Advance Access publication February 13, 2013

Political Analysis (2013) 21:252–265
doi:10.1093/pan/mps061

Qualitative Comparative Analysis: How Inductive Use and Measurement Error Lead to Problematic Inference

Simon Hug

Département de science politique et relations internationales, Faculté des sciences économiques et sociales, Université de Genève, 40 Bd du Pont d'Arve, 1211 Genève 4, Switzerland
e-mail: simon.hug@unige.ch (corresponding author)

Edited by Jonathan Katz

An increasing number of analyses in various subfields of political science employ Boolean algebra as proposed by Ragin's qualitative comparative analysis (QCA). This type of analysis is perfectly justifiable if the goal is to test deterministic hypotheses under the assumption of error-free measures of the employed variables. My contention is, however, that only in a very few research areas are our theories sufficiently advanced to yield deterministic hypotheses. Also, given the nature of our objects of study, error-free measures are largely an illusion. Hence, it is unsurprising that many studies employ QCA inductively and gloss over possible measurement errors. In this article, I address these issues and demonstrate the consequences of these problems with simple empirical examples. In an analysis similar to Monte Carlo simulation, I show that using Boolean algebra in an exploratory fashion without considering possible measurement errors may lead to dramatically misleading inferences. I then suggest remedies that help researchers to circumvent some of these pitfalls.

1 Introduction

With some delay, Ragin's (1987) proposal to use Boolean algebra in qualitative comparative analysis (QCA) has found increasing resonance in the literature. More and more studies rely on this method to study phenomena in many subfields of political science and related social sciences (for overviews, see for instance Ragin, Berg-Schlosser, and de Meur 1996; Yamasaki and Rihoux 2008). While the required rigor that such analyses impose on the researcher is certainly a welcome change in comparative research, many researchers use these methods[1] without the necessary care. This carelessness relates in my view to two main issues.

First of all, contrary to the original design of QCA as a tool for testing propositions of deterministic necessary and sufficient conditions, many scholars implicitly or explicitly employ it as a tool for inductive theory generation. Second, very few scholars employing QCA consider possible measurement error. Contrary to the users of most other empirical methods employed in the social sciences, scholars employing QCA rarely reflect on the possibility that the data they have gathered and used in their analysis might be error-prone and thus affect their conclusions.

Author's note: Comments by Matthew Gabel, Jonathan Katz, Gerald Schneider, and anonymous reviewers as well as the proofreading by Joanne Richards are gratefully appreciated. Thanks are also due to Bear Braumoeller and Gary Goertz, who provided guidance in implementing their recommendations for testing necessary conditions, and to Jeff Gill for encouraging me to submit this article to *Political Analysis*. Supplementary materials for this article are available on the *Political Analysis* Web site.

[1]I use the plural here since the original QCA analysis proposed by Ragin (1987) has been extended by him (Ragin 2000) and others (see overviews in Ragin 1994, 2008; Ragin, Berg-Schlosser, and de Meur 1996; Rihoux 2006a; Schneider and Wagemann 2006, 2007; Rihoux and Ragin 2008b) in various ways. The problems I wish to discuss in this article apply, however, quite generally to all variants of this basic form of QCA, also often described as crisp-set QCA. For this reason, I will systematically use QCA in this article and not distinguish the various variants, except where needed (for a similar way to proceed, see Rihoux and Lobe 2009, 223).

Both of these issues create, however, quite considerable problems for using QCA, but seem to have been largely ignored by adherents of the method (see the discussions of the state of art and the critiques of QCA in, for instance, Schneider and Wagemann 2007; de Meur, Rihoux, and Yamasaki 2008; Ragin 2008; Rihoux 2008, 727–729) and deemed unworthy of the discussion in "good practice" guides (e.g., Schneider and Wagemann 2010). In this article, I wish to show how failing to address these two fundamental issues may lead to erroneous conclusions, and how one could try to circumvent these problems. The issues raised here relate to problems discussed in recent comparisons between QCA and quantitative statistical analysis (Seawright 2005b; Grofman and Schneider 2009), and in two studies focusing on the robustness of QCA (Skaaning 2011; Maggetti and Levi-Faur forthcoming). None of these recent contributions address directly, however, the issues raised in this article.[2]

In the next section, I briefly review the basic ideas of QCA and provide examples of how it has been employed in various research contexts. In Section 3, I highlight how in many studies QCA is either implicitly or explicitly used as an inductive tool for theory generation. I demonstrate that this use of QCA is first of all contrary to the original design of the method, and second that it is problematic in the absence of very detailed consideration of the scope of the theory (see Goertz and Mahoney 2009) to be tested and the logic of inference. Given that this problem also relates more generally to ideas of testing necessary and sufficient conditions, I also refer to this literature. In Section 4, I turn to the problem of measurement error, which may quite dramatically affect the insights drawn from QCA. I highlight the effects of such errors before proposing an analysis similar to Monte Carlo simulations of QCA in Section 5, both for the problem of using QCA inductively and for possible measurement errors. Section 6 discusses these results and suggests two ways that researchers can address these problems before concluding.

2 The Logic of QCA and Applications

The basic premise of QCA is that outcomes can be causally explained by combinations of necessary and sufficient conditions.[3] To test whether particular conditions are necessary or sufficient, the starting point is a so-called truth table that lists all possible combinations of conditions and outcomes (true or false) and shows how often they appear in the set of cases considered. To derive the so-called prime implicants, Ragin (1987, 85) suggests using Boolean algebra. These prime implicants state what combinations of conditions are necessary or sufficient for the outcome to occur or not.

In essence, the logic behind this approach relies on the methods of agreement and the difference proposed by Mill (1973 (1843)) more than one and a half centuries ago.[4] Interestingly enough, while many authors employing these methods refer proudly to this long tradition of thought (see most recently Rihoux 2006a, 688), few mention that Mill (1973 (1843), Book 6, chap. 7) rejected both methods as impractical, if not infeasible. The titles of three sections in the relevant chapter in Mill's (1973 (1843)) work should be illustrative enough:

> 3 The method of difference inapplicable in the social science
>
> . . .
>
> 4 The methods of agreement, and of concomitant variations, inconclusive in the social science
>
> . . .
>
> 5 The method of residues also inconclusive in the social science and presupposes deduction.

[2]See also the related article by Gelman et al. (2010) dealing with the relationship between deterministic and probabilistic models and Clark, Gilligan, and Golder's (2006) suggestions for how to test "asymmetric" hypotheses, that is, necessary and sufficient conditions (see below).

[3]A series of recent review articles and books (e.g., Ragin, Berg-Schlosser, and de Meur 1996; Rihoux 2006a; Schneider and Wagemann 2007; Ragin 2008; Rihoux and Ragin 2008b) discuss the basic ideas stated in Ragin (1987, 2000) in more detail and offer "best practices" for users.

[4]This relationship is nicely, though not exactly identically, discussed by Schneider and Wagemann (2007) and Caramani (2008). Surprisingly, however, Schneider and Wagemann (2007, 73–7) limit their discussion to the "method of agreement" and the "method of difference."

