Comparative Political Studies
1–39

DOI: 10.1177/0010414014564851
cps.sagepub.com
SAGE

Set-Theoretic Comparative Methods: Less Distinctive Than Claimed

Jack Paine[1]

Abstract

Proponents of set-theoretic comparative methods (STCM) sharply *differentiate* their approach from quantitative analysis—unlike many researchers who focus on *integrating* qualitative and quantitative methods. This article engages these opposing views by demonstrating shared foundations between STCM and quantitative techniques. First, it shows how the quantitative practice of analyzing cases that exhibit variation on both the explanatory conditions and the outcome—for example, all four cells of a 2 × 2 table—guards against misleading conclusions about necessary/sufficient conditions. Hence, conventional statistical ideas about association are relevant for STCM. Second, STCM's tools for analyzing causal complexity share important features with regression interaction terms. Third, scrutinizing these shared foundations suggests how stronger theoretical and empirical standards for causal inference with deterministic hypotheses can be established. Focusing on shared foundations and recognizing that STCM does not genuinely break new inferential ground facilitate new opportunities for strengthening comparative research tools, rather than unproductively overemphasizing differences from mainstream methods.

Keywords

causal inference, comparative method, determinism, necessary and sufficient conditions, QCA, qualitative comparative analysis, qualitative methods, regression, set theory

[1]University of California, Berkeley, CA, USA

Corresponding Author:
Jack Paine, Department of Political Science, University of California, 210 Barrows Hall, Berkeley, CA 94720, USA.
Email: jackpaine@berkeley.edu

The relationship between qualitative and quantitative methods—similarities, contrasts, and the feasibility of integration—is the focus of a major current debate in comparative politics. On one hand, many works stress the distinctive strengths of qualitative methods while also emphasizing that qualitative and quantitative traditions share many inferential goals. Brady and Collier (2010) advocate "shared standards" for different research traditions, Lieberman (2005) presents an iterative method for combining statistical and case study findings, Dunning (2012) demonstrates the crucial contribution of qualitative work to evaluating natural experimental designs, and Seawright (2014a) advocates an integrative approach that bridges these two approaches.

On the other hand, several recent and influential books emphasize fundamental differences between how qualitative and quantitative scholars approach social phenomena. Ragin (2008), Goertz and Mahoney (2012), and Schneider and Wagemann (2012) argue the core goal of qualitative research is to evaluate "complex" combinations of necessary and/or sufficient conditions. They propose a set-theoretic approach for studying these relationships. Given the focus of these three books and related work on set theory and cross-case comparisons, I refer to the techniques as set-theoretic comparative methods (STCM).

STCM scholars argue that there are sharp differences between their tools and quantitative methods. Mahoney, Goertz, and Ragin (2013) suggest that "in the social sciences, statistical and set-theoretic scholars adopt different approaches to causal analysis" (p. 75). Goertz and Mahoney (2012) maintain that "overcoming the quantitative-qualitative division in the social sciences is significantly a matter of better understanding the methodological differences between these two traditions along with the reasons why those differences exist" (p. 5).

Ragin (2008) begins his book by emphasizing sharp differences between STCM and quantitative methods. He underscores the contrast between "set-theoretic versus correlational connections" (pp. 6-10). A key example is found in Ragin's argument that scholars, when studying necessary and sufficient conditions, should not incorporate cases that exhibit variation on both the posited causal conditions and the outcome—and that one of the cells of a 2×2 table is *never* needed to analyze either necessity or sufficiency (Ragin, 2000, p. 96; 2008, pp. 20-23).[1] This approach is presented as sharply differentiating STCM from standard ideas of correlation and association in quantitative methods, as well as with quantitative case selection practices. Furthermore, in discussing "configurations of conditions versus 'independent' variables," Ragin calls for a focus on alternative constellations of causal conditions. He claims that regression, by contrast, is limited to examining the

net effect of individual variables—a shortcoming Ragin sees as inherent even in regression with interaction terms.

STCM has commendably brought core issues in comparative analysis to the forefront of methodological attention. Its proponents routinely emphasize many foundational points about data analysis that should be heeded in all work, qualitative or quantitative. These considerations include careful case selection, close attention to measurement, and the inherent conditionality of most causal relationships. For these reasons, this approach has become an influential part of the broader reinvigoration of qualitative methods research in response to, and in part in reaction against, King, Keohane, and Verba ("KKV," 1994).

However, by emphasizing the *differences* between their methods and quantitative research, STCM scholars have also posed an important challenge to "post-KKV" qualitative methods that seek *integration* with quantitative techniques—as with the studies just noted of Brady and Collier, Lieberman, Dunning, and Seawright.

This article attempts to advance this important debate by demonstrating shared foundations between STCM and quantitative methods—which, consequentially, counters STCM's emphasis on differences. The core STCM procedures calculate associational measures for complex combinations of posited necessary/sufficient conditions. Contrary to STCM arguments, these procedures share core similarities with conventional quantitative techniques. Understanding these similarities not only helps scholars from different traditions better understand each other's research but also shows that STCM do not genuinely break new inferential ground. By strongly emphasizing differences, STCM has unproductively isolated itself from mainstream methods. Furthermore, the focus in this article on common foundations also highlights a central shared challenge in much social scientific research, including STCM: moving from association patterns to infer causal relationships, on the basis of observational data.

Notably, the specific comparison with conventional quantitative methods employed below is vital not because the present analysis advocates using conventional quantitative tools. Rather, the fact STCM scholars frequently evoke this contrast to justify their own approach necessitates scrutinizing how STCM's focus on differences may obscure shared foundations.

Following the overview section that summarizes key features of STCM, detailing the virtues of analyzing cross-case variance provides the first example of a crucial similarity between a core analytic procedure of STCM and quantitative techniques. Evaluating necessary condition hypotheses requires analyzing cases that exhibit variation on both the causal conditions and the outcome—as in quantitative research but contrary to the consensus of STCM

scholars. To understand the importance of analyzing all four cells (in a 2 × 2 setting), suppose the hypothesis states $X = 1$ is a necessary condition for $Y = 1$. To evaluate this hypothesis, STCM calculates the percentage of $Y = 1$ cases with $X = 1$, hence only using two cells.[2] Using the STCM metric, the data support the necessary condition hypothesis if there are few $Y = 1$ cases with $X = 0$.

The problem with only using two cells, however, is that there may be few $Y = 1$ cases that are also $X = 0$ cases, for two reasons: (1) The data actually support the necessary condition hypothesis or (2) there are simply few $X = 0$ cases relative to $X = 1$ cases—which makes it essentially impossible to falsify the hypothesis using the STCM metric. Using information from all four cells ensures that consideration (2) will not create false positives about necessary relationships—in particular, by incorporating information from a cell that STCM scholars claim is *never* relevant for assessing necessity/sufficiency. An identical argument applies to evaluating the "relevance" of a sufficient condition hypothesis.

The next argument demonstrates inherent similarities between STCM and quantitative approaches to studying complex relationships, a central focus of STCM research.[3] This argument compares STCM analysis of a truth table with regression analysis that employs multiplicative interaction terms. The analysis demonstrates how regression can be used to convey the same information about necessary/sufficient condition hypotheses as a truth table—in contrast to strong STCM skepticism about multiplicative regression interaction terms.

Researchers who acknowledge these core similarities will not be surprised that STCM and quantitative methods also share a crucial common limitation. This involves the shared challenge in social science of inferring *causal* conclusions from *associational* relations. Examples of associational measures include STCM metrics for measuring the consistency and relevance of complex conditions as necessary or sufficient. Whereas current best-practice in quantitative textbooks and research focuses centrally on this shared challenge, it has received less attention from STCM scholars. The final major argument therefore posits a best-case scenario in which a researcher has data perfectly consistent with a necessary/sufficient condition hypothesis (possibly involving complex conditions). It focuses on two issues that must be addressed before a compelling causal inference can be achieved: Freedman's (2010) standard for comparing a deterministic hypothesis to a probabilistic benchmark and Waldner's (2005) standard for evaluating hypothetical counterexamples.

The concluding section argues that these common foundations—including both core similarities and shared challenges—between STCM and quantitative methods raise a series of pressing issues. The bridging attempt offered by this article will hopefully clarify future methodological debates in

comparative politics. These not only include the relationship between STCM and quantitative methods but also between STCM and traditional qualitative techniques.

STCM Techniques for Analyzing Necessary/ Sufficient Conditions

This section briefly outlines key elements of STCM, which provides building blocks for the subsequent discussion. Following the focus of STCM texts on 2×2 tables to argue for differences, I restrict attention to binary conditions throughout the article. The concluding section discusses possible extensions to multi-valued fuzzy sets.

STCM scholars view natural language, qualitative theory in social science, and qualitative research in general as inherently set-theoretic in structure. Stemming from this view, STCM researchers focus on studying the necessary and sufficient conditions for set membership in the outcome of interest. If a condition $X = 1$ is necessary for a particular outcome $Y = 1$, then every $Y = 1$ case is also an $X = 1$ case. Therefore, $X = 1$ is a superset of $Y = 1$. If condition $X = 1$ is sufficient for $Y = 1$, then every $X = 1$ case is also a $Y = 1$ case. Therefore, $X = 1$ is a subset of $Y = 1$. Framing necessary and sufficient conditions in terms of sets has led STCM scholars to adopt the broader mathematical language of set theory.

These scholars argue that it is reasonable to consider necessary and sufficient condition hypotheses even when empirical counterexamples to the hypothesis exist. For example, Braumoeller and Goertz (2000, 2002) underscore the importance of taking measurement error into account when evaluating such hypotheses. STCM scholars have developed two novel measures for studying necessary/sufficient relations in noisy social science data. These measures can be used for either a single or for multiple conditions.

Consistency Scores and Subsets/Supersets

A consistency score measures the extent to which the data support a claim of either necessity or sufficiency. The higher the percentage of $X = 1$ cases in the data that also achieve $Y = 1$—that is, the extent to which $X = 1$ is a subset of $Y = 1$—the more consistent the data are with a claim that $X = 1$ is sufficient for $Y = 1$. The higher the percentage of $Y = 1$ cases with the condition $X = 1$—that is, the extent to which $X = 1$ is a superset of $Y = 1$—the more consistent the data are with a claim that $X = 1$ is necessary for $Y = 1$ according to conventional STCM metrics.

Coverage Scores and Triviality

The coverage score assesses what STCM scholars call the "triviality," or "relevance," of a condition (Schneider & Wagemann, 2012, p. 144). As an example of a trivial sufficient condition, suppose that every $X = 1$ case achieves $Y = 1$, meaning the data are perfectly consistent with the hypothesis. However, suppose that every $X = 0$ case also achieves $Y = 1$. Despite the high sufficiency consistency score for $X = 1$, we would intuitively think of $X = 1$ as a trivial sufficient condition for $Y = 1$. Because $Y = 1$ will always occur in the data set, regardless of the value of X, the absence of the trivial sufficient condition does not change the outcome. The coverage score for a sufficient condition is identical to the consistency score for a necessary condition. In this example, if we assume that there are an equal number of $X = 0$ and $X = 1$ cases, the coverage score is 0.5 because cases with $X = 1$ account for only half of the $Y = 1$ outcomes.

