Case Selection and the Validity of Causal Inferences

in Qualitative Comparative Research

Thomas Plümper[a], Vera Troeger[b] and Eric Neumayer[c]

[a] Vienna University of Economics, Department of Socioeconomics, Welthandelsplatz 1, 1020 Vienna, Austria, thomas.pluemper@wu.ac.uk (corresponding author)
[b] University of Warrick, Department of Economics, v.e.troeger@warwick.ac.uk
[c] London School of Economics and Political Science (LSE), Department of Geography and Environment, e.neumayer@lse.ac.uk

Abstract:

Traditionally, social scientists perceived causality as regularity. As a consequence, qualitative comparative research was regarded as unsuitable for drawing causal inferences since few cases cannot establish regularity. The dominant perception of causality has changed, however. Nowadays, social scientists define and identify causality through the counterfactual effect of a treatment. This brings causal inference in qualitative comparative research back on the agenda. We argue that the validity of causal inferences from the comparative study of cases depends on the case selection algorithm. We employ Monte Carlo techniques to demonstrate that different case selection rules strongly differ in their ex ante reliability for making valid causal inferences.

1. Introduction

The counterfactual turn in the conception of causality implies that one can infer causality from comparative case studies. We demonstrate that the validity of such inferences depends on the case-selection rule. In fact, we show that under identifiable conditions, qualitative comparative case studies can be used to infer causality with high validity.

The claim that qualitative research allows making causal inferences seems to be the methodological equivalent of stirring up a hornets' nest – or to be more precise: two hornets' nests. The first nest is inhabited by the tribe of quantitative researchers who believe that they have irreversibly won the debate with the qualitative tribe and that the presumed impossibility of deriving causal inferences by qualitative research methods was one important element of this victory. In the second are qualitative researchers who might well feel that the focus on causal inference in this paper suggests we do not understand the main purpose of qualitative methods. It may alienate qualitative researchers even more that we subject case-selection rules to Monte Carlo simulations, which are quantitative by nature (Vose 1996). Apparently, this amounts to heresy in the eyes of our qualitative peers, seemingly proving that we do not do justice to qualitative methods.

Naturally, we disagree – at least partly – with both camps. The quantitative critique is wrong and outdated because the concept of causality as regularity (Hume 1748, Mill 1843, Popper 1959) has been superseded by the concept of causality as counterfactual effect (Neyman 1923, Rubin 1974, Holland 1986). In fact, the counterfactual concept of causation requires only a single case for causal inference if only it were possible to observe the counterfactual (Rescher 1964, Sayer 2000, Gelman 2011). In the absence of directly observable counterfactual outcomes, the closest methodological equivalents according to the

'identification school' are randomization of treatment (McGinnis 1958) and stratification of treatment and control group (Särndal et al. 2003) through case-selection. It is this latter research strategy of – rule- or model-based case-selection – that demands a re-evaluation of qualitative comparative designs.

The qualitative critique does not convince us either. True, though some methodologists explicitly argue that qualitative methods can be employed for causal inferences,[1] case studies can serve many other purposes and are, arguably, better suited for inductive purposes such as theory and concept development (Bennett 2004, Gerring 2007). However, our analysis enables those qualitative researchers who do wish to make causal inferences based on the comparative analysis of cases to understand how case-selection rules differ in respect to their *ex ante* reliability for detecting the direction and strength of a causal effect. "Wait a minute", our critics will say: applied qualitative researchers hardly ever estimate an effect strength and thus an analysis of effect strengths must be irrelevant for comparative qualitative research.[2] However, we do not "compute the effect strength from a comparison of two cases" in order to tempt qualitative researchers to quantify effect strengths. We merely compute the effect

[1] Savolainen (1994), Mahoney (2000), Abell (2001), Maxwell (2004), Caren and Panofsky (2005), Flyvbjerg (2006), and Rohlfing (2014) have used different arguments for why qualitative case studies can be used for "drawing big conclusions" (Savolainen 1994: 1217). The most compelling argument, in our view, is based on the idea that single theories (ought to) allow researchers to derive multiple predictions. In principle, it must be possible to derive so many hypotheses from a theory that no other theory can explain all the predictions. If this is the case, then a single case study that supports all predictions would be powerful – though of course critics may always claim that an unknown theory may explain the same observed outcomes. In our view though, you cannot beat something with nothing. Flyvbjerg (2006) makes a broader argument and argues that causal inference from case study design is generally possible. He supports his view with examples from physics. Yet, all examples selected by Flyvbjerg's are deterministic in nature and homogeneous in the effect strength. These examples therefore do not provide evidence for the validity of causal inference from qualitative designs when causal mechanisms are probabilistic and effects are weak – as it is usually the case in the social sciences.

[2] But see Rosenthal (1996).

strength and compare it to the assumed true effect size to have an indicator against which we can judge the *ex ante* reliability of selection algorithms. Computing the effect size is a tool, not the goal. Even if comparative qualitative researchers only intended to make inferences on the direction of a causal effect, they should agree that the expected deviation of an implied effect strength estimate from the truth – called root mean squared error by the quantitative tribe – is a good indicator for the relative *ex ante* reliability of case-selection algorithms: The larger this deviation, the more likely that even the inferred direction of an effect is wrong.

We demonstrate that the validity of causal inferences based on qualitative comparison of cases depends on the data-generating process and on the choice of case-selection algorithm. While the first factor is beyond the influence of scientists, researchers can freely choose the algorithm that determines the selection of cases. Of course, methodologists have long since been aware of the importance of case-selection for qualitative research (Lijphart 1971, Eckstein 1975, Seawright and Gerring 2008). One can trace back systematic theoretical and methodological reasoning on case-selection to at least John Stuart Mill (1872). After all this time, one might expect that the optimal case-selection algorithms are known. Yet, this is not the case and we offer the first rigorous analysis of the relative performance of both simple and more complex case-selection rules under conditions of relevance to real world comparative research.