These sections discuss in detail why Mill (1973 (1843)) considers these methods to be "inconclusive," especially in the context of multiple causal paths leading to a particular outcome.[5] The title of the last section cited already suggests the possible solution Mill (1973 (1843)) wishes to propose. More precisely, given that experimental methods are often difficult to employ in the social sciences, only deductive, theoretically guided research is possible. Although this is consistent with the original intent of QCA, this important point, as Lieberson (1991) argues, is often missed. Only theory can delimit the set of causal factors and specify which interactions among them (i.e., which causal paths) are theoretically sound and should be empirically evaluated.

Consequently, Ragin's (1987) proposal to use Boolean algebra may be conceived as building upon Mill's (1973 (1843)) self-criticism. If QCA is used to evaluate theoretically deduced multiple causal paths, it directly addresses Mill's (1973 (1843)) worries of being unable, with his methods, to allow for "equifinality." These causal paths, however, have to be arrived at based on theoretical considerations (see above). An issue that neither Mill (1973 (1843)) nor Ragin's (1987) proposal address, however, is that measurement in science in general, and in the social sciences in particular, is error-prone. Relying on Mill's (1973 (1843)) or Ragin's (1987) method does not allow us to directly take measurement error into account.

A brief glance at some studies employing QCA suggests, however, that these well-known issues, raised in this long historical tradition, have been widely forgotten.[6] Hicks, Misra, and Ng (1995), for instance, employ QCA to study the emergence and consolidation of the welfare state, whereas Pennings (2005) uses a similar methodology to explore the reasons for welfare state reform.[7] These questions are also addressed in a study by Kittel, Obinger, and Wagschal (2000). In all three cases, it is unclear whether the endeavor is inductive or deductive. Also, none of these three sets of authors consider the possibility that the variables they use might be affected by measurement error.

Two other studies employing QCA refer to measurement error, but do so in two very different ways. Ebbinghaus and Visser (1999) in their study of union density find that Italy does not conform to their explanation (i.e., forms a contradiction in their QCA) and suspect, among other things, measurement error. Such a possibility is, however, not entertained for any of the other cases and explanatory variables. As the consequences of such measurement errors are quite different to those known from statistical methods (see below), proceeding in such a way is problematic. Stokke (2004, 108), on the other hand, in his study on regime effectiveness in international relations, discusses the issue of "[v]ulnerability to [m]iscodings" and suggests that "... the Boolean version of QCA ... makes strong assumptions on the accuracy of the data."[8] He then proceeds in his illustrative study to assess how a recoding of a case affects his conclusions.[9]

This short discussion of some prominent studies relying on QCA suggests that the necessary deductive underpinnings of causal paths are rarely offered. In addition, many if not most QCA studies do not consider measurement error and if they do so, as QCA does not offer a direct way to address this problem, employ hardly any remedies, or suggest some that has no foundation in the premises of the method (for such recommendations for fuzzy set QCAs, see Maggetti and Levi-Faur forthcoming).

[5]Lieberson's (1991) critique of the careless application of Mill's method underlines this point even more forcefully, as do critiques by Little (1995) and Goldthorpe (1997). While Mill (1973 (1843)) mentions the problems created by multiple causal paths, these are not the only issues (see, for instance, the discussion of these issues by Bennett 2004, 32).

[6]Rihoux (2006a) and Yamasaki and Rihoux (2008) provide a much more extensive discussion and list of works employing QCA. The points raised below for a selected number of applications also apply, however, to those listed by these two sets of authors.

[7]Pennings (2005) employs Ragin's (2000) fuzzy set analysis that, as discussed below, also has difficulties coping with measurement error.

[8]Stokke (2004, 109) then goes on to suggest that "the fuzzy-set version permits ... procedures for addressing measurement errors." What these procedures are, however, is discussed neither by Stokke (2004) nor by Ebbinghaus (2005), who makes a similar claim. As discussed below, Ragin (2000) also neglects to offer any discussion of how measurement error might be addressed in a fuzzy set QCA.

[9]As this is one of the suggestions I present below to address measurement error (see also Braumoeller and Goertz 2000), I postpone a detailed discussion of this point until later.

3 QCA as Inductive Tool

As the few examples alluded to above suggest, often authors employing QCA use it either implicitly or explicitly as an inductive tool to generate theoretical propositions. This use of QCA goes, however, explicitly against the original design of the method. Both Mill (1973 (1843), Book 6, chap. 7) and Ragin (1987) highlight the importance of deduction. Hence, Ragin (1987, 45) is quite explicit about the deductive nature of the case-oriented approach:[10]

> It is deductive because initial theoretical notions serve as guides in the examination of causally relevant similarities and differences. (Without theoretical guides, the search for similarities and differences could go on forever.)[11]

When it comes to causal processes, the same insistence on deduction appears in Little's (1995, 48) critique of Mill's (1973 (1843)) methods:

> . . . [I]t is not sufficient to permit the researcher to infer the complete underlying causal structure. Rather, it would be necessary to arrive at a hypothesis about the causal relations among these conditions; and such a hypothesis most naturally emerges from a substantive theory of the causal mechanisms that are at work in the social phenomena under consideration.

The necessarily deductive nature of this enterprise is often, however, neglected in practical research. For instance, in Osa and Corduneanu-Huci's (2003) work on mobilization efforts in nondemocracies, indicators for political opportunities are combined with a measure for the presence of social networks without giving any theoretical thought to the causal patterns in which these relate to mobilization. Similarly, the elimination of other possible explanatory factors, suggested by the social movement literature, is hardly discussed.[12]

The same reservations may apply to other studies employing QCA, particularly when it is far from clear what the theoretical bases are.

Using a QCA as an inductive theory-generating tool contradicts its original purpose, and, as Mill (1973 (1843), Book 3, chap. 10) implicitly argues, raises the issue of the applicability (or scope) of a theory.[13] If no theoretical considerations on the scope of the theoretical arguments are made, then quite clearly QCA yields only a description of the data at hand (e.g., Rihoux 2006a) and nothing more.[14] If we all the same insist on making the scope of the inductive theory correspond to the set of countries studied, there are obviously two major concerns: first, while certain conditions may perhaps go hand in hand with a particular outcome (for instance, being registered in a course to get an A, to

[10]See also Ragin, Berg-Schlosser, and de Meur (1996) and Rihoux (2008, 729), who states that "[t]he use of QCA should be both case informed. . . and theory informed." Interesting to note is that in more recent discussions of this approach by Rihoux (2006a, 683) (see also Rihoux 2008, 728) and Rihoux and Ragin (2008a), for instance, this central use of QCA is complemented by four additional usages, namely, theory generation, data description, data checking, and tests of assumptions. While the latter three are certainly helpful usages of QCA for exploratory purposes, the first one goes squarely against the original design and poses considerable problems as demonstrated below. Oddly, Schneider and Wagemann (2010, 3–4) in their discussion of "good practice" consider these five aims as the "original" ones.

[11]Ragin (1987, 45) continues by stating that "[i]t is inductive because the investigator determines which of the theoretically relevant similarities and differences are operative by examining empirical cases." His use of the term "inductive" is misleading, since what he describes is a simple process of theory-testing in Popper's (1959) sense, and certainly not in the way the latter author describes inductivism. Similarly, the end of Chapter 3 in Ragin (1987), while suggesting a back and forth between theory and empirics in QCA, nowhere states that QCA could be carried out without theory. This also implies that theory is needed to specify not only the components of a causal path, but also the latter's precise configuration.

