Interesting future work could consist of sensitivity analyses of Elkins' classification strategies. For instance, how much would the classification of, say, Russia as a presidential country change if we excluded, for example, Guatemala from the analysis? Conceptually, Guatemala should not matter for the classification of Russia. Yet, data-driven classification strategies seem to be potentially subject to such validity threats.

[17] This is more difficult if a set is calibrated based on more than one indicator and/or non-standardized indicators, such as historical sources, interviews, accumulated case knowledge etc. Yet, even then, a researcher is obliged to explicitly outline the calibration rationale in a transparent fashion.

[18] To illustrate our point, imagine a situation in the early days of

the newsletter suggests that the problems of set-theoretic research are beyond remedy.

One avenue for the future development of set-theoretic research and QCA in particular could be the parsing out of the assumptions that must be met in order to derive valid results. What we need, so to speak, is the equivalent to the Gauss-Markov conditions for OLS regressions. Engagement with assumptions and the failure to meet them should also take stock of the fact that set-relational data can be processed with algorithms other than Quine-McCluskey. For instance, Baumgartner recently proposed the Coincidence Algorithm (Ambuehl, Baumgartner, Epple, Kauffmann, and Thiem 2014; Baumgartner 2009). Although the Quine-McCluskey algorithm has been at the heart of QCA since its formulation in 1987, QCA is not wedded to it, in our reading. Any other algorithm is suitable as long as it is able to discern set-relational patterns in a dataset and facilitates the dialogue between case knowledge and cross-case inference. In our view, future developments and improvements should consist of the development of QCA's theoretical basis; the incorporation of second-order fuzzy sets in order to model uncertainty about the anchors; the modeling of common threats to causal inference such as common causes, omitted causes and measurement problems (see Thiem in this issue); a more systematic analysis of necessary conditions, an issue also largely neglected by QCA critics in the present symposium; updated formulas for the parameters of fit to make them less sensitive to skewed set membership scores; set-theoretic multi-method research putting the combining cross-case and within-case analyses on firm analytic grounds; and easy-to-use software packages so as to ensure that advances in methods find their way into empirical research, just to mention a few.

There is enough on the plate and we believe the present exchange, together with similar discussions in other outlets, is just the beginning and not the end of a long road towards further enhancing set-theoretic research.

References

Ambuehl, Mathias, Michael Baumgartner, Ruedi Epple, Alexis Kauffmann and Alrik Thiem. 2014. Cna: A Package for Coincidence Analysis. R Package Version 1.0-0. http://cran.r-project.org/web/packages/cna/index.html.

Baumgartner, Michael. 2009. "Inferring Causal Complexity." *Sociological Methods & Research* 38 (1): 71–101.

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

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*. Benoît Rihoux and Heike Grimm, eds. (New York: Springer), 171–192.

regression analysis when critics pointed to problems such as multicollinearity, omitted variable bias, etc. Where would science stand today if the diagnosis of these problems had led to the abandonment of regression analysis? Instead of giving up on the new tools, we now know the conditions under which OLS produces valid results, and what the consequences are when a specific assumption is violated. QCA should be put on a similar trajectory.

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

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

Bennett, Andrew and Colin Elman. 2006. "Qualitative Research: Recent Developments in Case Study Methods." *Annual Review of Political Science* 9: 455–476.

Berg-Schlosser, Dirk, Gisèle De Meur, Charles C. Ragin, and Benoît Rihoux. 2008. "Qualitative Comparative Analysis as an Approach." In *Configurational Comparative Methods*. Benoit Rihoux and Charles Ragin, eds. (Thousand Oaks: Sage), 1–18.

Brady, Henry E. 2013. "Do Two Research Cultures Imply Two Scientific Paradigms?" *Comparative Political Studies* 46 (2): 252–265.

Brady, Henry E. and David Collier. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 2nd ed. Lanham: Rowman & Littlefield Publishers.

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

Collier, David. 2014. "Problematic Tools: Introduction to Symposium on Set Theory in Social Science." *Qualitative & Multi-Method Research* 12 (1): 2–9.

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

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

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

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

Freedman, David A. 1991. "Statistical Models and Shoe Leather." *Sociological Methodology* 21: 291–313.

Goertz, Gary and James Mahoney. 2012. *A Tale of Two Cultures: Contrasting Qualitative and Quantitative Paradigms*. Princeton: 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.

Goodwin, Geoffrey P. and Philip N. Johnson-Laird. 2013. "The Acquisition of Boolean Concepts." *Trends in Cognitive Sciences* 17 (3): 128–133.

Hall, Peter A. 2003. "Aligning Ontology and Methodology in Comparative Research." In *Comparative Historical Analysis in the Social Sciences*. James Mahoney and Dietrich Rueschemeyer, eds. (Cambridge: Cambridge University Press), 373–404.