Specifically, we vary the size of the total set of cases from which specific cases are selected, we vary the degrees to which the causal factor of interest is correlated with confounding factors, and we vary the "signal-to-noise ratio", that is, the (relative) strength of the effect of the causal factor of interest.[3] Using a Monte Carlo design we compare the relative performance of 11 case-selection algorithms, partly following suggestions of qualitative

[3] Using the example of QCA analysis, Hug (2013) has demonstrated that noise can lead to wrong inferences. We agree with Hug that noise is ubiquitous and therefore include noise into the data generating process.

methodologists and partly derived from common practice in comparative case analyses. The very best case-selection algorithm results in an estimated average effect that is almost a hundred times closer to the true effect than the worst algorithm. We also evaluate the conditions conducive to higher validity of causal inferences from qualitative research. We find that the best selection algorithms exhibit relatively high ex ante reliability for making valid inferences if: a) the explanatory variable of interest exerts a strong influence on the dependent variable relative to random noise and confounding factors, b) the variable of interest is not too strongly correlated with confounding variables, and c) the dependent variable is not dichotomous. More importantly, while the best algorithms are still fairly reliable even in the presence of strong stochastic influences on the dependent variable and other complications, the worst algorithms are highly unreliable even if the conditions under which qualitative research works best are met.

The paper is organized as follows: the next section shows that the dominant modern concept of causality as counterfactual analysis implies that one can make causal inferences based on comparative qualitative analysis. One cannot make such inferences with certainty, however, and the validity of inferences will crucially depend on how cases are selected. We review what methodologists have advised on the selection of cases in qualitative comparative research in section 3. This informs our choice of selection algorithms that we subject to Monte Carlo analysis, though we also add some original algorithms to test whether and, if so, how much better they can perform. Section 4 describes these algorithms, the Monte Carlo design and how we evaluate the relative performance of the case-selection algorithms. Section 5 presents results from the three sets of Monte Carlo simulations already mentioned.

2. Causal Inference and Qualitative Comparative Research

Causality as regularity dominated the philosophy of science at least from Hume to Popper. Hume (1748) argued that scientists cannot have knowledge of causality beyond observed regularities in associations of events. He therefore suggests inferring causality through a systematic comparison of situations in which the presumed causal factor is present or absent, or varies in strength. The concept of causality as regularity became the central element of Hempel and Oppenheim's (1948) deductive-nomological model of scientific explanation. Hempel also was the first to develop the concept further to include statistical inference (Salmon 1989). In Popper's conception of a non-degenerative research program (Popper 1934/1959), a single falsification effectively leads to the rejection of the tested hypothesis or, worse, the theory from which the hypothesis derives. The "regularity" perspective culminates in the definition of science as "unbroken, natural regularity" (Ruse 1982: 74).

This "strict regularity" concept of causality had ambiguous implications for social science qualitative researchers' ability to make causal inferences. On the one hand, the analysis of a small number of cases cannot establish regularity. On the other hand, if, conversely, a single deviant case suffices to refute a causal claim or even a theory, as Popper believes, then strength in numbers does not exist.[4] The "strict regularity" perspective is dead, however, because a) not all regularities are causal ("correlation is not causation") and b) causality can be probabilistic rather than deterministic and can thus exist without strict regularity.

Probabilistic causal mechanisms paved the way for an interpretation of regularity as statistical regularity. Accordingly, it was the combination of causality as regularity and the acceptance

[4] See, for example, Goffman (1961). For a broader discussion, see Gobo (2004); or for a broader perspective on 'crucial cases', see Gerring (2007).

of probabilistic causal effects (the idea of statistical inference)[5] that inevitably led to the conclusion that one cannot derive causal inferences from qualitative research. Yet, not even the brilliant idea of statistical inference saved the regularity concept of causality. If correlation is not causality, then high correlation does not imply causality either and low correlation and statistical insignificance may indicate low-probability causality and a lack of sufficient variation rather than the absence of causality. Eventually, this insight eliminated the support for the causality as regularity view.

Over the last three decades, the concept of causality as regularity was replaced by the counterfactual concept of causality, also called the potential outcomes framework. Its understanding of causality is tautological: causality exists if a cause exerts a causal effect on the outcome, and a cause exerts a causal effect on the outcome when the relation is causal. This tautology seems to be the main reason why scholars advancing (Rubin 1974; Holland 1986; Pearl 2000; Hidalgo and Sekhon 2011) or supporting (Morgan and Winship 2007) the counterfactual perspective focus on causal inference and the identification of causal effects[6] rather than on causality itself (Pearl 2015).

According to the counterfactual concept of causality, causality is perfectly identified if one observes the outcome given treatment and the outcome given no treatment at the same time for the same person(s). Naturally, this is impossible. Hence, a counterfactual analysis starts

[5] Causal inference became statistical inference through hypothesis testing based on statistical significance and p-values (Fisher 1925) and statistical inference appears inconsistent with qualitative research. In order to estimate standard errors, a large number of cases have to be analyzed and the analysis requires a quantification of all analyzed pieces of information.

[6] Causal inference is more than the identification of causation and consists of four distinct elements: the identification of a causal relationship between two variables (cause and effect), the estimation or computation of the strength of the effect, the identification or formulation of a causal mechanism and, lastly, the generalization from the observed cases to the population. We focus on the first two aspects here.

with a "missing data2 problem and then immediately turns to "second-best2 options for inferring causality. If one cannot observe the potential or counterfactual outcome for any one single case, then one needs to resort to comparing the outcomes of different cases. This raises the challenge that either one must make sure that the cases compared are equal or sufficiently similar in all dimensions that matter or that one can render the influence of all potential confounders irrelevant. Otherwise, no causal effect has been 'identified'.[7]

The approach generally preferred by identification scholars – what they call the 2gold standard" – aspires to render potential confounders irrelevant by randomizing treatment across a very large number of cases in a controlled experiment.[8] Though practically all actual experiments fall way short of the ideal of experimental designs, the randomization of treatment in a sample where N approaches infinity guarantees that treatment will be uncorrelated with both observable and, crucially, unobservable confounders. Because of this lack of correlation with any potential confounder, any observable difference in outcomes between the two groups must be due to the treatment. If one assumes causal homogeneity among cases and assumes away that potential confounders might condition the effect of treatment, then ideal experiments will not only have identified a cause-effect relationship but will also allow the calculation of the unbiased effect size.