[12]The problem of omitted variables is discussed in detail by Seawright (2005b), who shows that very strong assumptions need to hold for such omissions not to affect the results of a QCA (see also the debate among Achen 2005; Ragin 2005; Seawright 2005a). I use Osa and Corduneanu-Huci (2003) as a second example presented in more detail in the replication materials.

[13]Goldthorpe (1997, 15) makes the same point, while Goertz and Mahoney (2009) provide insightful discussions on the scope of a theory and the problem of causal homogeneity (see also Mahoney and Goertz 2006). Schneider and Wagemann (2010, 5), on the other hand, only briefly refer to scope conditions, arguing essentially that after having obtained results from a QCA, scope conditions have to be clearly specified. This seems, especially in a deductive perspective, hardly the appropriate way to deal with scope conditions.

[14]Whether or not QCA allows for causal explanations also seems to be very often unclear in the minds of QCA defenders. Hence, in their discussion of possible critiques of QCA, we first learn from de Meur, Rihoux, and Yamasaki (2008) that "[w]hen the number of cases is small. . . , this strategy tends to result in individualizing explanations" (153) only to be told two pages down that "QCA minimization algorithms do not produce 'explantions' of a given outcome" (155).

take Waldner's (2005) example), they hardly are causal explanations, and second, the elimination of contradictory cases may actually eliminate the real causal mechanisms from consideration.[15] These same problems of induction also appear in the discussion on empirical tests of necessary conditions. This is nicely illustrated in the debate between Seawright (2002a), Braumoeller and Goertz (2002), and Clarke (2002) (see also Seawright 2002b; Braumoeller 2003). Hence, it seems reasonable to subscribe to Mahoney and Goertz's (2006, 228) statement that

> . . . qualitative and quantitative scholars share the overarching goal of producing valid descriptive and causal inferences.[16]

4 QCA and Measurement Error

By its very nature, QCA is a tool to assess causal hypotheses of a very specific type. Clearly, the original QCA approach considered only deterministic causal hypotheses (Goldthorpe 1997, 5, 7), while more recent developments also focus on probabilistic causal hypotheses.[17]

Given this deterministic view of causality, it is surprising that measurement error has been little acknowledged in QCA studies. As a single case may lead us to reject a deterministic hypothesis, we should worry, it seems, considerably about possible errors. In part, this neglect is rather surprising, since many authors employing QCA emphasize the qualitative nature of the research in political science, for instance, and the considerable difficulties and subjectivities involved in measurement.[18] Many authors consider the extension of the QCA logic to fuzzy sets (Ragin 2000) as a response to such critiques, since fuzzy sets seem to allow for measurement error (e.g., Stokke 2004; Ebbinghaus 2005). Careful reading of Ragin's (2000) book shows, however, that the fuzzy sets are not employed to allow for measurement error, but only allow for graded coding.[19]

[15]As demonstrated below, these problems are obviously compounded by the possible presence of measurement error. See, however, the tests proposed by Braumoeller and Goertz (2000) to rule out trivial necessary conditions (as, for instance, the one presented by Waldner 2005) and their convincing argument for ruling out such conditions. Similarly, Ragin (2006) proposes the notion of "coverage" and a measure thereof to deal with, among others, trivial causal paths.

[16]It seems useful to recall that Mahoney and Goertz (2006, 228) employ here language from King, Keohane, and Verba's (1994, 7f) discussion, where inference denotes drawing conclusions about unobserved facts from observed ones. Consequently, employing QCA as a descriptive tool for the cases at hand does not qualify according to these terms as "descriptive inference."

[17]That this is done without any reference to, perhaps, the most authoritative discussion of a probabilistic conception of causality (summarized in Pearl 2001) may irritate more than one reader of this literature. This irritation may be heightened by the fact that, at no point in their presentation of QCA, do Berg-Schlosser et al. (2008, 9) discuss the fundamental principle of deterministic causation implied by QCA, except when noting that "QCA moves away, quite radically, from simplistic, probabilistic causal reasoning." Readers of more recent conceptions of causality (e.g., Pearl 2001) will be puzzled by this argument especially, because the specific conception of probabilistic causation is nowhere clearly stated, discussed, and defended. Odd to read in this context is also Waldner's (2005) skeptical assessment concerning probabilistic necessary conditions, since, according to him, classical logic has a hard time conceiving of such hypotheses. See also the rather strong position on this issue held by Little (1991) and Clark, Gilligan, and Golder (2006, 314).

[18]This critique, raised by Bollen, Entwistle, and Alderson (1993), is quoted by de Meur, Rihoux, and Yamasaki (2008, 149f), but not addressed. Interesting in this context is Rihoux's (2006a, 690) admission that "modifying the operationalization even only on one single condition . . . may very well bring profound modifications to the minimal formulae" (see on this point also Grofman and Schneider 2009; Skaaning 2011). He fails to note, however, that even without any changes in the operationalization, a small amount of measurement error has the same "profound" consequences, as I show below in more detail. Goldthorpe (1997, 20, note 8) refers to exactly the same problem when discussing Kangas's (1994) work: a simple recoding of a doubtful case would yield completely different "causal" results.

[19]This becomes quite transparent in the applications in Ragin (2000) and Pennings (2005), where continuous variables are simply rescaled to values between 0 and 1 to indicate the likelihood of belonging to the two main polar categories. This rescaling, obviously, cannot deal with the problem of measurement error in the underlying continuous scale or in the misclassification on the rescaled indicator, which may follow. See also the related discussion in Rihoux (2008, 729–33), Skaaning (2011), and Maggetti and Levi-Faur (forthcoming). Hence, Ragin (2000, 222–6) devotes exactly three pages to measurement error, but this is done without referring to the strategy of employing fuzzy sets. Nevertheless, both Stokke (2004) and Ebbinghaus (2005) argue that fuzzy sets allow for measurement error to be taken into account, despite any explicit statement of how this is done. Perhaps, this superficial (or even lacking) treatment of measurement error might be due to the view articulated by Waldner (2005), who argues that measurement error in qualitative research is dependent on the observer, whereas it is independent of the observer in most of the natural sciences. Obviously, this dichotomous perspective is hardly appropriate and certainly fails to account for differences between qualitative and quantitative studies (see, e.g., Benoit 2005).

Much more astute in this context are tests of necessary conditions proposed by Dion (1998), Braumoeller and Goertz (2000), and Clark, Gilligan, and Golder (2006). Dion (1998), in essence, adopts a Bayesian perspective where prior beliefs about necessary conditions are updated on the basis of new data according to Bayes's rule. What remains unspecified, however, is where the uncertainty is coming from if some case(s) contradict(s) the necessary condition being tested. Obviously, this uncertainty might come from measurement error, as is suggested by Braumoeller and Goertz (2000, 848). These authors present a simple test for necessary conditions relying on the idea that counterexamples to necessary conditions might be due to measurement error. If the extent of this measurement error can be assessed (e.g., through reliability tests), this information can be used to assess whether observed counterexamples to a necessary condition are likely to stem from measurement error. In this case, the necessary condition hypothesis can be maintained despite these counterexamples.[20]

Clark, Gilligan, and Golder (2006), on the other hand, suggest evaluating "asymmetric" hypotheses, that is, hypotheses suggesting that the presence and the absence of a condition do not have symmetric effects on the outcome to be explained, with traditional quantitative tools and the use of interaction effects.