A similar concept of triviality, or relevance, applies to necessary condition hypotheses. Assume that no $X = 0$ cases achieve $Y = 1$, implying the data are perfectly consistent with the hypothesis. But suppose that $Y = 1$ almost never occurs even when $X = 1$. For example, oxygen is a necessary condition for social revolution. The necessary condition coverage score—which is identical to the sufficient condition consistency score—detects that $X = 1$ is a trivial necessary condition because it will show that only a trivial percentage of $X = 1$ cases have $Y = 1$.

Multiple Conditions

In addition to the consistency and coverage scores for individual conditions, STCM also provides tools for analyzing clusters of multiple conditions. For example, a condition $A = 1$ may not be individually sufficient for $Y = 1$, but the conjunctural condition $(A, C) = (1, 1)$ may be sufficient. This is an example of a complex sufficient condition. Furthermore, $(D, E) = (1, 0)$ may also be a sufficient condition for $Y = 1$. This causal process therefore exhibits equifinality because there are multiple paths to the outcome.

Thus, the basic ideas of consistency, coverage, and multiple conditions provide key building blocks for STCM analysis of necessity and sufficiency.

Core Similarity 1: The Virtues of Analyzing Cross-Case Variance

Empirically evaluating necessary condition hypotheses requires analyzing cases that exhibit variation on both the causal conditions and the outcome[4]—which demonstrates a crucial similarity to quantitative research. This claim

runs against STCM arguments that necessary condition hypotheses can only be meaningfully tested using designs that lack variation on the outcome. Problematically, this accepted STCM procedure may lead one to conclude data are highly consistent with the necessary condition hypothesis simply because few cases lack the posited necessary condition, which produces artificial support for the claim. In a 2 × 2 setting, guarding against this problem requires incorporating all four cells into the analysis. Appendix A demonstrates the STCM procedure for assessing the triviality of a sufficient condition hypothesis faces a similar shortcoming that can also be fixed by incorporating information from all four cells.

There is wide agreement in the STCM literature that scholars should only focus on two cells at a time. Ragin (2008) argues the quantitative tradition of combining all four cells of a 2 × 2 table will "conflate different kinds of causal assessment" (pp. 7, 22). Braumoeller and Goertz (2002) argue that sampling cases from all four cells to test necessary condition hypotheses is "pointless" because "the theories that [they] examine don't imply anything about the number or proportion of cases that should be found in [certain] cell[s]" (pp. 199, 200). Goertz and Mahoney (2012) state directly that "selection on the dependent variable when testing necessary conditions follows directly from the definition of a necessary condition" (p. 179). Correspondingly, the standard STCM calculations of consistency and coverage for both necessary and sufficient conditions incorporate only two cells at a time—in contrast to the quantitative approach of analyzing all four cells, such as with a regression coefficient or any standard measure of statistical association.

However, analyzing only two cells at a time can create misleading findings whenever the number of $X = 1$ and $X = 0$ cases differs. For example, consider a claim that $X = 1$ is necessary for $Y = 1$, implying that $X = 0$ cases with $Y = 1$ provide evidence against the hypothesis. If the number of $X = 1$ cases is large relative to the number of $X = 0$ cases, the data may appear to be highly consistent with the necessary condition hypothesis simply because there are relatively few $X = 0$ cases that *could* provide evidence against the hypothesis. This occurs because it is not *possible* for there to be many $(X, Y) = (0, 1)$ cases if there are few $X = 0$ cases. The easiest way to guard against this problem is to compare the *percentage* of $X = 1$ cases with $Y = 1$ to the *percentage* of $X = 0$ cases with $Y = 1$. This revised procedure incorporates all four cells.

The core argument complements and extends existing arguments that researchers should always incorporate all four cells when the data are available. Seawright (2002) advanced this argument using a Bayesian model. The present arguments address two major concerns from responses published in the same issue of *Political Analysis*. First, Clarke (2002) argues Seawright

Table 1. Notation for a 2 × 2 Table.

	$F = 0$	$F = 1$
$A = 1$	n_{fA}	n_{FA}
$A = 0$	n_{fa}	n_{Fa}

This table presents notation for each cell of a 2 × 2 table. In the subscript, uppercase refers to the presence of a condition and lowercase to its absence. F = fuel wealth; A = authoritarian rule.

used an implausible weighting procedure in his Bayesian likelihood function for introducing new cases into the analysis. In the three cross-tabulations presented below (Tables 2-4), both the total number of cases and the distribution of $X = 1$ and $X = 0$ cells are fixed for each table. Therefore Clarke's critique does not hold here.[5] Second, Braumoeller and Goertz (2002) critique Seawright for dismissing necessary condition hypotheses when deviant cases exist, whereas this section allows for the possibility of cases that are inconsistent with a deterministic relationship. Separately, Freedman (2008) disputes Goertz's (2008) advice to ignore certain cells in a 2 × 2 table. In contrast to the focus of this section, Freedman (2008) grounds his critique by appealing to process tracing:

> At least in my experience, it is often hard to see where the cases go until you study them . . . Great work can be done with one cell, or even one case. Isn't de Tocqueville's *Democracy in America* a classic example of within-case analysis? (pp. 15-16)

A Revised Approach to Calculating Necessary Condition Consistency Scores: The Importance of All Four Cells

The exposition follows notation from the generic 2 × 2 table depicted as Table 1. To use substantively interesting terminology, suppose condition F refers to "fuel wealth" and A refers to the outcome "authoritarian rule."

The consistency score for fuel wealth as a sufficient condition for authoritarian rule calculates the percentage of $F = 1$ cases with $A = 1$. Similarly, the consistency score for non-fuel wealth as a sufficient condition for authoritarian rule calculates the percentage of $F = 0$ cases with $A = 1$. Using the notation from Table 1, these two terms can be expressed as follows:

$$\mathrm{SC}\left(F=1, A=1\right) = \%\left(A=1 \mid F=1\right) = \frac{n_{FA}}{n_{FA} + n_{Fa}}, \tag{1}$$

$$\mathrm{SC}\left(F=0,A=1\right)=\%\left(A=1\mid F=0\right)=\frac{n_{fA}}{n_{fA}+n_{fa}}. \tag{2}$$

Examined individually, each term incorporates only two cells at a time.

In contrast, a bivariate regression coefficient combines information from all four cells. The regression coefficient for the association between fuel wealth and authoritarianism is defined as the covariance of F and A divided by the variance of F. In a 2 × 2 setting, this is simply the difference in average outcome value between fuel and non-fuel cases, which is expressed by

$$\beta_{\text{fuel}}=\frac{n_{FA}}{n_{FA}+n_{Fa}}-\frac{n_{fA}}{n_{fA}+n_{fa}}. \tag{3}$$

Comparing Equations 1 through 3 shows the regression coefficient combines the two sufficient condition consistency scores, and Equation 3 can be re-expressed as

$$\beta_{\text{fuel}}=\mathrm{SC}\left(F=1,A=1\right)-\mathrm{SC}\left(F=0,A=1\right). \tag{4}$$

As proposed by STCM scholars, the consistency score for necessary conditions only incorporates information from two cells. The consistency score for fuel wealth as a necessary condition for authoritarian rule conveys the percentage of $A = 1$ cases with $F = 1$:

$$\mathrm{NC}\left(F=1,A=1\right)=\frac{n_{FA}}{n_{FA}+n_{fA}}. \tag{5}$$

A crucial problem emerges here. The metric from Equation 5 can be large for two different reasons. First, n_{fA} may be small because the data strongly support the hypothesis. Second, n_{fA} may be small because there simply are not many cases that lack fuel wealth among the included cases, which implies there cannot be many $(F, A) = (0, 1)$ cases. In other words, without accounting for differences in the frequency of non-fuel cases relative to fuel cases, the STCM necessary condition consistency score can create misleading findings.

Tables 2 through 4 demonstrate the potential for misleading findings by illustrating the importance of how cases are distributed across $F = 0$ and $F = 1$. The tables present three hypothetical 2 × 2 cross-tabulations. Because the number of cases in both $F = 1$ cells is constant across the three tables, the STCM necessary condition coverage score is also the same. Therefore, none of the following concerns relate to this commonly used measure for triviality.[6]

With regard to whether fuel wealth is necessary for authoritarianism, according to Equation 5, Tables 2 and 3 have high consistency scores of 20/22 = 91% and 20/23 = 87%, respectively. In contrast, the consistency

Table 2. Hypothetical Example of Misleading Necessary Condition Consistency Scores.

	$F = 0$	$F = 1$
$A = 1$	2	20
$A = 0$	0	20

This table is a hypothetical 2 × 2 table. It represents a fixed group of cases that are distinct from those in Tables 3 and 4. F = fuel wealth; A = authoritarian rule.

Table 3.

	$F = 0$	$F = 1$
$A = 1$	3	20
$A = 0$	37	20

This table is a hypothetical 2 × 2 table. It represents a fixed group of cases that are distinct from those in Tables 2 and 4. F = fuel wealth; A = authoritarian rule.

Table 4.

	$F = 0$	$F = 1$
$A = 1$	20	20
$A = 0$	20	20

This table is a hypothetical 2 × 2 table. It represents a fixed group of cases that are distinct from those in Tables 2 and 3. F = fuel wealth; A = authoritarian rule.

score in Table 4 is low, 20/40 = 50%. To understand why these consistency scores are problematic, consider the differences among the three tables. Tables 3 and 4 each have 38 more cases with fuel wealth than Table 2. The difference between Tables 3 and 4 consists of how the $F = 0$ cases are distributed between the outcomes $A = 0$ and $A = 1$. In Table 3, almost all the additional cases are in the (0, 0) cell, whereas in Table 4, the two $F = 0$ cells have an identical number of cases.

It is puzzling that the consistency score for Table 2 is high—especially when compared with Table 3. In Table 2, *every* fuel-poor case goes against the necessary condition hypothesis, whereas only 7.5% do so in Table 3. In Table 3, there are an equal number of fuel-poor and fuel-rich cases. Therefore, the rarity of fuel-poor cases with authoritarian rule indicates support for the hypothesis, a reasonable conclusion.