Clearly, from the experimentalist viewpoint, qualitative small-N comparative research is useless for causal inferences. In fact, so is everything else. Its diehard proponents explicitly argue that experiments are a necessary condition for causal inference. For example, Light, Singer, and Willett (1990: 5-6, emphasis in original) claim that "to establish a causal link,

[7] It is perhaps the most important problem of the identification school to define 'identification' as a binary concept. Given that an effect is either identified or not, it is not possible to merely improve identification.

[8] For deviating views, see Kaptchuk (2004) and Cartwright (2007).

you must conduct an *experiment* (…). Only experimental inquiries allow you to determine whether a treatment *causes* an outcome to change." This claim wrongly assumes that identification is a necessary condition for causal inference, whereas in fact perfect identification is only a necessary condition for making causal inferences that are valid with certainty. The idea that one can only make causal inferences if scientists are certain about having identified a cause-effect relationship via experiments is absurd, however. If the claim was correct, social scientists would not be able to infer that more education causes higher lifetime income, or that smoking causes lung cancer. For that matter, we would not be able to explore much else of interest to social scientists and public policymakers. The quest for causal inference in the social science is not about certainty; it is about how to deal with model uncertainty and how much uncertainty about the validity of inferences can be tolerated (Neumayer and Plümper 2016).

More importantly, making certainty a prerequisite for causal inference runs into a logical problem for the social sciences because experiments that social scientists are able to conduct do not generate inferences that are valid with certainty. Even ignoring causal heterogeneity and potential conditionalities, the confounding-factors problem can only be solved asymptotically, that is, by increasing the sample size to infinity. With a finite number of participants, randomization of treatment does not suffice to render treatment uncorrelated to unobserved confounders like mood, experience, knowledge, intelligence, and often to even observed confounders like age, sex, income, education between treatment and control group. As a remedy, many experimenters control for observable differences in addition to randomizing treatment. Since it is impossible to control all factors that influence human behavior, not least because some of them may be unobserved, the problem of confounders can be reduced but not eliminated by experiments. Yet, if experiments only increase the

probability that causal inferences are correct, then the strict dichotomy between experiments and all other research methods that Light, Singer, and Willett make is unjustified.

The second approach to solving the "missing data" problem in the counterfactual concept of causality argues that causal effects are identified if cases can be selected so as to guarantee that all the relevant properties of the treatment group exactly match the properties of the control group (Sekhon 2008, 2009; Morgan and Harding 2006, Zhou and Xie 2016). Identification via selection on the properties of the treatment and control groups requires perfect knowledge of all the factors that influence outcomes and also that one can match cases on these properties. As with experiments, falling short of this ideal will mean that a causal effect has not been identified with certainty, but does not render causal inference impossible. For experimentalists, matching is far inferior to experiments because they doubt one can know all the relevant properties (one can know the so-called data-generating process) and even if one could know these properties, one cannot measure all of these properties, some of which are unobservable, and thus one cannot match on them.

This second approach substitutes impossible counterfactual analyses with a possible analysis of cases that have been carefully selected to be homogeneous with respect to confounding variables. This strategy is obviously encouraging for causal inference based on case comparison. Nothing in this matching approach suggests that the validity of causal inferences depends on the number of cases. If cases are homogeneous, causal inferences based on small-N comparative qualitative methods become possible, and the validity of these causal inferences depends on the employed selection rule.

Comparative qualitative researchers have always made arguments that closely resemble matching: if two cases are identical in all relevant dimensions but vary in the dimension of interest (the treatment), then it is possible to directly infer causality and to compute a causal

effect (Bryman 1984). This possibility does not imply that causal inference from comparative qualitative research is optimal or easy, however. Of course, there is the issue of knowing all relevant dimensions and finding at least two cases which are identical in all these dimensions. There are other difficulties, too: First, if causal processes are stochastic, as they are bound to be, then a single small-N comparative analysis, which cannot control for noise and random errors, will not reveal the truth but some random deviation from the truth. Matching cases in a quantitative analysis with large N therefore can be superior - though the greater difficulty of adequately matching a larger number of cases means that any positive effect on the validity of causal inferences from efficiency gains may be defeated by the negative effect due to problems in matching. Second, perfect homogeneity among cases on all confounding factors can only be achieved if researchers know the true data-generating process, which is unlikely to be the case even if qualitative researchers argue that their in-depth study of cases allow them to know much more about this process than quantitative researchers do (George and Bennett 1979, 2005). In the absence of knowledge of the true data-generating process, qualitative comparative researchers should make sure that selected cases do not differ in respect to known strong confounding factors. The potential for bias grows with the strength of the potentially confounding factor (for which no controls have been included), and the size of the correlation between the variable of interest and the confounder.

3. Case-selection and Qualitative Methodology

Methodological advice on the selection of cases in qualitative research stands in a long tradition. John Stuart Mill in his *A System of Logic*, first published in 1843, proposed five methods meant to enable researchers to make causal inferences: the method of agreement, the method of difference, the double method of agreement and difference, the method of residues, and the method of concomitant variation. Modern methodologists have questioned

and criticized the usefulness and general applicability of Mill's methods (see, for example, Lieberson 1991, 1994; Sekhon 2004).[9] However, without doubt Mill's proposals had a major and lasting impact on the development of the two most prominent modern methods, namely the "most similar" and "most different" comparative case-study designs (Przeworski and Teune 1970; Lijphart 1971, 1975; Meckstroth 1975).

Yet, as Seawright and Gerring (2008: 295) point out, these and other methods of case-selection are "poorly understood and often misapplied". Qualitative researchers mean very different things when they invoke the same terms "most similar" or "most different" and usually the description of their research design is not precise enough to allow readers to assess exactly how cases have been chosen. Seawright and Gerring (2008) have therefore provided a formal definition and classification of these and other techniques of case-selection. They suggest that "in its purest form" (Seawright and Gerring 2008: 304) the "most similar2 design chooses cases which appear to be identical on all controls (z) but different in the variable of interest (x). Lijphart (1975: 163) suggested what might be regarded a variant of this method that asks researchers to maximize "the ratio between the variance of the operative variables and the variance of the control variables".