As these authors nicely explain, necessary and sufficient conditions have exactly this characteristic, and this observation allows us to draw on methods for which insights on measurement error, for instance, are much more broadly available.

These suggestions for explicitly (or implicitly) addressing measurement error made by Dion (1998), Braumoeller and Goertz (2000), and Clark, Gilligan, and Golder (2006) have, however, found little or any resonance among scholars employing QCA.[21]

Most studies employing QCA, by not even mentioning the possibility of measurement error, implicitly eliminate this possibility by fiat. Quite likely, this is due to the fact that until now there has been no explicit way of incorporating such measurement error into QCA. And, as I demonstrate below, measurement error affects conclusions from QCAs quite dramatically.

5 A Seemingly Monte Carlo Simulation

Using QCA in an inductive fashion and with data that are not error-free might be considered a situation of "violations in assumption," as discussed in classical statistics textbooks. In this latter framework, the consequences of such violations (especially in small samples) are often assessed with Monte Carlo simulations. By specifying a particular data-generating process, for instance including various levels of measurement error in the dependent variable, a set of data sets are generated and the properties of the proposed estimator are evaluated. One might proceed similarly in assessing the consequences of measurement error and misspecified scope conditions in a QCA, but to remain closer to substantive examples, I proceed differently. I rely on existing data sets from QCA studies and modify these data sets to "simulate" possible measurement error and misspecified scope conditions. Generating, however, a large number of "new" data sets based, for instance, on a randomly changed coding in the dependent variable will lead to a set of data sets, where each will appear identically in approximately equal proportions. Hence, in the analyses reported below, I generate only one data set per change, whereas in the replication materials, I report on another Monte Carlo simulation, where the changes were done randomly, thus generating 1000 new data sets each.

[20]Braumoeller and Goertz (2000) also suggest tests to rule out trivial necessary conditions and to make sure that falsely accepting a necessary condition hypothesis (i.e., type II error) does not happen too frequently and to rule out trivial necessary conditions. I will discuss these tests in more detail below. Ragin (2006) proposes a related measure, "consistency," which is concerned with the number of deviant cases.

[21]The same also holds for the approach presented by Braumoeller (2003) to analyze situations of causal complexity. The reluctance to address measurement error in empirical tests of necessary and sufficient conditions or using QCA might be related to the issue raised by Sekhon (2005a), who argues that when moving from a deterministic view of necessary conditions to a probabilistic one, distributional properties need to be defined. This, however, is difficult and rarely if ever done (see also Pearl 2001; Sekhon 2005b).

Table 1 Data from (Ragin 2000) and (Grofman and Schneider 2009)[a]

Country	P	U	C	S	W
Austria	1	1	1	1	1
Denmark	1	1	1	1	1
Finland	1	1	1	1	1
Norway	1	1	1	1	1
Sweden	1	1	1	1	1
Australia	0	0	0	0	0
Canada	0	0	0	0	0
France	0	0	0	0	0
USA	0	0	0	0	0
Germany	0	0	1	0	0
Netherlands	0	0	1	0	0
Switzerland	0	0	1	0	0
Japan	0	0	0	1	0
New Zealand	0	1	0	0	0
Ireland	0	1	1	1	1
Belgium	1	1	1	0	1

[a]Legend: P, strong left party; U, strong unions; C, corporatist industrial system; S, sociocultural homogeneity; and W, strong welfare state.

The data used for the analyses that follow have been previously used by Ragin (2000, 292), Grofman and Schneider (2009), and Skaaning (2011).[22] These authors considered that four conditions should influence the presence of a strong welfare state, namely, a strong left party, strong unions, a corporatist industrial system, and sociocultural homogeneity. Following Grofman and Schneider (2009, 663), I use only sixteen of the eighteen cases (thus dropping Australia and Italy) used by Ragin (2000, 292). Table 1 depicts the data in tabular form. As Grofman and Schneider (2009, 664) nicely discuss, "... there are two sufficient paths leading to a generous welfare state: a strong corporatist industrial system (C) AND strong unions (U) AND a strong left party (P) OR strong unions (U) AND a strong corporatist industrial system (C) AND socioeconomic [*sic*] homogeneity (S) ..."[23] Or, more formally,

$$\mathrm{PUC} + \mathrm{UCS} \rightarrow W. \quad (1)$$

Using the QCA implementation in *R* by Dusa (2006, 2007), I first replicated Grofman and Schneider's (2009) analysis and obtained the same result.[24] Starting from this analysis, I then proceeded to a "seemingly" Monte Carlo simulation. More precisely, instead of employing randomly generated data, I rely on Grofman and Schneider's (2009) data and changed it in a systematic manner described below.

The first set of such analyses deals with the issue of using QCA as an inductive tool. As discussed above, this use of QCA goes against the original design of the method and, in the absence of a thorough discussion of the scope of the inductive results, is highly problematic. In order to demonstrate this, I assume that the scope of the theory to be inductively derived corresponds to the set of sixteen cases studied by Ragin (2000), Grofman and Schneider (2009), and Skaaning (2011). Consequently, and assuming even further that the explanatory variables are the correct ones, the

[22]The choice of the example is irrelevant, since for the analyses that follow, changes in the data will be studied. The nature of the original data is, however, of mostly secondary importance. In the replication materials, I provide an additional set of analyses based on a study by Osa and Corduneanu-Huci (2003), generating the same basic insights as those discussed in the main text.

[23] It needs to be noted that Grofman and Schneider (2009) omit logical remainders, i.e. combinations of conditions that are not present in the data as depicted in Table 1. While I adopt the same strategy below, an analysis including the logical remainders can be found in the replication materials. This latter analysis is based on Osa and Corduneanu-Huci's (2003) data. As the results show, even in this case the general conclusions regarding measurement errors and the inductive use of QCA still obtain.

[24]I have also carried out all analyses reported upon in this article by using a newer implementation of QCA by Huang (2011). None of the results reported here change. The necessary *R*-code may be obtained from the author upon request.

Table 2 QCA with one observation removed (remainder excluded)

Solution	N (%)
PUC + UCS	14 (87.5)
PUC	1 (6.25)
UCS	1 (6.25)

results reported above (equation (1)) are obviously the correct theory. Based on this, I carried out two Monte Carlo simulations. In the first, one observation was dropped from the data in a sequential way. In the second, two observations were dropped again sequentially. Both situations "simulate" instances in which the scope of the theory fails to correspond to the set of cases employed for theory generation.