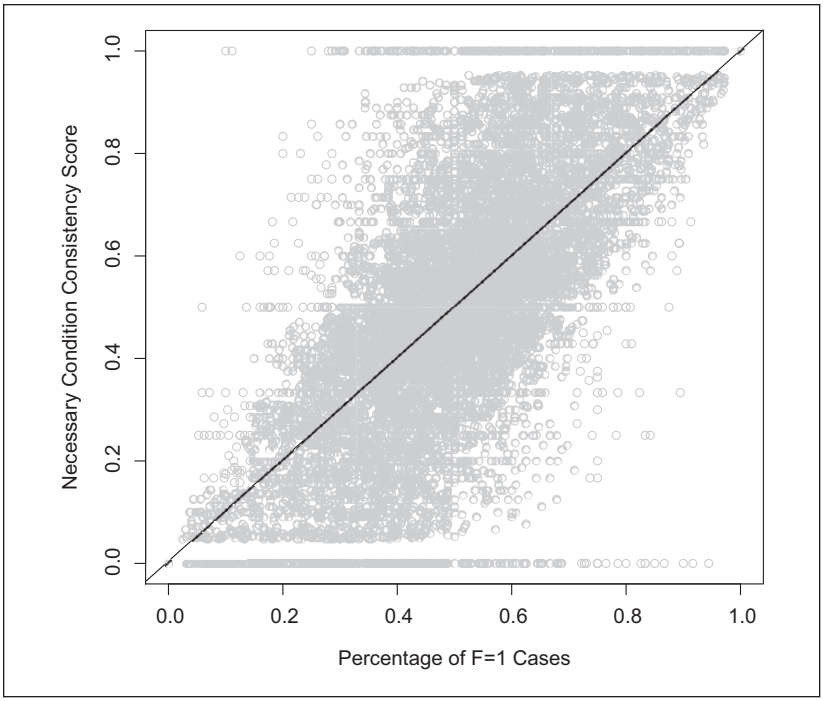

Figure 1. Simulated relationship between the percentage of $F = 1$ cases and STCM necessary condition consistency scores.
This figure demonstrates the systematic tendency for necessary condition consistency scores to be higher in data sets with a higher percentage of $F = 1$ cases.

In contrast, Table 2 simply does not contain many $F = 0$ cases—which explains the rarity of (0, 1) cases. In fact, in Table 2, there is evidence that fuel-poor cases are *more* likely than fuel-rich cases to experience authoritarian governance. Nonetheless, the STCM calculation that incorporates only two cells suggests the data in Table 2 are highly consistent with fuel wealth being necessary for authoritarianism.

In another puzzling finding, the consistency score is considerably higher in Table 2 than in Table 4, even though a *higher* percentage of fuel-poor cases contradict the hypothesis in Table 2. The main difference between Tables 2 and 4 is that the number of fuel-poor cases is greater in Table 4.

To provide more systematic evidence for why overlooking this crucial similarity with quantitative case selection practices causes problems, Figure 1 summarizes 10,000 randomly generated 2 × 2 tables. In every simulated table,

there is an equal chance for each of the four cells to be assigned anywhere between 0 and 20 cases. The horizontal axis presents the percentage of total cases assigned to $F = 1$. The vertical axis presents the STCM necessary condition consistency score for each of the 10,000 hypothetical tables.

An approach that incorporates only two cells at a time implies the distribution of cases across $F = 1$ and $F = 0$ is unimportant. Crucially, this distribution can only be calculated by incorporating information from all four cells. Therefore, if STCM scholars are correct that one does not lose valuable information by discarding particular cells, there should be *no systematic tendencies* in this randomly generated data and the best-fit line should be *flat*. Instead, the black best-fit line shows a sharply positive association between the percentage of cases assigned to $F = 1$ and the necessary condition consistency score. That is, one is more likely to find support for their hypothesis using the STCM metric simply because their data set has a high percentage of $F = 1$ cases. Thus, neglecting to study all four cells can easily engender misleading conclusions.

A necessary condition consistency score calculation that incorporates all four cells would overcome these problems. For example, consider the following metric:

$$\mathrm{NC}\left(F=1, A=1; 4\ \text{cells}\right) = \frac{\dfrac{n_{FA}}{n_{FA}+n_{Fa}}}{\dfrac{n_{FA}}{n_{FA}+n_{Fa}}+\dfrac{n_{fA}}{n_{fA}+n_{fa}}} = \frac{\mathrm{SC}\left(F=1, A=1\right)}{\mathrm{SC}\left(F=1, A=1\right)+\mathrm{SC}\left(F=0, A=1\right)}. \tag{6}$$

Equation 6 incorporates all four cells.[7] It adjusts for the frequency of fuel-poor cases relative to fuel-rich cases and does not report a high consistency score simply because there are few fuel-poor cases. Specifically, Equation 6 compares the *percentage* of fuel-rich cases with authoritarian governments to the *percentage* of fuel-poor cases with authoritarian rule. This contrasts with Equation 5, which compares the *number* of fuel-rich cases with authoritarian governments to the *number* of fuel-poor cases with authoritarian rule. By design, the revised measure only equals the STCM calculation when the numbers of fuel-poor and fuel-rich cases are equal. Comparing Equations 4 and 6 show that both a regression coefficient and the revised necessary condition consistency metric incorporate the sufficient condition consistency measures for both $F = 0$ and $F = 1$. The differences arise from the aggregation procedure, not from the number of cells used.

Using Equation 6 to calculate necessary condition consistency scores for Tables 2 through 4, respectively, yields 33%, 87%, and 50%. Because there are equal numbers of $F = 0$ and $F = 1$ cases in Tables 3 and 4, the estimated consistency scores from Equations 5 and 6 are identical. Thus, the revised measure does not yield misleading conclusions when applied to non-skewed data. But the revised measure estimates a sharply different consistency score for Table 2 (33% vs. STCM's 91%) by accounting for the unbalanced number of $F = 0$ and $F = 1$ cases, therefore guarding against misleading conclusions.

All Four Cells: Consistency or Triviality?

A potential counterargument to the position advocated here is that there are no inherent flaws with the existing STCM necessary condition *consistency* measure, and instead the concerns raised here apply only to the *triviality* of a necessary condition. Schneider and Wagemann (2012, pp. 235-237), for example, also discuss how differential numbers of $X = 1$ and $X = 0$ cases can create problems, but relate the concern to triviality rather than to consistency. This is not a compelling argument.

To use an illustrative example to support why skewed cases raise issues about consistency rather than triviality, consider every at-bat in Major League Baseball history as the set of cases. STCM procedures would conclude that not being Babe Ruth ($X = 0$) is a necessary condition for hitting a home run ($Y = 1$). After all, the overwhelming majority of home runs in baseball history have not been hit by Babe Ruth, that is, $Y = 1$ is nearly a perfect subset of $X = 0$.

The problem with this unwarranted conclusion does not stem from triviality. Home runs occur somewhat frequently when the posited necessary condition is present, that is, when players not named Babe Ruth come to bat.[8]

Instead, the problem with the conclusion is that the data are highly inconsistent with the hypothesis. Babe Ruth at-bats ($X = 1$) do not imply the negation of a home run ($Y = 0$), violating the contrapositive of the necessary condition hypothesis. In fact, the absence of the purported necessary condition predicts a home run with considerably *higher* probability than the presence of the condition. The STCM consistency metric misses this because the number of $X = 1$ cases is very small relative to the number of $X = 0$ cases. In contrast, a consistency measure that focuses on percentages rather than the number of cases easily avoids this problematic conclusion.

Furthermore, even those who reject the position that a skewed distribution of X cases is more closely related to consistency rather than to triviality *still* implicitly embrace the conclusion that one needs to scrutinize all four cells. The standard STCM consistency score calculation and Schneider and Wagemann's (2012, p. 236) alternate necessary condition triviality measure jointly use all

four cells.[9] This reinforces the conclusion that the only way to account for a skewed X distribution is to incorporate information from all four cells.

In sum, to assess a necessary condition hypothesis, analysts should always study cases that exhibit variation on both the posited causal condition and on the outcome. This practice, which is central to quantitative research, guards against misleading findings rather than leads analysts to "conflate" (Ragin 2000, p. 96) necessity conditions and sufficient conditions. STCM and quantitative methods share this crucial feature.

Core Similarity 2: Configurations of Conditions and Regression Interaction Terms

Regression models with multiplicative interaction terms share a core similarity with STCM analyses of complex necessary and sufficient conditions. Because the quantities used in STCM to study complex conditions can be derived from quantities estimated by an interactive regression model, the regression results necessarily contain all the same information as the STCM analysis. This finding opposes pointed STCM arguments that they offer a distinct—and superior—alternative for studying complex causal patterns.

There is wide agreement in the STCM literature that STCM provides a better alternative than regression for studying complex relationships. STCM studies configurations of conditions using a "truth table," which displays all logically possible combinations of the explanatory conditions under analysis by presenting one combination per row. A truth table provides the information needed to calculate the percentage of cases with each combination of conditions that achieves $Y = 1$, which is in turn used to compute consistency and coverage scores. In contrast, regression is purported to be unable to study complex relationships. Ragin (2008) characterizes regression as a method useful only for studying the net effects of individual conditions—in fact, one of his "four oppositions" between STCM and quantitative methods is "the analysis of causal complexity versus the analysis of net effects" (p. 9)—and he dismisses outright the ability of regression to analyze complex relationships. Ragin (2008) states that questions involving conditional relationships lie "outside the scope of conventional net-effects analyses, for those approaches are centered on the task of estimating context-independent net effects" (p. 181). Many other scholars, such as Hall (2003, pp. 382-383), Becker (2002, p. 250), and Brown (2009, p. 415) express similar concerns that regression analysis is unable to adequately analyze complex relationships.

These arguments overlook how interactive regression terms serve as a crucial bridge between quantitative and STCM approaches. When the interaction terms are properly interpreted, it becomes clear that a regression model

provides all the information needed to calculate necessary and sufficient condition consistency and coverage scores for any complex combination of conditions—despite the fact that regression analysis is best interpreted as studying the effects of particular interventions.[10] The analysis below demonstrates this point by analyzing a 2 × 2 table that involves a complex condition.[11] Because both regression and STCM can provide all the information needed to compute the number of cases in each cell of a 2 × 2 table, it follows that both approaches can be used to compute consistency and coverage scores for necessary and sufficient conditions—and therefore provide the *same* information for studying complex causal relationships.

To make the argument that truth tables and regression tables yield the same information, it is valuable to directly compare how a truth table and a regression table will display results from the same data set. The discussion focuses first on the truth table.

Truth Table: Computing the Number of Cases in Each Cell of a Complex 2 × 2 Table

Suppose there are two conditions of interest, S and F, and an outcome A. Continuing the substantive example from above, F and A refer to fuel wealth and authoritarianism, respectively. S represents strong bureaucratic institutions. Table 5, in which the first two columns correspond to a truth table, summarizes a hypothetical set of data. It presents the four logically possible combinations of strong institutions and fuel wealth, the number of cases with each combination, and the percentage of cases with each combination of conditions that experience authoritarian rule.

Table 5. Notation for a Truth Table that Incorporates the Percentage of Cases per Row With $A = 1$.

