Does set-relational causation fit into a potential outcomes framework?
An Assessment of Gerring's Proposal

Carsten Q. Schneider, Central European University (CEU), Budapest

&

Ingo Rohlfing, Cologne Graduate School, University of Cologne[1]

Published in Newsletter of the APSA Section Qualitative Methods and Multi-methods research, vol. 10, issue 1, pp. 8-14

1. INTRODUCTION

One of John Gerring's aims in his intriguing treatment of social science methodology is the development of a unified account for causal inference on the basis of the potential outcomes (PO) framework. Over the past two decades, the PO framework has become central in quantitative analyses (Morgan and Winship 2007). In qualitative research, in contrast, set theory and set-relational (SR) forms of causation and empirical research have started to play an ever more important role of unifying hitherto unrelated streams of qualitative literature (Goertz and Mahoney 2012; Ragin 1987; 2000; 2008; Rohlfing 2012, chap. 1; Schneider and Wagemann 2012). According to Gerring (2012, 337), the PO account is the more general framework and is able to accommodate SR causation such as necessity and sufficiency. In our contribution to this symposium, we discuss the viability of Gerring's proposal on how to perform SR research on the basis of the PO framework. We note here that the quest for

[1]Rotation principle. Both authors contributed equally to the paper. We link to thank Achim Goerres for helpful comments.

common methodological ground in causal analyses is important and Gerring's suggestions in this regard are novel. Our critical reflections are therefore not meant to question the quest for a unifying framework *per se*.

In section two, we briefly introduce some basics of the PO and SR framework. In section three, we outline Gerring's two-step proposal of how to analyze set relations from a PO perspective. The fourth section includes a critical reflection of Gerring's proposal, arguing that it falls short in correctly capturing several core features of SR causation. Most importantly, we show that the suggested procedure can produce false negatives – indicating the absence of a set relation when, in fact, one exists – and false positives – suggesting the presence of a set relation when there is none. In the concluding chapter, we detail some of the most important features of SR causality. If the PO and SR frameworks are truly compatible, all of these SR features must be transposed into the PO framework.

2. The Potential Outcomes Framework and Set-Relational Causation – Some Basics

The discussion of differences and similarities between PO and SR is best achieved by presenting them in 2x2 tables. In the simplest version of PO, both the treatment (the cause) and the outcome are binary. Likewise, the simplest SR approaches relate a single crisp-set condition (the cause) to a crisp-set outcome.[2] In a PO framework, the combination of a binary treatment and outcome yields four different combinations (table 1). The treatment can either be received (X) or not (~X), and one can or cannot observe the outcome for treated units (Y_X) and

[2] The PO and SR framework can work with more than one condition and these conditions and the outcome can be continuous (PO) and fuzzy sets (SR), respectively. In addition, SR subsumes notions of equifinality, conjunctural causation, and INUS and SUIN conditions. We postpone a discussion of some of these issues to a later section. For the time being, we follow Gerring in juxtaposing the basic setup.

non-treated units ($Y_{\sim X}$).[3] Two of these four scenarios can be observed, while the other two cannot (Morgan and Winship 2007, 35). We can observe the outcome for the treated for those units that were treated (cell A), and we can also observe the outcome for the untreated for those units that were not treated (cell C). However, we cannot observe the outcome for treated units if they were not treated in fact (cell B), and the outcome for untreated units if they were treated (cell D). The four scenarios are used for the assessment of *treatment effects*, that is, the difference the presence and absence of the treatment makes for the outcome. For example, we can ask: what difference does it make for the success of a party in national elections when it launches negative campaign advertisement as opposed to a campaign that lacks negative advertisement?

Table 1: Potential outcome framework

Outcome observed	For treated (Y_X)	D	A
	For untreated ($Y_{\sim X}$)	C	B
		No (~X)	Yes (X)
		Treatment received	

The PO literature has developed numerous types of treatment effects, some of which are summarized by Gerring (2012, 223). Given the implicit focus of Gerring's discussion, we limit ourselves to the *average treatment effect* (ATE).[4] The ATE captures the difference between the average outcome of the treated for treated units (cell A) and the average outcome of the untreated for non-treated units (cell C). Since, in our example, the outcome is binary, inquiries into treatment effects ask for the likelihood of observing the outcome, given that units receive or do not receive the treatment. In formal terms:

$$ATE = p(Y_X|X) - p(Y_{\sim X}|\sim X)$$

[3] In order to achieve some consistency between the notation in the PO framework and SR analyses, we use ~X to denote the absence of the treatment (~ meaning "not").