Naturally, the "most similar" technique is not easily applied because researchers find it difficult to match cases such that they are identical on all control variables. As Seawright and Gerring (2008: 305) concede: "Unfortunately, in most observational studies, the matching procedure described previously – known as exact matching – is impossible." This impossibility has three sources: first, researchers usually do not know the true model. Therefore they simply cannot match on all control variables. Second, even if known to affect

[9] We agree with Lieberson (1991, 1994) that Mill assumes deterministic, unconditional, and monocausal processes, which of course severely limits the usefulness of his suggestions. Thus, like everyone else, we do not take Mill literally, but rather 'reinvent' Mill, as Lieberson correctly describes our approach.

the dependent variable, many variables remain unobserved. And third, even if all necessary pieces of information are available, two cases that are identical in all excluded variables may not exist.

Qualitative researchers prefer the "most similar" technique, despite ambiguity in its definition and practical operationalization, to its main rival, the "most different" design. Seawright and Gerring (2008: 306) believe that this dominance of "most similar" over "most different" design is well justified. Defining the "most different" technique as choosing two cases that are identical in the outcome *y* and in the main variable of interest *x* but different in all control variables *z*, they argue that this technique does not generate much leverage.[10] They criticize three points: first, the chosen cases never represent the entire population (if *x can* in fact vary in the population). Second, the lack of variation in *x* renders it impossible to identify causal effects. And third, elimination of rival hypotheses is impossible.

For comparative case studies, Seawright and Gerring also identify a third selection technique, which they label the "diverse" technique. It selects cases so as to "represent the full range of values characterizing X, Y, or some particular X/Y relationship" (Seawright and Gerring 2008: 300). This definition is somewhat ambiguous and vague ("some particular relationship"), but one of the selection algorithms used below in our MC simulations captures the essence of this technique by simultaneously maximizing variation in *y* and *x*.

Perhaps surprisingly, King, Keohane and Verba's (1994) important contribution to qualitative research methodology discusses case-selection only from the perspective of unit homogeneity

[10] As Gerring (2004: 352) formulates poignantly: "There is little point in pursuing cross-unit analysis if the units in question do not exhibit variation on the dimensions of theoretical interest and/or the researcher cannot manage to hold other, potentially confounding, factors constant."

– broadly understood as constant effect assumption (King et al. 1994: 92)[11] – and selection bias – defined as non-random selection of cases that are not statistically representative of the population (Collier 1995: 462). Selecting cases in a way that does not avoid selection bias negatively affects the generalizability of inferences. Random sampling from the population of cases would clearly avoid selection bias. Thus, given the prominence of selection bias in King et al.'s discussion of case-selection, the absence of random sampling in comparative research may appear surprising. But it is not. Random selection of cases leads to inferences which are correct on average when the number of conducted case studies approaches infinity, but the sampling deviation is extremely large. As a consequence, the reliability of single studies of randomly sampled cases remains low. The advice King and his co-authors give on case-selection, then, lends additional credibility to commonly chosen practices by applied qualitative researchers, namely to avoid truncation of the dependent variable (p. 130, see also Collier et al. 2004: 91), to avoid selection on the dependent variable (p. 134, see also Collier et al. 2004: 88), while at the same time selecting according to the categories of the "key causal explanatory variable".[12]

While there is a growing consensus on the importance of case-selection qualitative research, as yet very little overall agreement has emerged concerning the use of central terminology and the relative advantages of different case-selection rules. Scholars agree that random sampling is unsuitable for qualitative research, but disagreement on sampling on the

[11] This interpretation deviates from Holland's definition of unit homogeneity, which requires that the conditional effect and the marginal effect are identical, implying that the size of the independent variable is the same (Holland 1986: 947). To satisfy King et al.'s definition of unit homogeneity, only the marginal effect has to be identical (King et al. 92-93).

[12] King et al. (1994) also repeatedly claim that increasing the number of observations makes causal inferences more reliable. Qualitative researchers have argued that this view, while correct in principle, does not do justice to qualitative research (Brady 2004, Bartels 2004, McKeown 2004). More importantly, they also suggest that the extent to which the logic of quantitative research can be superimposed on qualitative research designs has limits.

dependent variable, and the appropriate use of information from observable confounding factors persists. Our Monte Carlo analysis will shed light on this issue by exploring which selection algorithms are best suited under a variety of assumptions about the data-generating process.

4. A Monte Carlo Analysis of Case-selection Algorithms

In statistics, Monte Carlo experiments are employed to compare the performance of estimators. The term Monte Carlo experiments describes a broad set of techniques that randomly draw values from a probability distribution to add error to a predefined equation that serves as "data-generating process". Since the truth is known, it is straightforward to compare the estimated or computed effects to the true effects. An estimator performs the better the smaller the average distance between the estimated effect and the truth. This average distance is usually called the root mean squared error.

Our Monte Carlo experiments follow this common practice in statistics and merely replace the estimators by a case-selection rule or algorithm. We compare selection rules commonly used in applied qualitative research, as well as various simple permutations and extensions. Without loss of generality, we assume a data-generating process in which the dependent variable y is a linear function of a variable of interest x, a control variable z and an error term ε. Since we can interpret z as a vector of k control variables, we can generalize findings to analyses with multiple controls.[13]

Case-selection Algorithms

[13] However, we cannot generalize to selection algorithms that select two cases on one dimension and two other cases on another dimension of controls. We leave these issues to future research.

Ignoring for the time being standard advice against sampling on the dependent variable, researchers might wish to maximize variation of *y*, maximize variation of *x,* minimize variation of *z* or some combination thereof. Employing addition and substraction, the two most basic functions to aggregate information on more than one variable leads to seven permutations of information from which to choose; together with random sampling this results in eight simple case-selection algorithms – see table 1. The mathematical description of the selection algorithms, as shown in the last column of the table, relies on the set-up of the Monte Carlo analyses (described in the next section). In general, for each variable we generate Euclidean distance matrices, which are $N \times N$ matrices representing the difference or distance in a set of cases *i* and *j* forming the case-dyad *ij*. Starting from these distance matrices, we select two cases that follow a specific selection rule. For example, *max(x)* only considers the explanatory variable of interest, thereby ignoring the distance matrices for the dependent variable *y* and the control variable *z*. With *max(x)*, we select the two cases that represent the cell of the distance matrix with the largest distance value. We refrain from analyzing case-selection algorithms for qualitative research with more than two cases. Note, however, that all major results we show here carry over to selecting more than two cases based on a single algorithm. However, we do not yet know whether all our results carry over to analyses of more than two cases when researchers select cases based on different algorithms – a topic we will revisit in future research.[14]

[14] We also do not explore algorithms of case-selection when researchers are interested in more than one explanatory variable. We similarly leave this to future research.