S	F	Notation for number of cases	Notation for % with A = 1
0	0	n_{sf}	p_{sf}
0	1	n_{sF}	p_{sF}
1	0	n_{Sf}	p_{Sf}
1	1	n_{SF}	p_{SF}

For each logically possible combination of S and F, this table presents notation for the number of cases in the row and the percentage of those cases with $A = 1$. Denoting a generic condition as "X," the subscript for either the number (n) or percentage (p) contains an uppercase X if $X = 1$, and a lowercase x if $X = 0$. Some STCM scholars include percentages in their truth tables (e.g., Basedau & Richter, 2014, p. 561), whereas others do not (e.g., Schneider & Wagemann, 2012, p. 106). A = authoritarian rule; S = strong bureaucratic institutions; F = fuel wealth.

Given this information, it would be natural for a STCM scholar to use QCA software to reduce subset redundancy and to calculate consistency and coverage scores. However, for the purposes of comparing STCM and regression, it is useful to perform the calculations by hand. Suppose we wanted to learn about the necessary/sufficient consistency and coverage scores for having weak institutions and fuel wealth, that is, $(S, F) = (0, 1)$.

Based on the discussion from the previous section, we know a 2 × 2 table will provide all the relevant information. Table 6 incorporates the notation from Table 5 and shows how to use the information to compute the number of cases in each cell. A complex condition can be summarized in binary form because we are only interested in its presence and its absence for computing consistency and coverage scores. If a case has strong institutions and/or lacks fuel wealth, then it does not have $(S, F) = (0, 1)$. Therefore, [$S = 1$ OR $F = 0$] negates [$S = 0$ AND $F = 1$], and composes the other column in the table.

To understand why the symbols inside each cell represent how many cases would be observed in it, it is useful to begin with cell 1. The number of cases with weak institutions and fuel wealth that also experience authoritarian rule is simply the number of [$S = 0$ AND $F = 1$] cases multiplied by the percentage of cases with this combination of conditions that have $A = 1$, which is expressed by $p_{sF} \cdot n_{sF}$. This is shown in cell 1.

The number of cases in cell 2 is calculated using an identical procedure. If a case with weak institutions and fuel wealth does not have authoritarian rule, then it must have the negation of authoritarian rule. We know that $1 - p_{sF}$ is the percentage of cases with [$S = 0$ AND $F = 1$] that achieve $A = 0$. Multiplying this percentage by the total number of weak institutions/fuel wealth cases yields $(1 - p_{sF}) \cdot n_{sF}$, as shown in cell 2.

The terms in cells 3 and 4 are somewhat more cumbersome because there are three ways to have strong institutions and/or lack fuel wealth. The number of [$S = 1$ OR $F = 0$] cases is the sum of the number of cases that have either (a) weak institutions and no fuel, (b) strong institutions and no fuel, or (c) strong institutions and fuel. For cell 3, following an identical procedure as for the single condition in cell 1 yields $p_{sf} \cdot n_{sf} + p_{Sf} \cdot n_{Sf} + p_{SF} \cdot n_{SF}$. For cell 4, following an identical procedure as for the single condition in cell 2 leads to $(1 - p_{sf}) \cdot n_{sf} + (1 - p_{Sf}) \cdot n_{Sf} + (1 - p_{SF}) \cdot n_{SF}$.

Clearly, a truth table contains the information needed to fill in a 2 × 2 table for any combination of conditions the researcher desires to study.

Regression: Computing the Number of Cases in Each Cell of a Complex 2 × 2 Table

Regression can also yield identical information for computing cells in a 2 × 2 table—and therefore for computing necessary/sufficient condition consistency

Table 6. Summarizing the Truth Table as a 2 × 2 Table With a Complex Condition.

	$S = 1$ OR $F = 0$	$S = 0$ AND $F = 1$
$A = 1$	**(3)** $p_{sf} \cdot n_{sf}$+ $p_{Sf} \cdot n_{sf}$+ $p_{SF} \cdot n_{SF}$	**(1)** $p_{sF} \cdot n_{sF}$
$A = 0$	**(4)** $(1 - p_{sf}) \cdot n_{sf}$+ $(1 - p_{Sf}) \cdot n_{sf}$+ $(1 - p_{Sf}) \cdot n_{SF}$	**(2)** $(1-p_{sF}) \cdot n_{sF}$

This table takes information from Table 5 to calculate a 2 × 2 table in which the posited causal condition is complex, with each cell numbered by a term in parentheses. The second column provides the number of cases with weak institutions and fuel wealth. The first column provides the number of cases with strong institutions and/or lack fuel wealth. See Note 3 for the definition of a "complex" condition. S = strong bureaucratic institutions; F = fuel wealth; A = authoritarian rule; p = percentage; n = number.

and coverage scores—although regression presents the information in a different form. Continuing the example from above, consider the coefficient estimates for the following model:

$$A_i = \beta_0 + \beta_S \cdot S_i + \beta_F \cdot F_i + \beta_{SF} \cdot S_i \cdot F_i + \varepsilon_i . \tag{7}$$

This model is called a *fully saturated* regression because it contains a multiplicative interaction term for each combination of variables and includes all lower order terms. Showing how ordinary least squares (OLS) estimates each of β_0, β_S, β_F, and β_{SF} reveals how a fully saturated regression model provides the same information as a truth table for computing a complex 2 × 2 table, because the four β estimates collectively include all four percentages: p_{sf}, p_{Sf}, p_{sF}, and p_{SF}. Throughout the discussion, I assume causal homogeneity. This implies that the β coefficients (which capture average effects) carry the same implications for the effect of particular interventions for each case.

The estimated constant term for the statistical model is β_0. This term tells us the average outcome when both conditions equal 0. In other words, β_0 is the percentage of cases with weak institutions and without fuel wealth that experience authoritarian rule—which by assumption is p_{sf}—and can be expressed as follows:

$$\%\left[A = 1 \,|\, (S, F) = (0,0) \right] = p_{sf} . \tag{8}$$

This is the first of the four percentages we need from the regression results to calculate consistency and coverage scores.

To understand the other three regression coefficients, it is useful to think of regression as estimating the effects of particular interventions. The first hypothetical intervention of interest is what happens if a case with weak institutions and that lacks fuel wealth is changed to have strong institutions (while continuing to be fuel-poor). The expected effect of the intervention is the average outcome when (*S, F*) = (1, 0) minus the average outcome when (*S, F*) = (0, 0):

$$\beta_S = \%\left[A = 1 \mid (S,F) = (1,0)\right] - \%\left[A = 1 \mid (S,F) = (0,0)\right] = p_{Sf} - p_{sf}. \quad (9)$$

Because β_0 tells us p_{sf}, we can add β_0 to β_S to calculate p_{Sf} from the regression results. Thus, estimating β_S provides the second of the four percentage terms that we need to calculate consistency and coverage scores.

The second hypothetical intervention of interest is what happens if a case with weak institutions and that lacks fuel wealth is changed to become fuel-rich (while continuing to have weak institutions). The expected effect of the intervention is the average outcome when (*S, F*) = (0, 1) minus the average outcome when (*S, F*) = (0, 0):

$$\beta_F = \%\left[A = 1 \mid (S,F) = (0,1)\right] - \%\left[A = 1 \mid (S,F) = (0,0)\right] = p_{sF} - p_{sf}. \quad (10)$$

We can add β_0 to β_F to calculate p_{sF} from the regression results. Therefore, estimating β_F provides the third of the four percentage terms that we need to calculate consistency and coverage scores.

The third hypothetical intervention of interest is what happens if a case with weak institutions and that lacks fuel wealth is changed to have both strong institutions and fuel wealth. The expected effect of this intervention can be expressed by the sum of three terms. The first component is the effect of only changing *S*, which is β_S. The second component is the effect of only changing *F*, which is β_F. The third component expresses the extent to which simultaneously changing both *S* and *F* differs from the sum of the effects of each individual intervention and is captured by β_{SF}. Using the notation, this means that we can calculate the effect of the third intervention as

$$\%\left[A = 1 \mid (S,F) = (1,1)\right] - \%\left[A = 1 \mid (S,F) = (0,0)\right] = \beta_S + \beta_F + \beta_{SF}. \quad (11)$$

Because %[*A* = 1 | (*S, F*) = (1, 1)] = p_{SF} and %[*A* = 1 | (*S, F*) = (0, 0)] = β_0, we can solve Equation 11 to get

$$p_{SF} = \beta_0 + \beta_S + \beta_F + \beta_{SF}. \quad (12)$$

Thus, regression also allows us to calculate all four terms needed to compute consistency and coverage scores for necessary/sufficient conditions. Table 7 re-expresses Table 6, replacing the percentage terms from the truth table with the OLS coefficient estimates from the regression model in Equation 7.

Table 7. Summarizing the Regression Table as a 2 × 2 Table With a Complex Condition.

	$S = 1$ OR $F = 0$	$S = 0$ AND $F = 1$
$A = 1$	$\beta_0 \cdot n_{sf}$+ $(\beta_0 + \beta_S) \cdot n_{Sf}$+ $(\beta_0 + \beta_S + \beta_F + \beta_{SF}) \cdot n_{SF}$	$(\beta_0 + \beta_F) \cdot n_{sF}$
$A = 0$	$(1 - \beta_0) \cdot n_{sf}$+ $[1 - (\beta_0 + \beta_S)] \cdot n_{Sf}$+ $[1-(\beta_0 + \beta_S + \beta_F + \beta_{SF})] \cdot n_{SF}$	$[1-(\beta_0 + \beta_F)] \cdot n_{sF}$

This table replaces the percentage terms from the truth table with the coefficient estimates from the regression model in Equation 7. *S* = strong bureaucratic institutions; *F* = fuel wealth; *A* = authoritarian rule; *n* = number.

A closer examination of the β_{SF} coefficient also refutes the core STCM claim that conventional quantitative methods can only be used to estimate "context-independent net effects" (Ragin 2008, p. 181). Suppose fuel wealth promotes authoritarianism when institutions are weak, but exerts no effect on regime type when institutions are strong. In addition, suppose institutional quality does not exert an unconditional effect on authoritarianism and instead only modifies the effect of fuel wealth. Regression would reveal this relationship because β_F would be positive, β_S would be 0, and β_{SF} would equal $-\beta_F$. The negative β_{SF} captures the fact that fuel wealth positively affects authoritarianism when institutions are weak, but does not affect the baseline probability of authoritarian rule when institutions are strong. Thus, for a case that originally has weak institutions and no fuel wealth, only one of the three possible combinations of changing either or both the two conditions will alter the baseline probability of authoritarianism: changing fuel wealth but not institutions, because $\beta_F > 0$. Changing institutions but not fuel wealth has no effect because $\beta_S = 0$. Changing both conditions also has no effect: $\beta_F + \beta_S + \beta_{SF} = \beta_F - \beta_F = 0$. These findings are possible precisely because estimating a fully saturated regression assesses whether the relationships are conditional rather than constant.