Not surprisingly, the inductively derived "theories" differ in some cases. In Table 2, I report the results for the Monte Carlo simulations when only one observation is dropped.[25] As these results show, in slightly more than 10% of the cases, the original solution fails to materialize. More precisely, if either the Irish or Belgian case is dropped PUC, respectively UCS, appears as a unique solution and hence as unique sufficient condition. Hence, if Ireland is omitted from the analysis, socioeconomic homogeneity is no longer a necessary condition of a sufficient condition for a strong welfare state, whereas omitting Belgium removes the presence of a strong left party from the list of necessary conditions forming a sufficient one. Substantively, such changes are far from innocuous, as for instance, the link between socioeconomic homogeneity and the development of a welfare state is a hotly debated topic in the literature (e.g., Alesina, Baqir, and Easterly 1999). Dropping any of the other fourteen observations from the data set, however, fails to affect the result of the QCA.

When two observations are dropped, the share of Monte Carlo runs yielding different solutions that increases considerably, as Table 3 demonstrates. Obviously, the results reported in Table 3 are related to the results discussed above. Any pair of dropped observations involving Ireland and/or Belgium must by definition yield a different solution than the original one reported in equation (1). In the present analysis, in almost a quarter of all simulation runs the original solution fails to materialize. The same alternative solutions as found above, namely PUC and UCS, appear and in addition PUCS is found. The latter materializes when both Ireland and Belgium are removed from the data set. This solution would suggest that each of the original four conditions is necessary and jointly sufficient for the presence of a strong welfare state. Again, given the considerable debates about the social and the political ramifications of welfare states, hardly a minor change in the proposed explanation.

These simple Monte Carlo simulations suggest that using QCA as tool to inductively build theories is highly problematic. If only a single case for one reason or another (for instance, because of the missing data) is dropped from the analysis despite the fact that the scope of the theory should cover it, a sizeable probability exists that the inductive theory is wrong. The same obviously occurs if the scope of the theory is not correctly specified and fails to correspond to the set of cases used for theory generation. This probability, in the present Monte Carlo simulations being equal to approximately 0.1, rises to almost 0.25 when two cases are omitted. Given that it is extremely hard for inductively generated theories to define the latter's scope, QCA as a purely inductive tool is, as Ragin (1987, 45) has already noted, not appropriate.[26]

[25]Here and in subsequent tables, I refrain from reporting measures of "consistency" and "coverage." First, all causal paths presented in a basic QCA have a perfect consistency (i.e., no counterexamples). Second, while "coverage" and even more so "unique coverage" highlight the fragility of some inferences, they offer only a limited view, as discussed below. The *R*-code for these, as well as all other analyses, appears in the replication materials. As the additional example in the replication materials demonstrates, proceeding in a random fashion to create, for instance, 1000 data sets would lead to exactly the same set of changed outcomes and their relative frequency among the 1000 replications would be very close to those reported here in Tables 2–5.

[26]Hence, a QCA in such a context becomes a simple data description tool, however, given that it fails to allow for any theory generation, and for any inference (in the sense of King, Keohane, and Verba 1994, 7f), its scientific value for this purpose may be questioned. Interesting to note is that de Meur, Rihoux and Yamasaki (2008, 153) refer to King,

Table 3 QCA with two observations removed (remainder excluded)

Solution	N (%)
PUC + UCS	91 (75.83)
PUC	14 (11.67)
UCS	14 (11.67)
PUCS	1 (0.83)

Table 4 QCA with one error (remainder excluded)

Solution	N (%)
PUC + UCS	12 (75)
PUC + UCS + pucS	1 (6.25)
PUC + UCS + pUcs	1 (6.25)
PUC	1 (6.25)
UCS	1 (6.25)

The second set of Monte Carlo simulations deals with the possibility of measurement error in the dependent variable. In more traditional quantitative approaches, the consequences of such measurement errors are well understood, as any textbook, for instance Hanushek and Jackson (1977), will show (for less well-understood consequences, see Hausman, Abrevaya, and Scott-Morton 1998; Beger et al. 2009; Hug 2010). As measurement error in the dependent variable is part and parcel of the error term of the theoretical model in traditional quantitative approaches, it increases the uncertainty of our estimates as long as the measurement error is not correlated with our other explanatory variables.

Table 4 reports the results from the "seemingly" Monte Carlo simulations, where the value of the outcome variable was changed in the sequence once for each of the sixteen cases (*R*-code appears in the replication materials). As appears in the last column of this table, in approximately three-quarters of the cases such potential measurement errors would not affect the result. More precisely, if the outcome is mismeasured for any of the first twelve observations in Table 1, the result does not change. Changing any of the remaining four (out of sixteen) observations leads in part, however, to dramatically different results. Either (if Japan or New Zealand is mismeasured) an additional sufficient condition appears for a strong welfare state or (if Ireland or Belgium is mismeasured) one sufficient condition fails to materialize.[27] The former situation in particular leads to substantively different insights. Either the presence of socioeconomic homogeneity combined with the absence of a strong left party, weak unions, and a weak corporatist industrial system is sufficient to produce a strong welfare state, or alternatively, strong unions and the absence of the other three conditions are sufficient to produce the same outcome. Such results are very much at odds with the original solution obtained (equation (1)) when the data were considered error-free.

Table 5 reports the same type of analysis; however, this time in each run, the dependent variable for two observations was recoded. Now, we only find the same solution as reported in equation (1) in slightly less than 40% of the runs. In addition, the number of other types of solutions increases dramatically. Instead of four additional solutions, we suddenly have fourteen. Many of these are variants of those appearing in Table 4, but again some would considerably alter our conclusions. For instance, while both a strong leftist party and socioeconomic homogeneity are necessary

Keohane, and Verba's (1994) notion of inference only in relation to assumptions being made to logical remainders, i.e., combinations of conditions, which do not appear in the data.

[27]This is obviously related to what Ragin (2006, 304) calls "unique coverage"; namely, each of these sufficient conditions is only upheld by one single case. While Schneider and Wagemann (2010) suggest reporting the "coverage" of particular solutions, they offer no guidance on what to do if "unique coverage" is low.

Table 5 QCA with 2 errors (remainder excluded)

Solution	N (%)
pUCS + PUCs	45 (37.5)
pucS + pUCS + PUCs	5 (4.17)
pUcs + pUCS + PUCs	5 (4.17)
PUC	12 (10)
UCS	12 (10)
PUC + UCS	21 (17.50)
PUC + UCS + pucS	7 (5.83)
PUC + UCS + pUcs	7 (5.83)
PUC + UCS + pucS + pUcs	1 (0.83)
PUC + pucS	1 (0.83)
UCS + pucS	1 (0.83)
PUC + pUcs	1 (0.83)
UCS + pUcs	1 (0.83)
PUCS	1 (0.83)

conditions that form part of a sufficient one in almost all variants appearing in Table 4, this is no longer the case if we entertain the possibility of two measurement errors. Indeed, in 10% of all the iterations, socioeconomic homogeneity no longer forms a necessary condition of a sufficient one, whereas for strong leftist parties this percentage is slightly higher, at almost 12%. Again, such a difference is hardly innocuous in the current debate in the literature on welfare states, as discussed above.