Korjani, Mohammad M. and Jerry M. Mendel. 2014. "Interval Type-2 Fuzzy Set Qualitative Comparative Analysis (It2- Fsqca)." Paper presented at NAFIPCS 2014. Boston, MA, Paper #93.

Krogslund, Chris and Katherine Michel. 2014. "A Larger-N, Few Variables Problem? The Counterintuitive Sensitivity of QCA." *Qualitative & Multi-Method Research* 12 (1): 25–31.

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

Lakoff, George. 2014. "Set Theory and Fuzzy Sets: Their Relation-

ship to Natural Language." *Qualitative & Multi-Method Research* 12 (1): 9–14.

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.

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

Mahoney, James. 2012. "The Logic of Process Tracing Tests in the Social Sciences." *Sociological Methods & Research* 41 (1): 570–597.

Medzihorsky, Juraj and Mario Quaranta. 2014. Setmethods, R Package Version 1. http://CRAN.R-project.org/package=SetMethods.

Mendel, Jerry M. 2007. "Type-2 Fuzzy Sets and Systems: An Overview." *Computational Intelligence Magazine* 2 (1): 20–29.

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

Pinker, Steven. 2008. *The Stuff of Thought: Language as a Window into Human Nature*. New York: Penguine.

Ragin, Charles C. 2014. "Comment: Lucas and Szatrowski in Critical Perspective." *Sociological Methodology* 44 (1): 80–94.

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

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): 220–235.

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

Schedler, Andreas. 2012. "Judgment and Measurement in Political Science." *Perspectives on Politics* 10 (1): 21–36.

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 Ingo Rohlfing. 2014. "Case Studies Nested in Fuzzy-Set QCA on Sufficiency: Formalizing Case Selection and Causal Inference." *Sociological Methods & Research,* online first: http://dx.doi.org/10.1177/0049124114532446.

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*. Cambridge: Cambridge University Press.

Skaaning, Svend-Erik. 2011. "Assessing the Robustness of Crisp-Set and Fuzzy-Set QCA Results." *Sociological Methods & Research* 40 (2): 391–408.

Tanner, Sean. 2014. "QCA and Causal Inference: A Poor Match for Public Policy Research." *Qualitative & Multi-Method Research* 12 (1): 15–25.

Thiem, Alrik and Adrian Duşa. 2012. *Qualitative Comparative Analysis with R: A User's Guide*. New York: Springer.

Waldner, David. 2014. "What Makes Process Tracing Good? Causal Mechanisms, Causal Inference, and the Completeness Standard in Comparative Politics." In *Process Tracing in the Social Sciences: From Metaphor to Analytic Tool*. Andrew Bennett and Jeffrey Checkel, eds. (Cambridge: Cambridge University Press), 126–152.

Zadeh, Lotfi A. 1975. "The Concept of a Linguistic Variable and Its Application to Approximate Reasoning-I." *Information Sciences* 8 (3): 199–249.

Encountering Your IRB: What Political Scientists Need to Know

Dvora Yanow
Wageningen University, Netherlands
Dvora.Yanow@wur.nl

Peregrine Schwartz-Shea
University of Utah
psshea@poli-sci.utah.edu

Pre-script. After we finished preparing this essay, a field experiment concerning voting for judges in California, Montana, and New Hampshire made it even more relevant. Three political scientists—one at Dartmouth, two from Stanford—mailed potential voters about 300,000 flyers marked with the states' seals, containing information about the judges' ideologies. Aside from questions of research design and ethics, whether the research passed IRB review is not entirely clear (reports say it did not in Stanford but was at least submitted to the Dartmouth IRB; for those who missed the coverage, see www.nytimes.com/2014/10/29/upshot/professors-research-project-stirs-political-outrage-in-montana.html and political scientist Melissa Michelson's blog http://thewpsa. wordpress.com/2014/10/25/messing-with-montana-get-out-the-vote-experiment-raises-ethics-questions/ (both accessed November 3, 2014)). Two bits of information offer plausible explanations for what have been key points in the public discussion:

1. Stanford may have had a reliance agreement with Dartmouth, meaning that it would accept Dartmouth's IRB's review in lieu of its own separate review;

2. Stanford and Dartmouth may have "unchecked the box" (see part 2, section 9 below), relevant here because the experiments were not federally funded, meaning that IRB review is not mandated and that universities may devise their own review criteria.

Still, neither explains what appear to be lapses in ethical judgment in designing the research (among others, using the state seals without permission and thereby creating the appearance of an official document). We find this a stellar example of a point we raise in the essay: the discipline's lack of attention to research ethics, possibly due to reliance on IRBs and the compliance ethics that IRB practices have inculcated.

***