In sum, multiplicative interaction terms from regression analysis can be used to compute consistency and coverage scores for complex necessary and

sufficient condition relationships, and regression is not forced to assume constant effects—belying the strong claims from the STCM literature that regression is inherently unable to study causal complexity. Appendix B provides an additional example that features multiple paths to the outcome. The fact that the regression table presents the information differently than does a truth table should not obscure their inherent similarities. Furthermore, multiplicative interaction terms represent only one of many quantitative techniques for analyzing complex causality, with additional possibilities discussed in the conclusion.[12]

Shared Challenges: From Association to Causation

Researchers who acknowledge these core similarities will not be surprised that STCM and quantitative methods also share common limitations. Drawing *causal* conclusions from *associational* relations poses great difficulties—regardless of how many cells of a 2 × 2 table are used, and regardless of whether STCM or regression is used to examine complex relationships. Whereas current best-practice quantitative textbooks and research focuses centrally on this challenge, it has received less attention from STCM scholars.[13] Below I posit a best-case scenario in which the data are perfectly consistent with a necessary/sufficient hypothesis. This section focuses on two issues that must be addressed before a compelling causal inference can be achieved: Freedman's (2010) empirical standard for comparing a deterministic hypothesis to a probabilistic benchmark and Waldner's (2005) standard for evaluating hypothetical counterexamples.

Pointed critiques of conventional quantitative methods have led to a dramatic rethinking of how cross-case comparisons can be translated into convincing causal claims. Quantitative methodologists have responded by producing best-practice advice that focuses centrally on causal inference, whether for analyzing field experiments (Gerber & Green, 2013), "natural" experiments (Dunning, 2012), or observational data (Morgan & Winship, 2007; Rosenbaum, 2002).

Although STCM scholars have also discussed causal inference issues (e.g., Mahoney et al., 2013), this section focuses on two issues specifically pertaining to deterministic inferences that require concerted attention. First, even if the data are perfectly consistent with a necessary or sufficient condition hypothesis, it is useful to consider the likelihood the data could have been produced by an alternative probabilistic process. Applying this consideration—proposed by Freedman (2010)—demonstrates the generic difficulty of making convincing deterministic inferences using empirical evidence alone and the particular difficulties that arise with a small number of cases.

Furthermore, the value-added of stating a hypothesis as necessary or sufficient—as opposed to expressing a parallel probabilistic claim—is questionable without scrutinizing why it is reasonable to believe that *Y cannot* occur if a posited necessary condition *X* is absent, or that *Y must* occur if a posited sufficient condition *X* is present. I therefore also consider Waldner's (2005) standard for using hypothetical counterfactuals to evaluate deterministic hypotheses. Combining the empirical and theoretical approaches may suggest an avenue for creating stringent standards that more adequately assess deterministic hypotheses. Finally, because STCM scholars have raised important considerations regarding whether necessity/sufficiency claims are inherently deterministic, the end of the section engages this contentious topic.

Overall, scrutinizing the causal underpinnings of STCM hypotheses reveals the need for similar advancements as those that have improved quantitative research in recent decades.

Data-Generating Process (DGP) Vis-à-Vis Data

To distinguish associational patterns from causal inferences it is crucial to introduce the concept of a data-generating process (DGP) and to distinguish the DGP from the data actually observed. Suppose we have a best-case scenario for a deterministic hypothesis in which the cases are perfectly consistent with $X = 1$ being both necessary and sufficient for $Y = 1$. That is, every $X = 1$ case achieves $Y = 1$ whereas no $X = 0$ cases have $Y = 1$. If a scholar infers that X is in fact sufficient for Y from this associational pattern, they have implicitly made the following claim about the DGP:

$$\Pr\left(Y = 1 \mid X = 1\right) = 1. \tag{13}$$

In words, *Y must* occur if *X* is present. Therefore, any attempts to infer sufficiency must scrutinize how convincing the "must" component of their causal claim is.

Similarly, moving from association to inferring necessity implies:

$$\Pr\left(Y = 1 \mid X = 0\right) = 0.^{14} \tag{14}$$

In words, *Y cannot* occur if *X* is not present. Therefore, any attempts to infer necessity must scrutinize how convincing the "cannot" component of their causal claim is.

Crucially, we observe draws from the DGP but do not observe the DGP itself. Instead, we must impose assumptions to make inferences about the

DGP. The remainder of the section discusses standards for inferring that a DGP is deterministic.

Freedman's Standard for Assessing Probabilistic Alternatives

The data may appear to strongly support a necessary/sufficient condition hypothesis when there are no empirical counterexamples. However, even in this ideal circumstance an empirical approach to evaluating deterministic hypotheses can produce misleading conclusions when scholars do not carefully consider probabilistic alternatives that may be generating the observed data. It is never possible to fully disentangle deterministic from probabilistic alternatives solely on the basis of empirical observation[15]—an especially pressing concern when the number of cases is small. The following discussion elaborates upon Freedman's (2010) standards for comparing deterministic hypotheses to probabilistic alternatives, using an example from applied STCM to substantiate the importance of this consideration.

Mahoney (2010) claims that, for Spanish America, the combination of (a) lack of a strong colonial legacy and (b) prolonged warfare during the 19th century without a major victory (jointly denoted as $X = 1$) was sufficient for low levels of economic development ($Y = 0$). Four countries in his study possess this set of conditions, and all four have low levels of economic development.[16]

The pattern that Mahoney uncovers is fully consistent with a claim that the DGP involves $\Pr(Y = 0 \mid X = 1) = 1$.[17] However, suppose instead the DGP for $X = 1$ cases produced $Y = 0$ or $Y = 1$ with equal probability, that is, $\Pr(Y = 0 \mid X = 1) = \Pr(Y = 1 \mid X = 1) = .5$. We would *still* be fairly likely to observe four $X = 1$ cases achieve the same outcome (either all $Y = 0$ or all $Y = 1$), hence portraying a deterministic pattern. To create a best-case scenario for evaluating the sufficient condition hypothesis, assume the same DGP governs all four cases. This can be conceptualized as a large urn in which 50% of the balls are yellow and 50% are red, that is, Pr(yellow ball | posited SC is present) = Pr(red ball | posited SC is present) = .5. If we take four independent draws from the large urn, there is a $2 \times (0.5)^4 = 12.5\%$ chance of observing either all red or all yellow balls. Therefore, one out of every eight possible arrays of cases from the population of balls will be perfectly consistent with a deterministic process, even though this hypothetical DGP actually produces either type of ball with *equal* probability.

As an additional consideration, assuming the urn produces each type of ball with equal probability is particularly stringent for establishing evidence in favor of a probabilistic alternative. Suppose we instead assume that 90% of the balls are red and 10% are yellow: Pr(red ball | posited SC is present) = .9.

If we take four independent draws, there is a 65.5% chance every ball will have the same color. This alternative possible DGP favors red over yellow balls, but also is not deterministic. Hence, observing that all four cases sharing a particular trait also achieve the same outcome provides weak evidence in favor of a deterministic hypothesis. One could reach an identical conclusion about necessity if they had four cases that lacked the posited necessary condition, and all four of these cases failed to achieve the outcome.

Probabilistic alternatives can never be empirically rejected unless we draw an infinite number of balls from the data-generating urn. For example, suppose the large urn produces 99% red balls and 1% yellow balls. If we take 298 independent draws from the urn, there is still a slightly higher than 5% chance that no yellow balls will be observed. While the 5% threshold is arbitrary, it is notable because it is a commonly used statistical threshold for rejecting a null hypothesis of no effect. In the 99-1 urn example, different analysts could conceivably disagree on whether the 298 red draws provide strong evidence in favor of the deterministic hypothesis.[18]

For the present discussion, the main takeaway is that observing no deviant cases yields little support for a deterministic claim when there are few cases—regardless of how one states their probabilistic alternative. This is troubling because, at least in comparative politics, deterministic claims are usually made when analyzing few cases. In addition, because probabilistic alternatives can never be rejected solely on the basis of empirical observations, we should always use additional standards to evaluate a deterministic hypothesis.

Waldner's Standard for Evaluating Hypothetical Counterexamples

Waldner (2005, p. 28) argues that scholars should not proclaim strong support for a necessary condition hypothesis—even if there are no empirical counterexamples—without scrutinizing why *Y cannot* occur when *X* is absent. Even if no alternate paths to *Y* are observed empirically in the set of cases, one still needs to evaluate hypothetical counterexamples to assess the plausibility of the *cannot* claim. A similar consideration calls into question claims of sufficiency absent a compelling argument that *Y must* occur when *X* is present. If a combination of conditions is sufficient for an outcome, then it should be difficult to posit a plausible hypothetical scenario in which the outcome could fail to occur when those conditions are present. Thus, scrutinizing hypothetically possible scenarios composes an important component of evaluating a hypothesis that the DGP is deterministic.

Continuing the substantive example from above, the theory supporting the deterministic elements of Mahoney's (2010) claims is likewise not very convincing. Importantly, Mahoney (2010) *does* provide compelling theoretical justifications for why the conditions he studies may have greatly increased the probability a case would achieve one outcome or another. However, his theoretical discussion does not convincingly ground the "must" claim inherent in his sufficient condition hypotheses nor the "cannot" claim inherent in his posited necessary conditions.

One of Mahoney's (2010) complex sufficient condition hypotheses states that a weak tradition of mercantilist colonial institutions AND the presence of established liberal colonial institutions are jointly sufficient for high levels of development. Problematically for the claim, it is relatively easy to construct a hypothetical possibility that strongly suggests Pr (high development | posited SC is present) < 1, in contrast to Mahoney's implicit inference that Pr (high development | posited SC is present) = 1. Consider the out-of-region case Zimbabwe. This case illuminates how a country that possessed the conditions posited to be sufficient could have failed to achieve high development. Zimbabwe was imbued with British parliamentary institutions and a nascent industrial structure at independence but has been a developmental disaster—partly because the African majority government lashed out against the European settlers that originally established what are presumed to be "good" institutions.[19]

The point here is not to imply that Mahoney's argument should be evaluated in terms of how it can be generalized. Rather, Zimbabwe functions as a counterfactual consideration for the Spanish American cases Mahoney does consider. Without a strong argument that a trajectory resembling Zimbabwe's was not hypothetically possible for any Spanish American countries that possessed the purported sufficient conditions for prosperity, we have serious grounds for questioning the deterministic "must" element of Mahoney's (2010) hypothesis—which, again, is distinct from the issue of whether Mahoney analyzed conditions that strongly increased the likelihood of a particular outcome.

In an example involving a necessary condition, Mahoney (2010) claims that possessing either of the following two complex conditions was necessary for achieving high economic development among the countries he analyzes: (NC #1) being neglected by the Habsburgs AND being either a colonial center or semiperiphery during the Bourbon era OR (NC #2) being neglected during both colonial eras AND victorious in warfare during the 19th century. As with the Zimbabwe example, examining out-of-region cases illuminates the possibilities that must be considered to strongly establish a combination of conditions as necessary for an outcome. It is useful to observe that Brunei,

Qatar, and the United Arab Emirates all scored in the highest development category in the United Nations' Human Development Index in 2013, despite not inheriting strong institutions under British colonization nor achieving victory in warfare. Large oil reserves in these countries have provided an exceptional source of revenue that fall outside the standard "development" process. Thus, there exist conditions that could plausibly lead to development that do not require either Mahoney's NC #1 or NC #2.