Hence, not surprisingly, measurement error in the dependent variable may affect quite dramatically the results of QCA studies. When only slightly more than 5% of the cases are affected by measurement error, there is already a 25% chance of reaching the wrong conclusion. If the measurement error doubles to approximately 12%, the probability of drawing erroneous conclusions increases to more than 0.6. One might argue that the likelihood of measurement error in binary outcome variables is very often lower. In a study of job tenure, Hausman, Abrevaya, and Scott-Morton (1998) show, however, that approximately 10% of the respondents in a panel survey err when making statements on job change. Similarly, when reanalyzing several studies on civil war and social protest, Hug (2010)—using the approach proposed by Hausman, Abrevaya, and Scott-Morton (1998)—estimates that the amount of misclassification in binary dependent variables ranges all the way up to 20% (for another approach to address misclassification in models with binary dependent variables, see Beger et al. 2009). Hence, the extent of measurement error assumed for the Monte Carlo simulations here is quite representative of what is to be expected in most empirical studies.[28]

6 Discussion and Conclusion

In his review article, Rihoux (2006a, 683) lists five main purposes for using QCA in research. Only one of them corresponds to the originally stated purpose of the method, namely, the testing of theoretical propositions. The others include describing data, checking data, assessing certain assumptions, and generating theory. In this article, I have argued that QCA is perfectly appropriate for testing particular types of theories conceived in very specific terms of causality, provided measurement error can be ruled out. It is also obviously perfectly adequate as a data description tool, as long as the data are error-free. When it runs into trouble, however, it is in the theory-generating

[28]Interestingly, de Meur, Rihoux, and Yamasaki (2008, 151) explicitly cite the war–peace dichotomy as one that is unproblematic. Most scholars having dealt with this "dichotomy" in empirical studies might offer a dissenting view (see, for instance, Gates and Strand 2004). Obviously, I deal here only with measurement error in the dependent variable. It is likely that the consequences of measurement error in explanatory variables, as in the classical linear regression framework, are even harder to assess and correct for.

mode as I have demonstrated in this article, and more generally, if measurement error is present. Without any a priori conception of what the scope of a particular theory is, various subsets of a set of cases may yield completely different inductively generated theoretical statements.

Even in the theory-testing mode, however, the use of QCA faces considerable problems. While one might perhaps agree with authors like George and Bennett (2005) that measurement error may be reduced through the use of case studies, no sound mind would argue that such errors can completely be ruled out, especially in the social sciences. As demonstrated in the Monte Carlo simulations in this article, allowing for possible measurement error may considerably affect the results of tests of our theoretical propositions. Consequently, Stokke (2004, 109) suggests redoing a QCA after recoding doubtful cases and assessing the robustness of the results. While this already seems an improvement over most standard practices, it can hardly be considered as a rigorous test of measurement error.[29]

As discussed above, Braumoeller and Goertz (2000) provide tools that have hardly ever been employed in the literature, but which allow the testing of necessary condition hypotheses when faced with measurement error. I illustrate the use of these tools by assuming that the presence of a strong left party is assumed to be a necessary condition for the presence of a generous welfare state.[30] I use this condition given that there is one counterexample, which clearly appears in Table 1.[31] Given that I consider measurement error in the dependent variable, the one counterexample appears because among the ten cases where a strong left party is missing, a strong welfare state appeared all the same. Hence, the lower bound of a one-sided 95% confidence interval for this proportion of counterexamples based on a binomial distribution corresponds to 0.005.[32] Following Braumoeller and Goertz (2000, 848), the p_I-test would suggest rejecting the necessary condition hypothesis if this lower bound is higher than the error to be expected in the cell corresponding to the counterexamples in a two-by-two table. For the case at hand, there is no direct way to assess the reliability of the measures used, but assuming that one case might be on average misclassified (i.e., one out of sixteen), we already have an error rate of more than 0.06. Consequently, allowing for measurement error, we cannot reject the hypothesis that strong left parties are a necessary condition for a strong welfare state.[33] This result has to be related to the analyses reported upon above, where both measurement error and misspecified scope conditions led, sometimes, to the conclusion that a strong leftist party was not a necessary condition of a sufficient one to explain strong welfare states.

Braumoeller and Goertz (2000) also recommend an assessment of the power of this test, or, inversely, the type II error, namely the error of wrongly keeping the necessary condition hypothesis despite it being wrong. Proceeding as they suggest, by comparing the distribution as found in the data (one out of ten counterexamples) with a uniform distribution and assuming as significance level (α) 0.05, we find for the estimated power the value of 0.405, implying that our type II error is almost 0.6.[34] With the significance levels commonly accepted, one might be reluctant to accept the proposed necessary condition at face value.[35]

[29]Surprisingly, Schneider and Wagemann (2010) offer no advice concerning measurement error in their list of "good practices," whereas Maggetti and Levi-Faur (forthcoming) propose some hands-on fixes for fuzzy set QCA, which lack, however, any basis in the first principles of QCA.

[30]As any necessary condition can easily be transformed into a sufficient condition (and vice versa) (see, for instance, Braumoeller and Goertz 2000, 846), the same test could also be applied to the two sufficient conditions highlighted by Grofman and Schneider (2009).

[31]This table also shows that assuming strong unions and sociocultural homogeneity to be necessary conditions for a strong welfare state would yield no counterexamples. In the footnotes that follow, I also briefly mention the results for these two potential necessary conditions.

[32]For the cases with no counterexamples, this lower bound is obviously zero.

[33]Given that the lower bounds of the confidence intervals for the two other potential necessary conditions are zero, we cannot reject the hypotheses that they are actually necessary conditions.

[34]For the other two potential necessary conditions, these values are even lower, namely 0.280 for strong unions, and 0.160 for a corporatist industrial system.

[35]Braumoeller and Goertz (2000) also suggest that testing for the triviality of the necessary condition, by comparing the distribution of the values for X for each of the values of Y (left party strength, respectively, welfare state strength in the present case). Refraining from presenting the resulting crosstabulaton, it is easy to show that the respective χ^2 value of 12.343 allows us to reject (with $p = 0.000$) the triviality of the necessary condition (provided that we are to accept the

While Braumoeller and Goertz (2000) proposed that tests obviously only deal with one of the problems I have raised, namely measurement error, some of their results, together with those outlined in Dion (1998), can be drawn upon to at least partly alleviate the problems of using QCA as an inductive tool.[36] As the analyses presented above suggested, some solutions arrived at depend on the presence of one (or two) specific cases in the data set. To induce a theory (suggesting often a complex combination of conditions) on the basis of such a limited amount of information seems haphazard. This is also implied by Dion's (1998) finding that from a Bayesian perspective at least six cases are required to attain a sufficient level of certainty to maintain a necessary condition. Similarly, based on their p_{II}-test, Braumoeller and Goertz (2000) suggest that even without any counterexamples, necessary condition hypotheses need to be rejected if less than seven cases support them.[37] Relatedly, Ragin (2006, 304) suggests using "consistency" (i.e., the share of cases having the same combination of conditions that also display the outcome) as a useful criterion (see also Schneider and Wagemann 2010, 10). As Braumoeller and Goertz (2000) demonstrate, however, it is not only the share of consistent cases that is important, but also the absolute number.