[4] Most of our arguments also apply to other treatment effects.

With binary conditions and outcomes, $p(Y_X|X)$ is equivalent to the ratio of cases in cell A, $p(X \cap Y_X)$, relative to all cases in the right column, p(X). Likewise, $p(Y_{\sim X}|\sim X)$ is derived from the ratio of cases in cell C, $p(\sim X \cap \sim Y_{\sim X})$, relative to all cases in the left column, p(~X). Clearly, then, the ATE relies on a *symmetric* notion of causation because a treatment makes a difference if there is a significant difference in the outcome between cases of X vis-à-vis those of ~X.

The conventional understanding of SR causation, in contrast, is *asymmetric* (Ragin 1987; 2000; 2008; Schneider and Wagemann 2012) and thus diametrically opposed to the PO framework. The difference can be seen once we ask which cells matter for the assessment of SR causation.

Table 2: Sufficiency and necessity in 2x2 tables

Outcome	Present	D	A
	Absent	C	B
		Absent	Present
		Condition	

The presence of a condition is *necessary* for the presence of the outcome if all cases that lack the condition are also non-members of the outcome. Formally, this means that $p(\sim Y|\sim X) = 1$.[5] From an SR perspective, claims of necessity only generate expectations about cases that are not members of the condition, i.e. about cases in the left column of table 2 (Gerring 2012, 339).[6] We expect cell D to be empty and cell C to contain cases. Similarly, the presence of a condition is *sufficient* if all cases that share the condition are also members of the outcome. Formally, it holds that $p(Y|X) = 1$. For sufficiency, only cases that share the condition are relevant, i.e. cell B should be devoid of cases and cell A should contain cases (Gerring 2012, 340). Finally, a condition is both *necessary and sufficient* if all members of the outcome set are also members of the condition

[5] We drop subscripts to Y in the following, keeping in mind that we are discussing the ATE.

[6] We follow Gerring in using this classic definition from logic. An alternative and more common definition in the methods literature is $p(X|Y) = 1$ (Goertz and Mahoney 2012). Both are equivalent and our arguments are independent of the definition used.

set and vice versa. Formally: $p(\sim Y|\sim X) = 1 \cap p(Y|X) = 1$. All that is required for a condition to be both necessary and sufficient is that cells B and D are devoid of cases and that cells A and C contain cases.[7]

An assessment of Gerring's proposal on how to handle SR requires a short overview on how the latter are currently analyzed by SR scholars. In brief, SR researchers first assess whether the condition set *is* sufficient, necessary, or neither of the two. Those conditions that are either necessary or sufficient are assessed with regard to their *empirical importance*. For step one, the measure of *consistency* is employed, for step two that of *coverage* (Ragin 2006; Schneider and Wagemann 2012, chap. 5, 9). Consistency measures the degree to which a distribution of cases is in accord with a postuated set relation. When interested in the sufficiency of X for Y, consistency is calculated by the share of cases in cell A relative to all cases in the right-hand column of the 2x2 table. The more cases that fall into cell B, the less consistent is the empirical evidence with the set-theoretic statement of sufficiency. In applied SR, deviations from perfect consistency are allowed. It follows that if consistency is higher than a specified threshold value, X is interpreted as sufficient. If consistency is lower, X is not considered as sufficient for Y.[8] The *degree* of consistency, thus, is only used to differentiate sufficient from non-sufficient set relations and not to express the degree of importance of sufficient condition.

Instead, it is coverage that expresses the empirical importance of a sufficient condition by asking: what is the share of cases that are members of the outcome and the condition relative to all cases displaying the outcome? This question is answered by counting the number of cases in cell A relative to all cases in the upper row. The lower the coverage, the lower is the empirical importance of the sufficient condition because fewer cases are captured by the sufficient

---

[7] Conditions that are both necessary and sufficient are rare in social science theory and empirical research alike. We therefore do not discuss them any further here.

[8] The precise consistency threshold varies with features of the analysis at hand (Schneider and Wagemann 2012, chap. 5), but should never be lower than 0.75 for sufficiency (Ragin 2006).

condition at hand. In contrast to consistency, there is no lower threshold for coverage below which conditions need to be dismissed as sufficient conditions.