Crucially, the argument that an exceptional source of revenue can engender high levels of overall wealth—regardless of other development impediments faced by the country—is plausible for the countries Mahoney (2010) analyzes. In fact, at least until the 1980s, Venezuela provides within-region evidence of this assertion. Propelled by oil revenues, between 1920 and 1980 Venezuela was one of the fastest growing countries in Latin America (Hausmann, 2003, p. 245). This conclusion suggests Pr(high development | ¬ NC #1 AND ¬ NC #2) > 0, violating the "cannot" element of the necessary condition claim.

One way of revising the necessary condition hypothesis in response to considerations raised by these out-of-region cases would be to posit an additional qualification. The argument of necessity may hold, provided that the countries of concern do not have some exceptional source of revenue. However, then the worry arises that this qualification is simply another path to the outcome—which immediately changes the scope conditions for the original argument. Importantly, STCM scholars *are* very careful in thinking about scope conditions, as exemplified in Chapter 16 of Goertz and Mahoney (2012). However, claims of necessity and sufficiency require scrutinizing additional scope conditions about hypothetical possibilities that are not observed among the cases under inquiry, a topic STCM scholars have not addressed.

Toward a More Demanding Standard? Combining the Approaches

A further example, which focuses on physical impediments, suggests a possible avenue for creating viable standards. Here, Freedman's (2010) and Waldner's (2005) criteria for necessary/sufficient condition hypotheses are combined, resulting in a more demanding standard.

Mann's (1993) discussion of the effects of the Industrial Revolution on economic changes in Europe posits a necessary condition hypothesis with more compelling deterministic underpinnings than the examples considered above. He states that between 1760 and 1914 there was a "truly exponential

transformation in the logistics of collective power" (Mann, 1993, p. 12, emphasis added). The claim implicitly engages hypothetical counterexamples by suggesting that the outcome could not occur without this transformation. Mann (1993) additionally argues that transportation infrastructure, economic growth, and military capabilities exceeded "*all known* historical rhythms" (p. 13, emphasis added), a reference to empirical observations.

The hypothesis that industrialization was necessary for the changes Mann describes is supported empirically. Over the course of thousands of years, countless civilizations had existed without a single counterexample to the hypothesis. But what distinguishes the necessary condition claim as a compelling deterministic hypothesis—as opposed to a parallel probabilistic claim—is that physical constraints on human production made it nearly impossible for any earlier society to reach the per capita income levels achieved by several 19th-century European countries. Therefore, in the absence of the posited necessary condition, it is difficult to conceive of alternate paths to the outcome for countries prior to the 19th century. The combination of empirical evidence and hypothetical considerations suggests relatively strong evidence for the deterministic hypothesis.

Are Necessary/Sufficient Condition Hypotheses Deterministic?

Whereas the conclusions drawn from the two preceding sections apply regardless of the ontological status of necessary/sufficient condition hypotheses, the considerations levied in this section apply specifically to *deterministic* hypotheses. This is an important distinction because STCM scholars reject the standard position that necessity and sufficiency are inherently deterministic propositions (e.g., Collier, Brady, & Seawright, 2010, p. 145), and instead prefer the term *asymmetric*. As examples, Goertz (2005) argues it does not matter whether necessary condition hypotheses are treated as deterministic or probabilistic, and Ragin frequently appeals to notions of "almost always" sufficiency (e.g., Ragin 2008, p. 49). Thus, it is important to address why STCM should be considered a deterministic method. Although STCM scholars have raised valuable points, they have yet to convincingly address critiques levied by scholars from diverse backgrounds.

Braumoeller and Goertz (2000) have articulated the clearest perspective on why it may still be useful to think in terms of necessity and sufficiency when the data are not perfectly consistent with a deterministic hypothesis, by appealing to measurement error. However, their defense does not invalidate the concerns presented in this section. Even if a data set contains empirical counterexamples and the researcher imposes the permissive assumption that *all* the deviant cases arose from measurement error, it is still useful to assess

the likelihood of observing the non-deviant observations under different probabilistic DGPs. Furthermore, the theoretical basis of the "must" or "cannot" claims should still be carefully scrutinized.

The standards presented here would need to be altered to accommodate determinism-with-exception positions that do not rely on measurement error. However, in their current form, alternative determinism-with-exception arguments face at least three main shortcomings. First, if we strip the "cannot" and "must" components of necessary/sufficient condition hypotheses, Waldner (2005) argues it is not clear what value-added these concepts retain. Any thresholds used for assessing "necessary enough" or "sufficient enough" will be arbitrary, and it is not entirely clear how to distinguish such claims from ones that would be made using quantitative techniques:

> If stating that $P = .99$ permits us to claim that the relationship is "extremely likely," while $P = .51$ prompts us to claim that X is "necessary more often than not," then there is no non-arbitrary reason to prohibit me from claiming that when $P = .25$, X is "hardly ever necessary" for Y and when $P = .01$, X is "virtually never necessary" for Y. All of these would count as necessary condition hypotheses. (Waldner, 2005, p. 28)

Freedman (2010) makes a similar point in reference to whether outcomes can occur in the absence of almost necessary conditions: "'Impossibility' might just mean that the likelihood is below the cutpoint of 0.5 . . . Selecting cutpoints is another famous problem" (p. 108).

Second, if STCM scholars want necessary/sufficient condition hypotheses to be evaluated as probabilistic, this choice requires explicitly modeling a stochastic component (Sekhon, 2005). None of the models presented in Ragin (2008), Goertz and Mahoney (2012), or Schneider and Wagemann (2012) contain a stochastic component. This implies the assumed causal structure is deterministic and renders as ambiguous the exact interpretation of "almost always" necessity or sufficiency.

Third, those who reject the standards presented in this section because they are only appropriate for deterministic hypotheses must replace them with viable alternatives. Without further steps, one cannot accept a claim that the DGP "almost always" produces $Y = 1$ when $X = 1$ is present—which entails an inference about the causal process—any more readily than one can accept that X increases the probability of Y simply by observing a positive regression coefficient linking the two. Expanding on the hypothetical urn examples from above, perhaps an almost-always sufficiency claim entails Pr(outcome | posited "almost always" SC is present) $\geq .9$. What assumptions should we make to infer this process generated the data? Should we retain a

weaker version of the "must" claim from a sufficiency hypothesis—which, again, is implied directly by the definition of a sufficient condition? Alternatively, a claim of "almost always" sufficiency could imply Pr(outcome | posited "almost always" SC is present) = 1 for 90% of the cases analyzed but not the other 10%. If this is indeed what the analyst means, then all the standards proposed above apply fully for 90% of the cases—plus the additional requirement for the researcher to characterize which cases fit the scope condition for the sufficiency hypothesis.

Inferring any type of causal relationship is difficult with any method, and certainly none of this discussion presumes agreement with prominent quantitative scholars who argue that deterministic arguments are inherently uninteresting or untestable (Clark, Gilligan, & Golder, 2006; Sekhon, 2004). However, a skeptical note is warranted. Without systematically employing demanding standards—perhaps similar to the ones discussed here or perhaps others—there is no reason to believe that associational patterns consistent with necessity or sufficiency allow us to infer the DGP is indeed deterministic. The stringent association-causation critiques levied on conventional quantitative methods cannot be dismissed in STCM. Expounding the similarities between STCM and quantitative methods clearly demonstrates the shared challenges for making compelling causal claims. This reveals vital considerations for future STCM research.

Conclusion

Set-theoretic comparative methods (STCM) have crafted "a language for all those scholars that do not feel at ease with applying statistical principles and practices to their research" (Schneider & Wagemann, 2013a, p. 5). This article argues that while the language may well be distinctive, the research procedures are more similar than has been recognized. In many regards, STCM share common foundations with quantitative research. Comprehending these areas of overlap not only helps scholars from different research traditions to better understand each other's research but also highlights shared inferential challenges faced by both STCM and quantitative methods.

The arguments of shared foundations offered here raise three questions for future research. First, how strong are the common foundations in non-binary versions of STCM? Second, in addressing the shared goal of analyzing interactions and causal complexity, what can be learned by using tools employed in recent quantitative work? Third, to what extent is STCM an advance over conventional qualitative methods?

First, this article focuses on the binary version of STCM. Although it will be valuable for future work to evaluate possible extensions of shared

foundations to non-binary (fuzzy-set and multi-value) versions of STCM, it is important to point out why this discussion has focused on the binary version. This has been done (a) in the interest of expositional clarity; (b) because researchers outside the STCM tradition routinely conceive of necessary/sufficient conditions as inherently dichotomous concepts, implying the binary version commands wide interest; and (c) STCM scholars commonly focus on the 2 × 2 table to make their crucial argument that standard statistical ideas of correlation or association do not apply to STCM.

To the extent that divergences between STCM and quantitative methods emerge outside the binary setting, it will be important to (a) understand the circumstances in which STCM assumptions about taking the maximum or minimum of fuzzy sets should be preferred over alternative assumptions and (b) for STCM to address recent critiques of aggregating fuzzy sets to evaluate necessity and sufficiency (Braumoeller, 2013; Dunning, 2013).

Second, with regard to causal complexity, recent work on quantitative methods has devoted considerable attention to this topic. Kam and Franzese (2007) carefully review discussions that date back to the 1970s, underscoring obstacles to analyzing interactions and clarifying standards for applied work. Imai, Keele, Tingley, and Yamamoto (2011) and Glynn (2012) focus on causal mediation effects, and Hainmueller and Hazlett (2014) present a least squares method that relaxes linearity and additivity assumptions. Perhaps most directly related to STCM, Grimmer, Messing, and Westwood (2014) discuss machine learning techniques for analyzing complex relationships. Especially considering that many recent simulation studies demonstrate QCA algorithmic results are unreliable under various minor perturbations (Hug, 2013; Krogslund, Choi, & Poertner, 2014; Lucas & Szatrowski, 2014; Seawright, 2014b), it will be valuable for STCM scholars to address how their techniques relate to these recent quantitative advances.

Third, the argument that STCM is an improvement over traditional *qualitative* methods is questionable. Work such as Goertz and Mahoney (2012) claims that because qualitative researchers routinely state hypotheses that either explicitly or implicitly reference necessary and sufficient conditions, most qualitative research should be recast in set-theoretic terms. But, as shown above, the ideas of subsets and supersets that are so important for STCM work can lead to misleading conclusions about associational relationships when the distribution of cases is skewed. Furthermore, STCM has not convincingly addressed the recent rethinking of causal inference that is so fundamental to the quantitative field. It is against the backdrop of this deficit that studies such as Mahoney (2010) stand out for meticulously using process tracing and structured case comparisons, which are staples of traditional qualitative methods. In this regard, new work on combining STCM with

process tracing (Rohlfing & Schneider, 2013; Schneider & Rohlfing, 2013) also has great merit. Yet in these valuable studies, the distinctive contribution of the STCM component—as opposed to the contribution of traditional qualitative tools—still needs to be demonstrated.