These two suggestions may help partly address the problems raised in this article and demonstrated with a simple Monte Carlo–like simulation. While explicitly taking into account the possibility of measurement error seems a necessity, and the tools provided by Braumoeller and Goertz (2000) seem appropriate, avoiding the problems of "inductivism" seems more difficult, even when one does not agree with all of Popper's (1959) criticisms. Given, however, that there seems to be some agreement on King, Keohane, and Verba's (1994) argument that quantitative and qualitative (positivist) researches have to rely on similar basic inferential goals (see also Mahoney and Goertz 2006, 228),[38] perhaps the use of QCA for inductive purposes will simply fade away.

References

Achen, Christopher H. 2005. Two cheers for Charles Ragin. *Studies in Comparative International Development* 40:27–32.

Alesina, Alberto, Reza Baqir, and William Easterly. 1999. Public goods and ethnic divisions. *Quarterly Journal of Economics* 114:1243–84.

Beger, Andreas, Jacqueline H. R. DeMeritt, Wonjae Hwang, and Will H. Moore. 2009. *The split population logit (SPopLogit): Modeling measurement bias in binary data*. Department of Political Science, Florida State University, and Department of Political Science, University of Tennessee.

Bennett, Andrew. 2004. Case study methods: Design use and comparative advantages. In *Models, numbers, and cases: Methods for studying international relations*, eds. Detlef F Sprinz and Yael Wolinsky-Nahmias, 19–55. Ann Arbor, MI: University of Michigan Press.

———. 2006. Stirring the frequentist pot with a dash of Bayes. *Political Analysis* 14(3):339–44.

Benoit, Kenneth R. 2005. How qualitative research really counts. *Qualitative Methods Newsletter* 3(1):9–12.

Berg-Schlosser, Dirk, Gisèle de Meur, Benoît Rihoux, and Charles C Ragin. 2008. Qualitative comparative analysis (QCA) as an approach. In *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques*, eds. Benoît Rihoux and Charles C. Ragin, 1–18. Thousand Oaks, CA: Sage.

Bollen, Kenneth A., Barbara Entwistle, and Arthur S. Alderson. 1993. Macrocomparative research methods. *Annual Review of Sociology* 19:321–51.

latter as a hypothesis on the basis of the previous two tests). For the two other conditions, the χ^2 values are, respectively, 12.444 (strong unions) and 7.467 (corporatist industrial system) with p-values of 0.000 and 0.006.

[36]As the number of cases is rather small in the example I use here, I refrain from demonstrating Clark, Gilligan, and Golder's (2006) tests of "asymmetric" hypotheses.

[37]For the analysis based on Osa and Corduneanu-Huci's (2003) data and presented in the replication materials, the condition E: SOCIAL, which is supported by more than a dozen cases, seems a safe bet, while condition B: ACCES, being supported by a single case, should never be "inductively" considered in a theory-building effort. It is useful to note that Dion's (1998) and Braumoeller and Goertz's (2000) criteria are very demanding. Using as an example Hicks, Misra, and Ng's (1995) study of the welfare state (I use this example, since it figures prominently in Mahoney 2007), it follows that these authors do not have a sufficient number of cases to support their claims. More precisely, they identify three paths leading to a consolidated welfare state (consisting of different combinations of conditions). Following Dion's (1998) and Braumoeller and Goertz's (2000) rules, at least eighteen (respectively twenty-one) cases would be needed to provide evidence in support for these three paths, but Hicks, Misra, and Ng (1995) have only fifteen (see, for a short discussion, Mahoney 2007, 135).

[38]For the most recent debate about this issue, consider the contributions by Bennett (2006), Brady, Collier, and Seawright (2006), Shively (2006), Schrodt (2006), and Rihoux (2006b).

Brady, Henry E., David Collier, and Jason Seawright. 2006. Toward a pluralistic vision of methodology. *Political Analysis* 14(3):353–68.

Braumoeller, Bear F. 2003. Causal complexity and the study of politics. *Political Analysis* 11(3):209–33.

Braumoeller, Bear F., and Gary D. Goertz. 2000. The methodology of necessary conditions. *American Journal of Political Science* 44:844–58.

———. 2002. Watching your posterior: Comment on Seawright. *Political Analysis* 10(2):198–203.

Caramani, Daniele. 2008. *Introduction to the comparative method with Boolean algebra*. Thousand Oaks, CA: Sage.

Clark, William Roberts, Michael J. Gilligan, and Matt Golder. 2006. A simple multivariate test for asymmetric hypotheses. *Political Analysis* 14(3):311–31.

Clarke, Kevin A. 2002. The reverend and the ravens: Comment on Seawright. *Political Analysis* 10(2):194–7.

de Meur, Gisèle, Benoît Rihoux, and Sakura Yamasaki. 2008. Addressing the critiques of QCA. In *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques*, eds. Benoît Rihoux and Charles C Ragin, 147–65. Thousand Oaks, CA: Sage.

Dion, Douglas. 1998. Evidence and inference in the comparative case study. *Comparative Politics* 30(2):127–45.

Dusa, Adrian. 2006. *QCA package for R version 0.1-3*. Romanian Social Data Archive, University of Bucharest.

———. 2007. User manual for the QCA(GUI) package in R. *Journal of Business Research* 60(5):576–86.

Ebbinghaus, Bernhard. 2005. When less is more: Selection problems in large-N and small-N cross-national comparisons. *International Sociology* 20(2):133–52.

Ebbinghaus, Bernhard, and Jelle Visser. 1999. When institutions matter: Union growth and decline in Western Europe, 1950–1995. *European Sociological Review* 15(2):1–24.

Gates, Scott, and Havard Strand. 2004. Modeling the duration of civil wars: Measurement and estimation issues. Paper prepared for presentation at the Joint Session of Workshops of the ECPR, Uppsala.

Gelman, Andrew, Iwin Leenen, Iven Van Mechelen, Paul De Boeck, and Jeroen Poblome. 2010. Bridges between deterministic and probabilistic models for binary data. *Statistical Methodology* 7:187–209.

George, Alexander L., and Andrew Bennett. 2005. *Case studies and theory development*. Cambridge, MA: MIT Press.

Goertz, Gary, and James Mahoney. 2009. Scope in case study research. In *Sage handbook of case-based methods*, eds. David Byrne and Charles C. Ragin, 307–17. Thousand Oaks, CA: Sage.

Goldthorpe, John H. 1997. Current issues in comparative macrosociology: A debate on methodological issues. *Comparative Social Research* 16:1–26.

Grofman, Bernard, and Carsten Schneider. 2009. An introduction to crisp-set QCA, with a comparison to binary logistic regression. *Political Research Quarterly* 62(4):662–72.

Hanushek, Eric A., and John E. Jackson. 1977. *Statistical methods for social scientists*. New York: Academic Press.

Hausman, Jerry, Jason Abrevaya, and Fiona Scott-Morton. 1998. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics* 87:239–69.

Hicks, Alexander, Joya Misra, and Tang Nah Ng. 1995. The programmatic emergence of the social security state. *American Sociological Review* 60(3):329–49.

Huang, Ronggui. 2011. *QCA3: Yet another package for qualitative comparative analysis*. R package version 0.0-5. http://cran.r-project.org/web/packages/QCA3/.