Consistency has the same meaning in analyses of necessity. It expresses the degree to which the empirical evidence is in line with the statement that X is necessary for Y and is derived by the share of cases in cell C relative to all cases in the left column. Only conditions that pass a consistency threshold are interpreted as necessary conditions.[9] Conditions deemed adequately consistent to be interpreted as necessary are subjected to an evaluation of their empirical relevance. The standard assessment involves calculating the share of cases that are members of both the condition and the outcome (cell A) relative to all cases that are members of the condition.[10] As will become apparent below, consistency and coverage play a central role in our assessment of Gerring's proposal for a PO-centered analysis of set relations, which is introduced in the following section.

3. GERRING'S PROCEDURE FOR A POTENTIAL OUTCOMES ANALYSIS OF SET RELATIONS

Gerring (2012, 337-342) proposes a two-step approach for the analysis of set relations within a PO framework. In *step one*, the ATE is calculated to determine if X makes a difference to Y. For sufficiency, the rationale is that if X is fully consistent with a pattern of sufficiency, some cases are located in cell A while cell B is empty. Moreover, the expectation is that some cases are located in cells C and D. We therefore observe a difference in the distribution of cases in the left-hand and right-hand column. Since the column-wise distributions can be captured as conditional probabilities, which, in turn, are the components of the ATE, a sufficient condition thus gives rise

[9] For necessary conditions, the minimum consistency value should not be lower than 0.9 (Schneider and Wagemann 2012, chap. 9).

[10] Goertz (2006) suggests a slightly different procedure, whereas Schneider and Wagemann (2012) propose a formula for assessing the relevance of a necessary condition that integrates and elaborates on both Goertz' and Ragin's (2006) proposal.

to an ATE that is different from zero. The reverse reasoning applies for necessity. If X is a fully consistent necessary condition, there should be cases in cell C and no cases in cell D. The absence of X makes a difference if cells A and B contain some cases, again creating a difference between the distribution of cases in the left and right column and a non-zero ATE. However, a non-zero ATE alone does not suffice to infer *whether or not* X is a cause of Y, as this hinges on whether the ATE is *statistically* significant (Gerring 2012, 340).

Step two of the protocol aims at determining the *degree* to which X is necessary and sufficient, respectively. A significant ATE is an inadequate tool for this because its calculation draws on all four cells of the 2x2 table, whereas the analysis of any set relation only relies on two of the four cells. Since the same ATE can result from many different distributions of cases across the 2x2 table (see below), it is mandatory to look at how the cases are distributed across either the left or the right column. The left column captures the degree of necessity, expressed by the likelihood p(~Y|~X). The closer it is to zero, the more the pattern is in line with a statement of necessity. Correspondingly, the degree of sufficiency is expressed by the likelihood p(Y|X), The closer it is to one, the higher the degree of sufficiency (Gerring 2012, 340).

For an illustration of the two-step procedure utilized in an analysis of sufficiency, table 3 reproduces an empirical example provided by Gerring (2012, 338-340). The ATE is calculated by computing the probabilities of observing Y under X and ~X, respectively. These probabilities are 1, p(Y|X), and 0.5, p(Y|~X), yielding an ATE of 1-0.5 = 0.5. Since the ATE is statistically significant with t-score of about -22, we proceed to step two. A look at the two columns shows that we are dealing with a condition that is fully in line with a pattern of sufficiency due to a conditional probability of p(Y|X) = 1. At the same time, X can hardly be considered a necessary condition because the likelihood p(Y|~X) is 0.5 and thus very low.

Table 3: ATE for sufficiency for Gerring's hypothetical example

<table>
<tr><td rowspan="2">Outcome</td><td>Present</td><td>250</td><td>500</td></tr>
<tr><td>Absent</td><td>250</td><td>0</td></tr>
<tr><td></td><td></td><td>Absent
p(Y|~X) = 0.5</td><td>Present
p(Y|X)=1</td></tr>
<tr><td></td><td></td><td colspan="2">Condition</td></tr>
</table>

The example shows that the two-step approach is built on a division labor. Step one answers "the question of causality" (Gering 2012: 340) by differentiating between causes and non-causes. If the ATE is statistically significant, then X is a cause (if supplemented by causal mechanisms). Otherwise, it is not. Step two addresses "the question of probability" (Gering 2012: 340) then only asks for the degree to which X is sufficient and necessary, respectively. This implies four salient issues that are important to our attempt at evaluating the two-step proposal from an SR perspective.