In sum, future research on comparative methods will be well-served by (a) exploring the nature and scope of potential shared foundations of the nonbinary versions of STCM and conventional quantitative methods; (b) examining the implications for STCM of new quantitative work on complex relationships; and (c) parsing out—in new work by STCM scholars that seeks to incorporate traditional qualitative tools—the actual value-added of the STCM component.

Appendix A

Assessing Sufficient Condition Triviality With Skewed Cases

The section on the first core similarity focused on problems with the STCM necessary condition consistency metric. The STCM procedure for evaluating the triviality of a sufficient condition hypothesis faces a similar limitation. Consider the hypothetical example in Table A1.

Table A1. Hypothetical Example of Assessing Sufficient Condition Triviality With Skewed Cases.

	$F = 0$	$F = 1$
$A = 1$	2	20
$A = 0$	0	0

F = fuel wealth; A = authoritarian rule.

In this 2 × 2 table, the STCM consistency score for $F = 1$ as a sufficient condition for $A = 1$ is 1. This is a very sensible conclusion about an associational pattern based on the data at hand—although, as discussed in the section on shared challenges, additional steps remain to achieve a compelling inference about the fundamental "must" claim.

But the STCM sufficient condition coverage score (identical to the STCM necessary condition consistency score presented in Equation 5) will lead to the misleading conclusion that $F = 1$ is not a trivial sufficient condition for $A = 1$. As can be seen by examining the $F = 0$ cases, every $F = 0$ case also achieves $A = 1$. But because conventional STCM metrics only incorporate two cells at a time and therefore do not consider imbalance among $F = 0$ and

$F = 1$ cases, the STCM sufficient condition coverage score equals 20 / 22 = 91%. In contrast, Equation 6 corrects this problem and reports a coverage score of 1 / (1 + 1) = 50%. This revised calculation accurately reports that $F = 1$ is a trivial sufficient condition for $A = 1$.

Appendix B

An Example of Studying Complex and Equifinal Relationships With Regression

Extending the discussion from the section on the second core similarity, the following numerical example features equifinality to further clarify how regression can be used to study complex relationships. An equifinal causal process implies that for at least one outcome, there is more than one grouping of conditions that has a high sufficient condition consistency score, which I refer to below as "paths." In other words, there are multiple paths to certain outcomes.

Levitsky and Way (2010) examine the causes of stable authoritarianism (A) among competitive authoritarian regimes. They focus on three conditions: the strength of Western linkage (W), the level of organizational power (O), and the strength of Western leverage (L). They claim there are two paths to stable authoritarianism: (W, O) = (0, 1) and (W, L) = (0, 0). They also claim there are two paths to the absence of stable authoritarianism: $W = 1$ and (O, L) = (0, 1). I represent their claims with hypothetical data, assuming that cases with conditions representing a particular path achieve that outcome 90% of the time. Table B1 presents the coefficient estimates that result from estimating a fully saturated regression model using ordinary least squares (OLS).

The following discussion explains how each of the coefficients are estimated and shows how using information from the regression table accurately captures a country's pathway—that is, its probability of stable authoritarianism—after particular hypothetical interventions. As above, I assume causal homogeneity.

Constant. Because (W, L) = (0, 0) is a pathway to authoritarianism, low levels of all three conditions lead to $A = 1$ in 90% of cases. Thus, the constant term $\beta_0 = .9$.

Linkage. The hypothetical intervention in which a country that originally lacked all three conditions gains high Western linkage changes its path from stable authoritarianism to unstable authoritarianism. High Western linkage

Table B1. Regression Table for Hypothetical Representation of Levitsky and Way (2010).

	Dependent variable is authoritarian stability
Constant (β_0)	.9
Linkage (β_W)	−.8
Organizational power (β_O)	0
Leverage (β_L)	−.8
Linkage × Organizational power (β_{WO})	0
Linkage × Leverage (β_{WL})	.8
Organizational power × Leverage (β_{OL})	.8
Linkage × Organizational power × Leverage (β_{WOL})	−.8

(regardless of the value of other conditions) is a path to unstable authoritarianism. Hence, this intervention lowers the probability of stable authoritarianism by .8 and $\beta_0 + \beta_W = .1$.

Organizational power. The hypothetical intervention in which a country that originally lacked all three conditions gains high organizational power does not change the expected outcome. Both with and without the intervention, the country is on a path to stable authoritarianism: $\beta_0 + \beta_O = .9$.

Leverage. The hypothetical intervention in which a country that originally lacked all three conditions gains high leverage changes its path from stable authoritarianism to unstable authoritarianism. High leverage is a path to unstable authoritarianism when organizational power is low: $\beta_0 + \beta_L = .1$.

Linkage × Organizational power. The hypothetical intervention in which a country that originally lacked all three conditions simultaneously gains both high Western linkage and high organizational power has the same effect as the hypothetical intervention in which a country that originally lacked all three conditions gains only high Western linkage. High Western linkage (regardless of the value of other conditions) is a path to non-stable authoritarianism: $\beta_0 + \beta_W + \beta_O + \beta_{WO} = .1$.

Linkage × Leverage. The hypothetical intervention in which a country that originally lacked all three conditions simultaneously gains both high Western linkage and high leverage has the same effect as either individual intervention. As discussed above, either individual intervention switches a country

that originally lacked all three conditions from a stable authoritarianism path to an unstable authoritarianism path. Thus, β_{WL} is positive to counteract that both β_W and β_L are negative: $\beta_0 + \beta_W + \beta_L + \beta_{WL} = .1$.

Organizational power × Leverage. The hypothetical intervention of high organizational power counteracts the negative impact of the high leverage intervention. As discussed above, manipulating only leverage for a country that originally lacked all three conditions would switch its path from stable authoritarianism to unstable authoritarianism. However, additionally manipulating organizational power means the country remains on a stable authoritarianism path. Hence, β_{OL} is positive to counteract the negative β_L term. $\beta_0 + \beta_O + \beta_L + \beta_{OL} = .9$.

Linkage × Organizational power × Leverage. The hypothetical intervention in which a country that originally lacked all three conditions gains all three conditions has the same effect as the hypothetical intervention in which a country that originally lacked all three conditions gains only high Western linkage. High Western linkage (regardless of the value of other conditions) is a path to non-stable authoritarianism: $\beta_0 + \beta_W + \beta_O + \beta_L + \beta_{WO} + \beta_{WL} + \beta_{OL} + \beta_{WOL} = .1$.

Acknowledgment

The author thanks David Collier for tireless assistance and feedback. The article has also benefited greatly from discussions with and comments by Danny Choi, Kevin Clarke, Ruth Berins Collier, Thad Dunning, Zachary Elkins, Kenneth Greene, Simon Hug, Chris Krogslund, Marcus Kurtz, Sebastian Mazzuca, Katherine Michel, Gerardo Munck, Mathias Poertner, Roxanna Ramzipoor, Ingo Rohlfing, Jason Seawright, Sean Tanner, Guadalupe Tuñón, Kim Twist, Alison Varney, four anonymous reviewers, and editor Ben Ansell.

Author's Note

Any remaining mistakes are the sole responsibility of the author.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Notes

1. It is rare to find either STCM or quantitative articles that solely evaluate the relationship between a single binary condition and a binary outcome. However, STCM texts frequently use this setting to argue for fundamental differences. To directly compare arguments, I also examine a 2 × 2 setting in the analysis below.
2. The presence of deviant cases are not sufficient to eliminate the necessary condition hypothesis if one accepts STCM arguments that measurement error must be taken into account when analyzing necessary and sufficient relationships among social scientific data.
3. A complex relationship is one in which multiple conditions must be combined to be either necessary or sufficient. Mahoney (2008) provides a useful discussion of INUS conditions ("insufficient but necessary part of a condition which is itself unnecessary but sufficient for the result," p. 418) and SUIN conditions ("sufficient but unnecessary part of a factor that is insufficient but necessary for an outcome," p. 419), the core concepts behind complex necessary/sufficient conditions.
4. This is not true if one accepts a single counterexample to invalidate a claim of necessity of sufficiency. In that case, only a single cell is relevant for the hypothesis. However, as stated above, STCM scholars frequently analyze data sets with one or more counterexamples, to which the following argument applies.
5. Of course, researchers rarely confront a "fixed" set of cases. There is considerable scholarship in political science on how researchers should select cases to collect data for, which the present analysis does not contribute to. The argument here focuses on the next step in the research process and shows why—assuming a researcher already has data for all four cells—researchers should indeed examine all four cells from this fixed set of data. This contrasts Ragin's (2008) and Goertz and Mahoney's (2012) arguments about numerous examples of 2 × 2 tables they present: Scholars should ignore certain cells from their fixed set of cases.
6. The conventional necessary condition coverage score is identical to the sufficient condition consistency score presented in Equation 1. Schneider and Wagemann (2012, p. 236) present an alternate triviality measure that I discuss below.
7. This proposal for an alternative measure to compute necessary condition consistency scores should be primarily viewed as *illustrative* for demonstrating the importance of considering all four cells. Hopefully this equation will be useful as a starting point in future work for determining the best way to incorporate the concerns raised in this section.
8. Comparing the example presented above of oxygen as a necessary condition for social revolution—which exemplifies a trivial necessary condition—to the Babe Ruth example reveals important differences. The distribution of $X = 0$ and $X = 1$ cells is not particularly important in the former because there are no cases without oxygen that experience social revolution. Instead, the problem is that the presence of oxygen offers almost no predictive power for whether a social revolution will occur. In the Babe Ruth example, the skewed distribution of $X =$

0 and $X = 1$ cases completely drives the misleading conclusion. In this example, the absence of the necessary condition predicts a home run with considerably *higher* probability than the presence of the condition—in contrast to the absence of oxygen.