Hug, Simon. 2010. The effect of misclassifications in probit models. Monte Carlo simulations and applications. *Political Analysis* 18(1):78–102.

———. 2013. Replication data for Qualitative Comparative Analysis: How inductive use and measurement error lead to problematic inference. http://hdl.handle.net/1902.1/19489 IQSS Dataverse Network [Distributor] V1 [Version].

Kangas, Olli. 1994. Macrosociological and comparative methodology. On regression, qualititative comparisons, and cluster analysis in the politics of social security. In *The Comparative political economy of the welfare state*, eds. Thomas Janoski and Alexander M. Hicks, 346–64. Cambridge, UK: Cambridge University Press.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

Kittel, Bernhard, Herbert Obinger, and Uwe Wagschal. 2000. Wohlfahrtsstaaten im Internationalen Vergleich. Politisch-Institutionelle Faktoren der Entstehung und Entwicklungsdynamik. In *Der gezügelte Wohlfahrtsstaat: Sozialpolitik in Australien, Japan, Schweiz, Kanada, Neuseeland and den Vereinigten Staaten*, eds. Herbert Obinger and Uwe Wagschal, 329–64. Frankfurt: Campus.

Lieberson, Stanley. 1991. Small N's and big conclusions: An examination of the reasoning in comparative studies based on a small number of cases. *Social Forces* 70(2):307–20.

Little, Daniel. 1991. *Varieties of social explanation: An introduction to the philosophy of social science*. Boulder, CO: Westview.

———. 1995. Causal explanation in the social sciences. *Southern Journal of Philosophy Supplement* 34:31–56.

Maggetti, Martino, and David Levi-Faur. Forthcoming. Dealing with errors in QCA. *Political Research Quarterly*.

Mahoney, James. 2007. Qualitative methodology and comparative politics. *Comparative Political Studies* 40(2):122–44.

Mahoney, James, and Gary Goertz. 2006. A tale of two cultures: Contrasting quantitative and qualitative research. *Political Analysis* 14(3):227–49.

Mill, John Stuart. 1973. Of the four methods of experimental inquiry. In *A system of logic, ratiocinative and inductive: Collected works of John Stuart Mill (Volume VII)*, ed. John Stuart Mill, 388–406. Toronto: University of Toronto Press.

———. 1988. *The logic of moral sciences*. La Salle, IL: Open Court.

Osa, Maryjane, and Cristina Corduneanu-Huci. 2003. Running uphill: Political opportunity in non-democracies. *Comparative Sociology* 2(4):605–29.

Pearl, Judea. 2001. *Causality*. New York: Cambridge University Press.

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

Popper, Karl R. 1959. *The logic of scientific discovery*. New York: Harper & Row.

Ragin, Charles. 1987. *The comparative method: Moving beyond qualitative and quantative strategies*. Berkeley, CA: University of California Press.

———. 1994. A qualitative comparative analysis of pension systems. In *The comparative political economy of the welfare state*, eds. Thomas Janoski and Alexander M. Hicks, 320–45. Cambridge, MA: Cambridge University Press.

———. 2000. *Fuzzy-set social science*. Berkeley, CA: University of California Press.

———. 2005. Core versus tangential assumptions in comparative research. *Studies in Comparative International Development* 40:33–8.

———. 2006. Set relations in social research: Evaluating their consistency and coverage. *Political Analysis* 14(3):291–310.

———. 2008. *Redesigning social inquiry: Fuzzy sets and beyond*. Chicago: University of Chicago Press.

Ragin, Charles C., Dirk Berg-Schlosser, and Gisèle de Meur. 1996. Political methodology: Qualitative methods. In *A new handbook of political science*, eds. Robert E. Goodin and Hans-Dieter Klingemann, 749–68. Oxford: Oxford University Press.

Rihoux, Benoît. 2006a. Qualitative comparative analysis (QCA) and related systematic comparative methods: Recent advances and remaining challenges for social science research. *International Sociology* 21(5):679–706.

———. 2006b. Two methodological worlds apart? Praises and critiques from a European comparativist. *Political Analysis* 14(3):332–35.

———. 2008. Case-orientated configurational research using QCA. In *The Oxford handbook of political methodology*, eds. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, 722–36. Oxford: Oxford University Press.

Rihoux, Benoit, and Bojana Lobe. 2009. The case for qualitative comparative analysis (QCA). Adding leverage for thick cross-case comparison. In *Sage handbook of case-based methods*, eds. David Byrne and Charles C. Ragin, 222–42. Thousand Oaks, CA: Sage.

Rihoux, Benoît, and Charles C. Ragin. 2008a. Introduction. In *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques*, eds. Benoît Rihoux, and Charles C. Ragin, xvii–xxv. Thousand Oaks, CA: Sage.

———. 2008b. *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques*. Thousand Oaks, CA: Sage.

Schneider, Carsten Q., and Claudius Wagemann. 2006. Reducing complexity in qualitative comparative analysis (QCA): Remote and proximate factors and the consolidation of democracy. *European Journal of Political Research* 45(5):751–86.

———. 2007. *Qualitative comparative analysis (QCA) und fuzzy sets. Ein Lehrbuch für Anweder und jene, die es werden wollen*. Opladen, Germany: Barbara Budrich Verlag Anton Hain.

———. 2010. Standards of good practice in qualitative comparative analysis (QCA) and fuzzy-sets. *Comparative Sociology* 9(3):397–418.

Schrodt, Philip A. 2006. Beyond the linear frequentist orthodoxy. *Political Analysis* 14(3):335–9.

Seawright, Jason. 2002a. Testing for necessary and/or sufficient causation: Which cases are relevant? *Political Analysis* 10(2):178–93.

———. 2002b. What counts as evidence? Reply. *Political Analysis* 10(2):204–7.

———. 2005a. Assumptions, causal inference, and the goals of QCA. *Studies in Comparative International Development* 40(1):39–42.

———. 2005b. Qualitative comparative analysis vis-à-vis regression. *Studies in Comparative International Development* 40(1):3–26.

Sekhon, Jasjeet S. 2005a. Probability tests require distribution. *Qualitative Methods Newsletter* 3(1):30–1.

———. 2005b. Quality meets quantity: Case studies, conditional probability and counterfactuals. *Perspectives on Politics* 2(2):281–93.

Shively, W. Phillips. 2006. Case selection: Insights from rethinking social inquiry. *Political Analysis* 14(3):344–7.

Skaaning, Svend-Erik. 2011. Assessing the robustness of crisp-set and fuzzy-set QCA results. *Sociological Methods & Research* 40:391–408.

Stokke, Olav Schram. 2004. Boolean analysis, mechanisms, and the study of regime effectiveness. In *Regime consequences: Methodological challenges and research strategies*, eds. Arild Underdal and Oran R. Young, 87–120. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Waldner, David. 2005. It ain't necessarily so—or is it? *Qualitative Methods Newsletter* 3(1):27–30.

Yamasaki, Sakura, and Benoît Rihoux. 2008. A commented review of applications. In *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques*, eds. Benoît Rihoux and Charles C. Ragin, 123–45. Thousand Oaks, CA: Sage.