First, the p-score of the ATE is the *only* basis for inferring causality (Gerring 2012, 340). If the ATE is significant, then the only unanswered question is how much X conforms to a specific form of set relation. If the ATE is insignificant, the analysis stops with the conclusion that there is no SR cause. Second, if a statistically significant ATE is detected, the column-wise conditional likelihoods only capture the degree to which we observe a set relation. Third, the two previous points imply that the *size* of the ATE is irrelevant. Both – whether X is an SR cause and to what extent – is captured by the p-score and a column-bound conditional likelihood. This implies that the *significance* of the ATE is the only PO framework element that carries some analytic weight in the analysis of SR causation. Fourth, the question of the empirical importance of an SR cause is left unaddressed by the two-step procedure as it exclusively focuses on statistical significance.

4. FIVE CRITICAL REFLECTIONS AT A GLANCE

Although compelling at first glance, we see five critical issues with the two-step procedure. Taking a broad view, the criticisms can be sorted into two rubrics. First, we question the claim that the PO framework can accommodate SR causation. Second, we cast doubt on the argument that the two-step protocol produces valid inferences on set relations.

Less unifying than it seems

Our skepticism concerning the unifying potential of the PO account is based on four observations. First and quite obvious, the PO framework alone does not suffice for the analysis of set relations: it is integral to step one, but irrelevant to step two in which only ordinary conditional probabilities are calculated. Second and relatedly, one cannot distinguish between patterns of sufficiency or necessity based on the ATE alone. In the presence of a significant ATE, one must resort to column-wise conditional probabilities in order to determine the degree to which a set relation is present.

Third, step two of the procedure resembles the assessment of consistency in SR research. This is problematic because step two would reinvent a set-theoretic wheel and unnecessarily unsettles the semantic field (Gerring 2001, 39-40). More importantly, it inverts the sequence of analyzing set relations as practiced in SR research wherein consistency tests must go first because they carry the information on whether or not X can be considered as a SR cause at all. Instead, in the two-step procedure, the conditional probabilities are not used to decide whether X *is* a set relational cause, but to express the degree to which a set relation is given. The notion of degree of sufficiency suggests that there is no threshold probability below which X ceases to be a sufficient condition. The two-step procedure, thus, can lead to statements that the degree of sufficiency is, say, 0.2. This is nonsensical from an SR perspective because a condition with a consistency of 0.2 definitely cannot be considered an SR cause.

Fourth, we argue that empirical research should also make statements about the empirical importance of conditions, something the two-step procedure does not allow for but is easily accommodated in SR framework through the parameter of coverage.

PO and valid set-relational inferences

The fifth and perhaps most important critique is that the significance test of the ATE is uninformative about the presence of absence of set relations. To illustrate this for the case of sufficiency, we provide empirical examples showing that the ATE can be insignificant in the presence of a sufficient condition – a *false negative* – and that the ATE can be significant in the presence of a non-sufficient condition – a *false positive*. This is particularly troublesome because the significance of the ATE is the only element in the PO framework that assumes an inferential role in the assessment of set relations.

In order to claim that the significance of the ATE cannot correctly identify set relations, we need to rely on a measure that can do so. We deem the consistency score as a plausible measure because it is in line with the meaning of necessity and sufficiency as defined in set theory and formal logic. We think it is fair to argue that if the two-step procedure and the current SR best practice prompt different conclusions, then this casts doubts on the two-step procedure because its set-theoretic pedigree is less clear than that of consistency (and coverage).[11] In the following, we use hypothetical data and address these criticisms in more detail.[12]

[11] This still leaves room for the conclusion that SR causes are not causes at all but simply descriptive statements about the social world (King and Powell 2008).

[12] Gerring emphasizes that his procedure ought to be applied only to conditions that are not trivial. All of the example we provide fulfill this criterion.

5. THE AMBIGUITY OF THE ATE: ILLUSTRATIONS

Significant ATE, unclear set relation

Table 4 includes two distributions that produce an ATE of the same size and with the same p-score. However, the panels display two very different set relations, namely a fully consistent pattern of necessity in the upper panel and a fully consistent pattern of sufficiency in the lower panel.