9. Using the notation from Table 1, Schneider and Wagemann's proposed necessity relevance measure equals $(n_{fa} + n_{fA}) / (n_{fa} + n_{fA} + n_{Fa})$. Juxtaposing this term with Equation 5 demonstrates that all four cells are used when evaluating both the consistency and triviality of a necessary condition hypothesis.
10. Of course, this interpretation of regression results requires that the conditions of interest either have been manipulated by the researcher or are at least hypothetically manipulable. It is difficult to attach a causal interpretation to conditions that do not possess this crucial property (Gerring, 2012, p. 207-212)—regardless of whether one thinks in terms of average effects or non-trivial necessary/sufficient conditions.
11. As an example of how a complex condition can be studied using a 2 × 2 table, suppose the posited necessary/sufficient condition is [$S = 0$ AND $F = 1$]. Cases with this combination of conditions would be listed in one column of the table (with one cell for $Y = 1$ and one for $Y = 0$), whereas cases possessing the negation of [$S = 0$ AND $F = 1$] would compose the other column in the table.
12. Importantly, the discussion above assumes that all logically possible combinations of conditions are empirically observed. In STCM terms, this means that the data do not exhibit "limited diversity." In contrast, when limited diversity is present, regression cannot compute a coefficient for each interactive and lower order term. Both STCM and regression are forced to use strong and unverifiable assumptions to make inferences about the data when logically possible combinations of conditions are not empirically observed. Possibly an important avenue for future research would be to compare the assumptions imposed by each method and to assess circumstances in which one should be preferred over another. Schneider and Wagemann (2013b) provide a recent contribution that focuses on the STCM approach to handling limited diversity.
13. Collier (2014) provides a similar argument.
14. Equation 14 states the contrapositive of a necessary condition hypothesis in probability terms. It is logically equivalent to state the direct definition in probability terms: $\Pr(X = 1 \mid Y = 1) = 1$. However, it does not make sense to condition on Y when modeling a causal process because X must occur temporally prior to Y for X to be a cause of Y.
15. A similar concern applies to quantitative results: Even strong correlations should not be accepted as causal without a plausible research design or other supporting evidence for the hypothesis.
16. The *only* issue raised here about Mahoney's (2010) contribution relates to his claims that he has identified and found strong evidence that a particular set of conditions is necessary and sufficient for his outcomes. He presents a nuanced historical framework accompanied by careful process tracing. As shown here, however, these commendable attributes of his research do not strongly support his claims of *necessity* or *sufficiency*.

17. To avoid confusion, note that Equation 13 states the data-generating process (DGP) if $X = 1$ is sufficient for $Y = 1$, whereas the Mahoney example focuses on whether $X = 1$ is sufficient for $Y = 0$. In addition, the "test" proposed in this section is intended to *illustrate* the core point about the importance of probabilistic alternatives, as opposed to suggesting that this exact test should be used for assessing probabilistic alternatives. There are many possible ways to model a probabilistic DGP, and additional work is needed to scrutinize the most appropriate tests for assessing probabilistic alternatives.
18. Freedman (2010) argues empirical data will never produce strong evidence for a deterministic hypothesis. Using the term "population" instead of "data-generating process" and referring to observed draws from the population as a "sample," he asks what we can conclude about observations with trait *U* if the sample does not include any observations with this trait. "If the fraction of *U*'s in the sample is small, that proves *U* is rare in the population (modulo the usual qualifications). However, unless we make further assumptions, it is impossible to demonstrate by sampling theory that there are no *U*'s in the population" (Freedman, 2010, pp. 110-111).
19. See Good (1976, p. 605) for evidence on colonial economic development, and Compagnon (2011, ch. 6) for evidence on post-independence institutional dismantling.

References

Basedau, M., & Richter, T. (2014). Why do some oil exporters experience civil war but others do not? Investigating the conditional effects of oil. *European Political Science Review*, *6*, 549-574.

Becker, H. S. (2002). Comment. *Contemporary Sociology*, *31*(2), 250.

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

Braumoeller, B. F., & Goertz, G. (2000). The methodology of necessary conditions. *American Journal of Political Science*, *44*, 844-858.

Braumoeller, B. F., & Goertz, G. (2002). Watching your posterior: Comment on Seawright. *Political Analysis*, *10*, 198-203.

Braumoeller, B. (2013, August 29 - September 1). *Aggregation Bias and the Analysis of Sufficient Conditions in fs/QCA*. Presented at the 109th annual convention of the American Political Science Association, Chicago, Illinois.

Brown, D. K. (2009). Review of "redesigning social inquiry: Fuzzy sets and beyond by Charles C. Ragin." *Teaching Sociology*, *37*, 414-416.

Clark, W. R., Gilligan, M. J., & Golder, M. (2006). A simple multivariate test for asymmetric hypotheses. *Political Analysis*, *14*, 311-331.

Clarke, K. A. (2002). The reverend and the ravens: Comment on Seawright. *Political Analysis*, *10*, 194-197.

Collier, D. (2014). Comment: QCA should set aside the algorithms. *Sociological Methodology*, *44*, 122-126.

Collier, D., Brady, H. E., & Seawright, J. (2010). Sources of leverage in causal inference: Toward an alternative view of methodology. In H. E. Brady & D. Collier (Eds.), *Rethinking social inquiry: Diverse tools, shared standards* (pp. 161-199). Lanham, MD: Rowman & Littlefield.

Compagnon, D. (2011). *Predictable tragedy: Robert Mugabe and the collapse of Zimbabwe*. Philadelphia: University of Pennsylvania Press.

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

Dunning, T. (2013, August 29 - September 1). *Contributions of fuzzy-set/qualitative comparative analysis: Some questions and misgivings*. Presented at the 109th annual convention of the American Political Science Association, Chicago, Illinois.

Freedman, D. A. (2008). Rejoinder. *Newsletter of the American Political Science Association Organized Section on Qualitative Methods*, *6*(2), 14-16.

Freedman, D. A. (2010). Black ravens, White shoes, and case selection: Inference with categorical variables. In D. Collier, J. S. Sekhon, & P. B. Stark (Eds.), *Statistical models and causal inference: A dialogue with the social sciences* (pp. 105-114). New York, NY: Cambridge University Press.

Gerber, A. S., & Green, D. P. (2013). *Field experiments: Design, analysis, and interpretation*. New York, NY: W.W. Norton.

Gerring, J. (2012). *Social science methodology: A unified framework*. New York, NY: Cambridge University Press.

Glynn, A. N. (2012). The product and difference fallacies for indirect effects. *American Journal of Political Science*, *56*, 257-269.

Goertz, G. (2005). Necessary condition hypotheses as deterministic or probabilistic: Does it matter? *Newsletter of the American Political Science Association Organized Section on Qualitative Methods*, *3*(1), 22-27.

Goertz, G. (2008). Choosing cases for case studies: A qualitative logic. *Newsletter of the American Political Science Association Organized Section on Qualitative Methods*, *6*(2), 11-14.

Goertz, G., & Mahoney, J. (2012). *A tale of two cultures: Qualitative and quantitative research in the social sciences*. Princeton, NJ: Princeton University Press.

Good, K. (1976). Settler colonialism: Economic development and class formation. *The Journal of Modern African Studies*, *14*, 597-620.

Grimmer, J., Messing, S., & Westwood, S. J. (2014). *Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods* (Working paper). Stanford, CA: Stanford University. Retrieved from http://stanford.edu/~jgrimmer/het.pdf

Hainmueller, J., & Hazlett, C. (2014). Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, *22*, 143-168.

Hall, P. A. (2003). Aligning ontology and methodology in comparative research. In J. Mahoney & D. Rueschemeyer (Eds.), *Comparative historical analysis in the social sciences* (pp. 373-404). New York, NY: Cambridge University Press.

Hausmann, R. (2003). Venezuela's growth implosion: A neoclassical story? In D. Rodrik (Ed.), *In search of prosperity: Analytic narratives on economic growth* (pp. 244-270). Princeton, NJ: Princeton University Press.

Hug, S. (2013). Qualitative comparative analysis: How inductive use and measurement error lead to problematic inference. *Political Analysis*, *21*, 252-265.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *The American Political Science Review*, *105*, 765-789.

Kam, C. D., & Franzese, R. J., Jr. (2007). *Modeling and interpreting interactive hypotheses in regression analysis*. Ann Arbor: University of Michigan Press.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

Krogslund, C., Danny Choi, D., & Poertner, M. (2014). Fuzzy sets on shaky ground: Parameter sensitivity and confirmation bias in fsQCA. *Political Analysis*. In press.

Levitsky, S., & Way, L. A. (2010). *Competitive authoritarianism: Hybrid regimes after the cold war*. New York, NY: Cambridge University Press.

Lieberman, E. S. (2005). Nested analysis as a mixed-method strategy for comparative research. *American Political Science Review*, *99*, 435-452.

Lucas, S. R., & Szatrowski, A. (2014). Qualitative comparative analysis in critical perspective. *Sociological Methodology*, *44*, 1-79.

Mahoney, J. (2008). Toward a unified theory of causality. *Comparative Political Studies*, *41*, 412-436.

Mahoney, J. (2010). *Colonialism and postcolonial development: Spanish America in comparative perspective*. New York, NY: Cambridge University Press.

Mahoney, J., Goertz, G., & Ragin, C. C. (2013). Causal models and counterfactuals. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 75-90). New York, NY: Springer.

Mann, M. (1993). *The sources of social power: Volume 2, The rise of classes and nation-states, 1760-1914*. New York, NY: Cambridge University Press.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. New York, NY: Cambridge University Press.

Ragin, C. C. (2000). *Fuzzy-set social science*. Chicago, IL: Chicago University Press.

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

Rohlfing, I., & Schneider, C. Q. (2013). Improving research on necessary conditions: Formalized case selection for process tracing after QCA. *Political Research Quarterly*, *66*, 220-230.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer-Verlag.

Schneider, C. Q., & Rohlfing, I. (2013). Combining QCA and process tracing in set-theoretic multi-method research. *Sociological Methods & Research*, *42*, 559-597.

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

Schneider, C. Q., & Wagemann, C. (2013a). Are we all set? *Newsletter of the American Political Science Association Organized Section on Qualitative Methods*, *11*(1), 5-8.

Schneider, C. Q., & Wagemann, C. (2013b). Doing justice to logical remainders in QCA: Moving beyond the standard analysis. *Political Research Quarterly*, *66*, 211-220.

Seawright, J. (2002). Testing for necessary and/or sufficient causation: Which cases are relevant? *Political Analysis*, *10*, 178-193.

Seawright, J. (2014a). *Multi-method social science: Combining qualitative and quantitative tools*. New York, NY: Cambridge University Press. In press.

Seawright, J. (2014b). Comment: Limited diversity and the unreliability of QCA. *Sociological Methodology*, *44*, 118-121.

Sekhon, J. S. (2004). Quality meets quantity: Case studies, conditional probability, and counterfactuals. *Perspectives on Politics*, *2*, 281-293.

Sekhon, J. S. (2005). Probability tests require distributions. *Newsletter of the American Political Science Association Organized Section on Qualitative Methods*, *3*(1), 29-30.

United Nations Development Programme. (2013). *Indices & data: Human Development Reports (HDR)*. Retrieved from http://hdr.undp.org/en/statistics

Waldner, D. (2005). It ain't necessarily so—Or is it? *Newsletter of the American Political Science Association Organized Section on Qualitative Methods*, *3*(1), 27-29.

Author Biography

Jack Paine is a Ph.D. Candidate in the Department of Political Science at the University of California, Berkeley. His research focuses primarily on applying game theoretic models to study civil war and authoritarian politics. He is particularly interested in the effects of oil wealth and of pre-colonial/colonial legacies.