Table 1: Simple Case-selection Algorithms

	Name	sampling information: max dist(y)	max dist(x)	min dist(z)	selection algorithm
1	random	no	no	no	random draw
2	max(y)	yes	no	no	max dist(y)
3	max(x)	no	yes	no	max dist(x)
4	min(z)	no	no	yes	min dist(z)
5	max(y)max(x)	yes	yes	no	max [dist(y)+dist(x)]
6	max(y)min(z)	yes	no	yes	max [dist(y)-dist(z)]
7	max(x)min(z)	no	yes	yes	max [dist(x)-dist(z)]
8	max(y)max(x)min(z)	yes	yes	yes	max [dist(y)+dist(x)-dist(z)]

Algorithm 1 does not use information (other than that a case belongs to the population), and samples cases randomly. We include this algorithm for completeness and because qualitative methodologists argue that random sampling – the gold standard for sampling in quantitative research – does not work well in small-N comparative research (Seawright and Gerring 2008: 295; King et al. 1994: 124).

We incorporate the second algorithm – pure sampling on the dependent variable without regard to variation of either *x* or *z* – for the same completeness reason. Echoing Geddes (1990), many scholars have argued that sampling on the dependent variable biases the results (King et al. 1994: 129, Collier and Mahoney 1996, Collier et al. 2004: 99). Geddes demonstrates that "selecting on the dependent variable" lies at the core of invalid results generated from qualitative research in fields as diverse as economic development, social revolution, and inflation.

But does Geddes's compelling critique of sampling on the dependent variable imply that applied researchers should entirely ignore information on the dependent variable when they also use information on the variable of interest or the confounding factors? Algorithms 5, 6, and 8 help us to explore this question. These rules include selection on the dependent variable

in addition to selection on *x* and/or *z*. Theoretically, these algorithms should perform better than the algorithm 2, but we are more interested in analyzing how these biased algorithms perform in comparison to their counterparts, namely algorithms 3, 4, and 7, which, respectively, maximize variation of *x*, minimize variation of *z*, and simultaneously maximize variation of *x*, and minimize variation of *z*, just as algorithms 5, 6 and 8 do, but this time without regard to variation of *y*.

Theoretically, one would expect algorithm 7 to outperform algorithms 3 and 4. Qualitative methodologists such as Gerring (2007) and Seawright and Gerring (2008) certainly expect this outcome and we concur. Using more information must be preferable to using less information when it comes to sampling. This does not imply, however, that algorithm 7 necessarily offers the optimal selection rule for comparative qualitative research. Since information from at least two different variables has to be aggregated, researchers have at their disposal multiple possible algorithms that all aggregate information in different ways. For example, in addition to the simple unweighted sum (or difference) that we assume in table 1, one can aggregate by multiplying or dividing the distances, and one can also weight the individual components.[15]

Arend Lijphart (1975) has suggested an alternative function for aggregation, namely maximizing the ratio of the variance in *x* and *z*: *max[dist(x)/dist(z)]*. We include Lijphart's suggestion as our algorithm 9 even though it suffers from a simple problem which reduces its usefulness: when the variance of the control variable *z* is smaller than 1.0, the variance of what Lijphart calls the operative variable *x* becomes increasingly unimportant for case-selection (unless of course the variation of the control variables is very similar across different pairs of cases). We solve this problem by also including in the competition an

[15] Other functions for aggregating information such as logarithms, roots and so on are of course possible.

augmented version of Lijphart's suggestion. This algorithm 10 adds one to the denominator of the algorithm proposed by Lijphart: *max[dist(x)/(1+dist(z))]*. Observe that adding one to the denominator prevents the algorithm from converging to *min[dist(z)]* when *dist(z)* becomes small. Finally, we add a variance-weighted version of algorithm 7 as our final algorithm 11 to check whether weighting improves on the simple algorithms. Table 2 summarizes the additional analyzed algorithms that aggregate information using more complicated functions.

Table 2: Case-selection Algorithms with More Complicated Functions for Aggregating Information from More than One Variable

	Name	sampling information: max dist(y)	max dist(x)	min dist(z)	selection algorithm
9	lijphart	no	yes	yes	max [dist(x)/dist(z)]
10	augmented lijphart	no	yes	yes	max [dist(x)/(1+dist(z))]
11	weighted max(x)min(z)	no	yes	yes	$\max \left[\frac{\text{dist(x)}}{\text{maxdist(x)}} - \frac{\text{dist(z)}}{\text{maxdist(z)}} \right]$

Note that thus far we have given the selection algorithms formal and technical labels, avoiding terminology of case-selection rules commonly used in the literature. Nevertheless, there are connections between some of the above algorithms and the terminology commonly used in the qualitative literature. For example, algorithms 2, 3 and 5 are variants of selection rules described by Gerring (2007) and Seawright and Gerring (2008) as "diverse" case-selection rules. Algorithms 2, 5, 6, and 8 all use information on variation of the dependent variable and are thus variants of selection on the dependent variable. More importantly, algorithms 4 and 7 as well as algorithms 9 to 11 seem to be variants of the most similar design. However, we do not call any of these algorithms "selection on the dependent variable" or "most similar". The reason is that, as discussed above, there is a lack of consen-

sus on terminology and different scholars prefer different labels and often mean different things when they invoke rules such as "sampling on the dependent variable" or "most similar".

The Monte Carlo Design

The use of Monte Carlo techniques may appear to be strange to qualitative researchers. However, Monte Carlo simulations are perfectly suited for the purpose of exploring the *ex ante* reliability of case-selection algorithms. As we have explained above, Monte Carlo simulations provide insights into the expected accuracy of inferences given certain pre-defined properties of the data-generating process. While they are commonly used to compare estimators, one can equally use them to compare the performance of different sampling rules.