Table 4: Same ATE, different set relation

Outcome	Present	0	5
	Absent	10	5
		Absent $p(Y \mid {\sim}X) = 0$	Present $p(Y \mid X) = 0.5$
		Condition	
		ATE: 0.5; p = 0.001	

Outcome	Present	5	10
	Absent	5	0
		Absent $p(Y \mid {\sim}X) = 0.5$	Present $p(Y \mid X) = 1$
		Condition	
		ATE: 0.5; p = 0.001	

The significance of the ATE cannot discriminate between the two as it is insensitive to how the cases are distributed in the left and right column of the panel. It is precisely for this reason that step two is required in Gerring's protocol. The inspection of the column-wise conditional probabilities, however, is not an element of the PO framework. To us, this inability of the ATE to differentiate between necessity and sufficiency casts doubt on the more general claim that the PO framework can accommodate SR research.

We note that in the original exposition, step two of the protocol is only about the *degree* to which a set relation is present. This means that the two-step protocol does not require researchers to classify X as either a sufficient or a necessary condition. In this light, the problem that we identify here is not a problem because the upper and lower panels capture different degrees to which X is necessary and sufficient. However, we think that not classifying a condition as either necessary or sufficient or nothing is at odds with the established, qualitative view on set relations. While we agree that in applied empirical research it is warranted to allow for some deviation from perfectly consistent set relations, there nevertheless has to be a lower bound separating sufficient or necessary causes from non-SR causes. The argument that, given a significant ATE, X is always sufficient and necessary to some extent is very unusual.

Insignificant ATE and consistent set relations

A further problematic issue is that the (in-)significance of the ATE is a fallacious criterion for inferring the presence or absence of a set relation. For illustration, we use the identical distribution of a smaller number of cases, as in Gerring's example above. Our example only includes 12 cases, but these cases are distributed across the cells so that they produce the same column-wise conditional probabilities and the same ATE. A significance test shows that this reduces the t-score of the ATE from about -22 in the original example (N = 1000) to -1.83 with a p-value of .10 (two-sided). If we follow the convention and take .05 as the threshold for statistical significance, we have to conclude that X does not qualify as being sufficient. However, a set-theoretic researcher would surely draw a different conclusion because the lower-right cell is devoid of cases and X is a fully consistent sufficient condition for Y (and empirically relevant with a coverage score of 0.5).

Table 5: Insignficant ATE with fully consistent sufficiency

Outcome	Present	4	4
	Absent	4	0
		Absent p(Y|~X) = 0.5	Present p(Y|X) = 1
		Condition	
		ATE:0.5; p = 0.10	

When one subordinates the traditional understanding of set relations under the PO framework, the conclusion that X is not sufficient is unproblematic, for it resolves the question of SR causality by exclusive reliance on the significance test of the ATE. However, for those who are, like us, convinced that it is a good idea to follow the definition of set relations anchored in formal logic and set theory, the example casts serious doubt on the suitability of the two-step framework for set-theoretic analyses.

The more general insight is that if we keep the column-wise conditional probabilities fixed, SR inferences derived from the two-step procedure depend entirely on the number of cases. This is of practical importance because empirical SR research often operates on small to medium-sized Ns (Rihoux, Álamos, Bol, Marx and Rezsöhasy 2012). More precisely, the example shows that the two-step account can produce *false negatives*; the non-significant ATE makes us believe that X is not sufficient when, in fact, it clearly is from an SR perspective.

Significant ATE and inconsistent set relations

The two-step procedure also suffers from the pitfall of *false positives*, meaning that the ATE is significant in the absence of any set relation from an SR perspective. The prospect of false positives stems from the inappropriate handling of set-relational inconsistency by the PO framework. The example in table 6 yields a highly significant ATE of 0.40. However, the consistency of X as a sufficient condition is a dismal 0.60, far too low to interpret the condition as being sufficient according to standards in the current SR literature.

Table 6: Significant ATE with inconsistent set relations

Outcome	Present	20	60
	Absent	80	40
		Absent $p(Y \mid \sim X) = 0.2$	Present $p(Y \mid X) = 0.6$
		Condition	
		ATE: 0.4; $p < 0.001$; consistency: 0.6	

Why do the significance of the ATE and consistency convey different messages in this example? Apart from the large number of cases, which always boosts statistical significance, the difference stems from the way in which ATE and consistency treat the distribution of cases in the left column (~X). Consistency, staying true to the asymmetric nature of set relations, does not take the distribution in the left-hand column into account at all. The significance of the ATE, in contrast, also depends on how instances of ~X are distributed. In table 6, the ATE is significant because most instances of ~X are also cases of ~Y (80 out of 100). This means that the ATE is crucially shaped by a relatively large number of cases in the lower left cell. Thus, the inference that X is sufficient of Y is (largely) driven by cases that are neither members of X nor of Y.[13] From an SR perspective, this is a misguided procedure and inference. One might reject this criticism by pointing out that step two of the procedure only establishes the degree to which X is sufficient and necessary, respectively. In our empirical example, this would yield the statement that X is sufficient to a degree of 0.6 and necessary to a degree of 0.2. This, in our eyes, amounts to a non-sensical interpretation of set relations, though.

[13] We do not have a pattern of necessity here because consistency for necessity is only 0.80.

PO and empirical importance of conditions

The two-step procedure conveys information on whether a set relation is given and the degree of sufficiency and necessity. What is missing from the procedure is an assessment of empirical importance. In light of what we have said about the three elements of the PO framework above, the only quantity that serves no other purpose and could act as the measure of empirical importance is the *size* of the ATE. The larger the ATE, the more important the condition would be. However, the example in table 7 shows that the size of the ATE is not a valid measure for substantive importance in SR research as a *case-based* approach (Ragin 1987). With a size of 0.80, the ATE is relatively large and highly significant, which, in this example, matches a consistency score of 1.[14] However, the coverage of the sufficient condition X is very small with a score of 0.10.

Table 7: Large ATE, small coverage for sufficiency

Outcome	Present	90	10
	Absent	450	0
		Absent $p(Y \mid \sim X) = 0.2$	Present $p(Y \mid X) = 1$
		Condition	
		ATE: 0.8; $p < 0.001$; coverage: 0.1	

The difference between the size of the ATE and coverage is again due to the large number of cases in the lower-left cell. Ceteris paribus, the more cases that are located in this cell, the larger the size and significance of the ATE. In contrast, the coverage score remains entirely unaffected

[14] Gerring (2012, 338) issues the caveat that a cause should not be trivial. To us, it is not entirely clear what trivial means here. It seems that X counts as trivial when the column-wise conditional likelihoods are identical because then X does not make a difference to Y. However, it is not apparent why this should designate a trivial cause, as similar conditional likelihoods simply denote that X is a non-cause in a PO perspective. It is questionable to exclude all scenarios in which X does not make difference as trivial, as this arbitrarily limits the application of the two-step procedure to studies about which we know that X makes a difference. In any case, it is safe to argue that Gerring’s understanding of trivialness is different from the established SR understanding. In SR research, a condition is trivial when its coverage is negligible.

because its assessment only depends on cases in the upper row. SR research focuses on the upper row because it includes the cases that display the outcome, which is what we are interested in and seek to explain in SR analyses (Ragin 2008, chap. 11).

Summary

The two-step approach attempts to reconcile the asymmetry of SR causation with the symmetric notion of causation inherent to the PO framework. The critical issues that we raised above suggest that the PO and SR frameworks do not fit easily together, thereby reinforcing conclusions previously reached by other scholars in different contexts (see \Braumoeller, 2000 #206;Braumoeller, 2002 #375}(Clarke 2002; Mahoney 2008; Seawright 2002a; b).

One major reason for this misfit is that the first of the two-step procedure draws on cases from *all* cells of the 2x2 table, while set theory tells us that only one of the two columns is relevant for SR inferences. From a set-theoretic perspective, only the right column is of direct relevance for sufficiency, and only the left column for necessity. We do not claim that the two-step procedure and reliance on all four cells *always* produces false positives or false negatives. However, whether or not the two-step procedure and established SR analyses lead to the same conclusion (as it does in Gerring's example \, 2012 #6491, 339-340} depends on the distribution of the data and is therefore accidental.

6. Instead of a Conclusion – Where To Go From Here?

In concluding our contribution to the symposium, we deem it worthwhile to at least briefly address some of the hitherto unmentioned additional challenges in integrating the SR and PO framework. First, all examples discussed here and in Gerring (2012, 339) are drastically reductionist, for they assumed that SR researchers are interested in single conditions. In practice, however, researchers invoking SR causation usually do not so because they are interested in equifinality, unifinality, conjunctural causation, and INUS and SUIN conditions (Goertz and Mahoney 2012; Rohlfing 2012, chap. 2; Schneider and Wagemann 2012). This focus on *causal complexity* triggers a series of complications that cannot be discussed in detail here. For instance, the SR focus on conjunctions of conditions as opposed to single conditions cannot be adequately captured by interactions. Classic statistical analysis of interactions becomes overly complicated with more than three variables (Braumoeller 2004), whereas SR routinely manages higher-order conjunctions.[15] In contrast to Gerring (2012, 356), we also think that the factorial design does not capture the same essence as SR conjunctions for reasons too long to be exposed here.