Monte Carlo simulations allow us to systematically change the data-generating process, and to explore the comparative advantages of different selection algorithms depending on the assumptions we make about the data-generating process. Possible systematic changes include variation in the assumed level of correlation between explanatory variables, the relative importance of uncertainty, the level of measurement error, and so on. Unsystematic changes are modelled by repeated random draws of the error term.

Specifically, we define various data-generating processes from which we draw a number of random samples, and then select two cases from each sample according to a specific algorithm, as defined above. As a consequence of the unaccounted error process, the computed effects from the various Monte Carlo simulations will deviate somewhat from the truth. Yet, since we confront all selection algorithms to the same set of data-generating processes, including the same error processes, performance differences must result from the algorithms themselves. These differences occur because different algorithms will select different pairs of cases *i* and *j*, and, as a consequence, the computed effect and the distance of

this effect from the true effect differ. Our analysis explores to what extent a comparison of two cases allows researchers to estimate the effect that one explanatory variable, called *x*, exerts on a dependent variable, called *y*. We assume that this dependent variable *y* is a function of *x*, a single control variable *z*, which is observed, and some error term ε:[16] $y_i = \beta x_i + \gamma z_i + \varepsilon_i$, where β, γ represent coefficients and ε is an iid error process. The process resembles what Gerring and McDermott (2007: 690) call a "spatial comparison" (a comparison across *n* observations), but our conclusions equally apply to "longitudinal" (a comparison across *t* periods) and "dynamic comparisons" (a comparison across $n \cdot t$ observations). We conducted simulations with both a continuous and a binary dependent variable. We report results for the continuous variable in detail in the next section and briefly summarize the results for the binary dependent variable with full results reported in the appendices.

There are different ways to think about the error term. First, usually scientists implicitly assume that the world is not perfectly determined and they allow for multiple equilibria which depend on random constellations or the free will of actors. In this respect, the error term accounts for the existence of behavioral randomness. Second, virtually all social scientists acknowledge the existence of systematic and unsystematic measurement error. The error term can be perceived as accounting for information that is partly uncertain. And third, the error term can be interpreted as model uncertainty - that is, as unobserved omitted variables also exerting an influence on the dependent variable. Only if randomness and free

[16] Obviously, as var(ε) approaches zero, the data-generating process becomes increasingly deterministic. We follow the convention of quantitative methodology and assume that the error term is randomly drawn from a standard normal distribution. Note, however, that since we are not interested in asymptotic properties of case-selection algorithms, we could as well draw the error term from different distributions. This would have no consequence other than adding systematic bias to all algorithms alike.

will, measurement error, and model uncertainty did not exist, would the inclusion of an error term make no sense. We always draw *x* and *z* from a normal distribution, but, of course, alternative assumptions are possible. Given the low number of observations, it comes without loss in generality that we draw ε from a normal distribution with mean zero and standard deviation of 1.5[17]; and, unless otherwise stated, all true coefficients take the value of 1.0; the standard deviation of variables is 1.0; correlations are 0.0; and the number of observations N, representing the size of the sample from which researchers can select cases, equals 100.

Evaluating the Results from the Monte Carlo Simulations

We compare the reliability of inference on effect strength. Specifically, the effect size of *x* on *y* from a comparative case study with two cases equals

$$\hat{\beta}(x) = \frac{y_i - y_j}{x_i - x_j} \ , \qquad (1)$$

where subscripts [*i*,*j*] represent the two selected cases. We take the root mean squared error (RMSE) as our measure for the reliability of causal inference as it reacts to both bias and inefficiency. The RMSE is defined as

$$RMSE = \sqrt{\frac{\sum\left(\hat{\beta} - \beta_{true}\right)^2}{N}} = \sqrt{Var\left(\hat{\beta}\right) + \left[Bias\left(\hat{\beta}, \beta_{true}\right)^2 \right]} \ . \qquad (2)$$

This criterion not only incorporates bias (the average deviation of the computed effect from the true effect), but also accounts for inefficiency, which is a measure of the sampling variation of the computed effect that reflects the influence of random noise on the computed effect. Qualitative researchers cannot appropriately control for the influence of noise on estimates. The best they can do to account for randomness is to choose a case-selection algorithm that responds less thank others to noise. Naturally, these are case-selection

[17] Thereby, we keep the R^2 at approximately 0.5 for the simulations with a continuous dependent variable.

algorithms that make best use of information. In quantitative research, the property characterizing the best use of information is called *efficiency,* and we see no reason to deviate from this terminology.

5. Results from the Monte Carlo Analysis of Case-selection Algorithms

We conduct three sets of MC simulations, in which we vary the parameters of the data-generating process, and evaluate the effect of this variation on the precision with which the algorithms approach the true coefficients together with the efficiency of the estimation. In each type of analysis we draw 1,000 samples from the underlying data-generating process. In the first set of simulations, we change the number of observations from which the two cases are chosen (i = 1,…N), thereby varying the size of the sample, i.e., the total number of cases from which researchers can select two cases. In the second set of simulations, we vary the correlation between x and z - that is, the correlation between the variable of interest and the confounding factor. In the final set of simulations, we vary the variance of x and thus the effect size or explanatory power of x relative to the effect size of the confounding factor z.

Analyzing the impact of varying the sample size on the validity of inferences in qualitative research may seem strange at first glance. After all, qualitative researchers usually study a fairly limited number of cases. In fact, in our Monte Carlo analyses we generate effects by looking at a single pair of cases selected by each of the case-selection algorithms. So why should the number of cases from which we select the two cases matter? The reason is that if qualitative researchers can choose from a larger number of cases about which they have theoretically relevant information, they will be able to select a better pair of cases given the chosen algorithm. The more information researchers have before they select cases, the more reliable their inferences should thus become. In other words, N does not represent the number

of cases analyzed, but the number of the total set of cases from which the analyzed cases are chosen.

By varying the correlation between *x* and the control variable *z* we can analyze the impact of confounding factors on the performance of the case-selection algorithms. With increasing correlation, inferences should become less reliable. Thereby, we look at the effect of potential model misspecification on the validity of inference in qualitative research. While quantitative researchers can eliminate the potential for bias from correlated control variables by including these on the right-hand-side of the regression model, qualitative researchers have to use appropriate case-selection rules to reduce the potential for bias.

Finally, in varying the standard deviation of *x* we analyze the impact of varying the strength of the effect of the variable of interest on the dependent variable. The larger this relative effect size of the variable of interest, the more reliable causal inferences should become.[18] The smaller the effect of the variable of interest *x* on *y* in comparison to the effect on *y* of the control or confounding variables *z*, the harder it is to identify the effect correctly, and the less valid the inferences are - especially when the researcher does not know the true specification of the model.

Table 3 reports the Monte Carlo results obtained when we only vary the size of the sample from which we draw the two cases we compare. In this set of simulations, we do not allow for systematic correlation between the variable of interest *x* and the confounding factor *z*. The deviations of computed effects from the true effect occur because of "normal" sampling error, and how efficiently the algorithm deals with the available information.

[18] We achieve this by changing the variance of the explanatory variable *x*, leaving the variance of the confounding factor *z* and the coefficients constant. Equivalently, one could leave the variance of *x* constant and vary the variance of *z.* Alternatively, one can leave both variances constant and change the coefficients of *x* and/or *z*.

Table 3: Monte Carlo results from varying 'sample size'

	Algorithm	N=20	N=40	N=60	N=80	N=100
1	random	9.137	13.846	6.008	6.860	17.349
2	max(y)	65.096	16.481	55.532	7.604	12.787
3	max(x)	0.575	0.488	0.447	0.429	0.411
4	min(z)	23.399	10.234	40.154	18.113	6.929
5	max(y)max(x)	3.213	7.608	35.725	1.935	2.047
6	max(y)min(z)	13.072	5.915	14.028	7.241	9.997
7	max(x)min(z)	0.522	0.438	0.419	0.387	0.360
8	max(y)max(x)min(z)	2.925	2.014	1.704	1.505	1.563
9	lijphart	1.754	1.544	1.400	1.548	1.416
10	augmented lijphart	0.536	0.479	0.442	0.407	0.389
11	weighted max(x)min(z)	0.521	0.442	0.417	0.388	0.359

Note: corr(x,z)=0, SD(x)=1
The table displays the root mean squared error. Smaller numbers indicate higher reliability.

Observe, first, that of the basic case-selection algorithms, *max(x)min(z)* performs up to 100 times better with respect to the average deviation from the true effect (the root mean squared error) than the poorest-performing competitors, namely *random*, which draws two cases randomly from the sample, and *max(y)*, which purely selects on the dependent variable. The drawback from selecting on the dependent variable declines if researchers additionally take into account variation of *x* and/or variation of *z*, but these algorithms 5, 6, and 8 are typically inferior to their counterparts 3, 4, and 7, which ignore variation of the dependent variable. Accordingly, selection on the dependent variable not only leads to unreliable and likely wrong inferences, it also makes other selection algorithms less reliable. Hence, researchers should not pay attention to variation in the dependent variable *y* when they select cases. By selecting cases on the variable of interest *x* while at the same time controlling for the influence of confounding factors, researchers are likely to choose cases which vary in their outcome if *x* indeed exerts an effect on *y*.

Maximizing variation of *x* while at the same time minimizing variation of *z* appears optimal. Algorithm 7 uses subtraction as a basic function for aggregating information from more than one variable. Would using a more complicated function dramatically improve the performance of case-selection? Results reported in table 3 show that, at least for this set of simulations, this is not the case. Algorithm 7 performs roughly 10 percent better than the augmented version of Lijphart's proposal (*augmented lijphart*), and while algorithm 11, the variance-weighted version of algorithm 7, is very slightly superior, not much separates the performance of the two.

Another interesting finding from table 3 is that only four algorithms become systematically more reliable when the sample size from which we draw two cases increases. These four algorithms are: *max(x)*, *max(x)min(z)* and its weighted variant, *weighted max(x)min(z)*, as well as *augmented lijphart*. Algorithms need to have a certain quality to generate, in expectation, improvements in the validity of causal inferences when the sample size becomes larger. Random selection, for example, only improves on average if the increase in sample size leads to relatively more "onliers" than "outliers". This may be the case, but there is no guarantee. When researchers use relatively reliable case-selection algorithms, however, an increase in the size of the sample, on which information is available, improves causal inferences unless one adds extreme outliers to the sample.[19]

The results so far support the arguments against random selection and King, Keohane and Verba's (1994) verdict against sampling on the dependent variable, but of course qualitative researchers hardly ever select cases randomly. Selection on the dependent variable may be

[19] Note, we say that inferences become more reliable if cases are selected from a larger sample of cases for which researchers have sufficient information. We are not making any normative claim about enlarging the sample size, because the improvements of enlarging the sample from which cases are selected has to be discounted by the deteriorations caused by an increase in case heterogeneity caused by an enlarged sample.

more common practice, even if researchers rarely admit to it. If researchers know, as they typically do, that both *x* and *y* vary in similar ways, and allow this variation to guide their case-selection, then the results are likely to simply confirm their theoretical priors. Selection rules must thus be strict, and should be guided by *verifiable rules* rather than discretion.

In table 4, we report the results of Monte Carlo simulations from varying the correlation between the variable of interest *x* and the confounding factor *z*.

Table 4: Monte Carlo results from varying the correlation between the variable of interest *x* and the confounding factor *z*

	Algorithm	corr=-0.9	corr=-0.7	corr=-0.3	corr=0	corr=0.3	corr=0.7	corr=0.9
1	random	10.849	6.188	11.002	7.301	10.535	5.783	7.420
2	max(y)	52.987	64.685	13.840	7.154	4.215	2.883	2.379
3	max(x)	0.891	0.733	0.465	0.401	0.472	0.734	0.930
4	min(z)	20.962	8.325	5.717	5.653	8.742	10.358	36.662
5	max(y)max(x)	2.801	2.777	2.177	1.944	1.929	1.799	1.822
6	max(y)min(z)	61.050	19.741	6.171	4.685	9.976	11.658	4.980
7	max(x)min(z)	0.741	0.486	0.369	0.364	0.383	0.475	0.711
8	max(y)max(x)min(z)	10.010	2.787	1.591	1.520	1.666	1.981	2.159
9	lijphart	3.426	2.202	1.671	1.575	1.505	2.072	3.778
10	augmented lijphart	0.869	0.551	0.397	0.372	0.411	0.543	0.829
11	weighted max(x)min(z)	0.736	0.481	0.369	0.363	0.383	0.472	0.701

Note: SD(x)=1.0, N=100, SD(z)=1.0, Varying Correlation (x,z).
The table displays the root mean squared error. Smaller numbers indicate higher reliability.