Second, those who attempt to unify PO and SR must also come to terms with the fact that the latter often operates in a *small-N to medium-N setting*, thus raising doubts about the usefulness of statistical significance as a means for detecting set relations. By this, we do not mean that notions of statistical significance cannot, or should not, be applied to set relations. The SR literature has made several suggestions in this regard (Dion 1998; Eliason and Stryker 2009; Ragin 2000). These tests, however, must be designed such that they only take into account those cases that matter from a set-theoretic perspective (Braumoeller and Goertz 2000). Third, and relatedly, unifying PO and SR requires keeping separate rather than conflating or neglecting the set-theoretically grounded parameters of fit *consistency and coverage*. Put differently, we need clarity with regard to the PO status of consistency and coverage. Fourth, advanced SR has, of course, long moved beyond dichotomous sets (crisp sets), and has embraced the notion of *fuzzy*

[15] Braumoeller's (2003) Boolean probit and logit seem more promising as they accommodate high-order interactions and equifinality. However, apart from the enormous data requirements in terms of the numbers of cases, one has to specify the conjunctions and the substitutability of conjunctions (a.k.a., equifinality) in advance of the empirical analysis. One might consider this beneficial from the viewpoint of hypothesis testing, but it increases the risk of model misspecification. This is different for SR research (and *Qualitative Comparative Analysis* in particular), allowing one to test specific models and realize more exploratory analyses deriving equifinal conjunctions (or single conditions) from a menu of conditions.

sets (Ragin 2000). Since fuzzy sets are distinct from continuous variables (Ragin 2008, chap. 4), the PO literature on continuous treatments is only of limited use when trying to gauge how PO and fuzzy sets relate to each other.

We conclude that PO and SR frameworks might be compatible and that it is absolutely worthwhile to find ways in which they can be integrated. However, we also think that unless it can be shown that all core notions of SR causation can be adequately managed within the PO framework, it might be wiser to work under the assumption that the two are not reconcilable.

REFERENCES

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

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

Braumoeller, Bear F. (2004): Hypothesis Testing and Multiplicative Interaction Terms. *International Organization* 58 (4): 807-820.

Clarke, Kevin A. (2002): The Reverend and the Ravens: Comment on Seawright. *Political Analysis* 10 (2): 194-197.

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

Eliason, Scott R. and Robin Stryker (2009): Goodness-of-Fit Tests and Descriptive Measures in Fuzzy-Set Analysis. *Sociological Methods Research* 38 (1): 102-146.

Gerring, John (2001): *Social Science Methodology: A Criterial Framework*. Cambridge: Cambridge University Press.

Gerring, John (2012): *Social Science Methodology: A Unified Framework*. Cambridge: Cambridge University Press.

Goertz, Gary (2006): Assessing the Trivialness, Relevance, and Relative Importance of Necessary and Sufficient Conditions in Social Science. *Studies in Comparative International Development* 41 (2): 88-109.

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

King, Gary and Eleanor Neff Powell (2008): How Not to Lie without Statistics. *Paper prepared for delivery at the 2008 Annual Meeting of the American Political Science Association*.

Mahoney, James (2008): Toward a Unified Theory of Causality. *Comparative Political Studies* 41 (4-5): 412-436.

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 Quantitative and Qualitative Strategies*. Berkeley: University of Berkeley Press.

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

Ragin, Charles C. (2006): Set Relations in Social Research: Evaluating Their Consistency and Coverage. *Political Analysis* 14 (3): 291-310.

Ragin, Charles C. (2008): *Redesigning Social Inquiry*. Chicago: Chicago University Press.

Rihoux, Benoît, Priscilla Álamos, Damien Bol, Axel Marx and Ilona Rezsöhasy (2012): From Niche to Mainstream Method? A Comprehensive Mapping of Qca Applications in Journal Articles from 1984 to 2011. *under review*.

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

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 (2002a): Testing for Necessary and/or Sufficient Conditions: Which Cases Are Relevant? *Political Analysis* 10: 178-193.

Seawright, Jason (2002b): What Counts as Evidence? Reply. *Political Analysis* 10 (2): 204-207.