Note that all substantive results from table 3 remain valid if we allow for correlation between the variable of interest and the confounding factor. In particular, algorithm 11, which weights the individual components of the best-performing simple case-selection algorithm 7, performs only very slightly better; while the performance gap between simple algorithm *max(x)min(z)*, based on subtraction, and the augmented Lijphart algorithm (*augmented lijphart*), which uses the ratio as aggregation function, increases marginally. Table 4 also demonstrates that correlation between the variable of interest and confounding factors renders causal inferences from qualitative research less reliable. Over all simulations and algorithms, the RMSE increases by at least 100 percent when the correlation between *x* and *z* increases from 0.0 to either -0.9 or +0.9.

Finally, we examine how algorithms respond to variation in the strength of the effect of the variable of interest. In this final set of simulations for which results are reported in table 5 we vary the standard deviation of the explanatory factor *x*; a small standard deviation indicates a small effect of *x* on *y* relative to the effect exerted from *z* on *y*. The results show that the performance of all case-selection algorithms suffers from a low "signal-to-noise" ratio. As one would expect, the smaller the effect of the variable of interest *x* on *y* relative to the effect of *z* on *y*, the less reliable the causal inferences from comparative case study research becomes. Yet, we find that the algorithms which performed best in the previous two sets of simulations also turn out to be least vulnerable to a small effect of the variable of interest. Accordingly, while inferences do become more unreliable when the effect of the variable of interest becomes small relative to the total variation of the dependent variable, comparative case studies are not simply confined to analyzing the main determinant of the phenomenon of interest if one of the top performing case-selection algorithms are used. As in the previous sets of simulations, we find that little is gained by employing more complicated functions

for aggregating information from more than one variable as, for example, the ratio (*augmented lijphart*) or weighting by the variance of *x* and *z* (*weighted max(x)min(z)*). Sticking to the most basic aggregation function has little cost, if any.

Table 5: Monte Carlo results from varying the strength of the effect of the variable of interest

	algorithm	SD(x)=0.3	SD(x)=0.7	SD(x)=1.0	SD(x)=1.5	SD(x)=2.0
1	random	20.796	13.926	7.301	4.701	12.342
2	max(y)	105.183	22.097	7.154	2.706	0.969
3	max(x)	1.390	0.597	0.401	0.274	0.200
4	min(z)	41.889	13.112	5.653	8.377	3.024
5	max(y)max(x)	56.402	6.168	1.944	0.803	0.456
6	max(y)min(z)	125.917	68.193	4.685	1.671	0.738
7	max(x)min(z)	1.291	0.521	0.364	0.236	0.177
8	max(y)max(x)min(z)	95.349	3.862	1.520	0.654	0.388
9	lijphart	4.842	2.153	1.575	0.956	0.730
10	augmented lijphart	1.293	0.542	0.372	0.259	0.197
11	weighted max(x)min(z)	1.242	0.522	0.363	0.233	0.177

Note: corr(x,z)=0.0, N=100, SD(z)=1.0, Varying SD(x)
The table displays the root mean squared error. Smaller numbers indicate higher reliability.

We now briefly report results from additional Monte Carlo simulations which we show in full in the appendix to the paper. First, weighting *x* and *z* by their respective sample range becomes more important when the data-generating process includes correlation between *x* and *z* and the effect of *x* on *y* is relatively small (see appendix table 1). In this case, weighting both the variation of *x* and *z* before using the *max(x)min(z)* selection rule for identifying two cases slightly increases the reliability of causal inferences.

Second, we also conducted the full range of Monte Carlo simulations with a dichotomous dependent variable (see appendix tables 2 to 5). We find that the algorithms that perform best with a continuous dependent variable also dominate with respect to reliability when we analyze dichotomous dependent variables. Yet, causal inferences from comparative qualitative case study

research become far less reliable when the dependent variable is dichotomous for all selection algorithms compared to the case of a continuous dependent variable. The root mean squared error roughly doubles for the better-performing algorithms. As a consequence, causal inferences with a binary dependent variable and an additional complication (either a non-trivial correlation between *x* and *z* or a relatively small effect of *x* on *y*) are not reliable. Accordingly, qualitative researchers should not throw away variation by dichotomizing their dependent variable. Where the dependent variable is dichotomous, qualitative research is confined to what most qualitative researchers actually do in these situations: trying to identify strong and deterministic relationships or necessary conditions (Dion 1998; Seawright 2002). In both cases, the strong deterministic effect of *x* on *y* compensates for the low level of information in the data.

6. Conclusion

Case-selection rules employed in qualitative research resemble 'matching' algorithms developed by identification scholars in quantitative research and thus can be employed to derive causal inferences. They also share their most important shortcoming: the extent to which causal inferences from selected samples are valid is partly determined by the extent of knowledge of the data-generating process. The more is known about the "true model", the better researchers can balance treated and control cases.

Our major contribution has been to guide qualitative comparative researchers on what are the selection rules with the highest *ex ante* reliability for the purpose of making valid causal inferences under a range of conditions regarding the underlying data-generating process. The validity of causal inferences from qualitative comparative research will necessarily always be uncertain but following

our guidance will allow comparative qualitative researchers to maximize the imperfect validity of their inferences.

Applied qualitative researchers can take away six important concrete lessons from our Monte Carlo simulations: First, ceteris paribus, selecting cases from a larger set of potential cases gives more reliable results. Qualitative researchers often deal with extremely small samples. Sometimes nothing can be done to increase sample size, but where there are no binding constraints it can well be worth the effort expanding the sample from which cases can be selected. Second, for all the better-performing selection algorithms, it holds that ignoring information on the dependent variable makes inferences much more reliable. Third, selecting cases based on both the variable of interest and confounding factors improves the *ex ante* reliability of causal inferences in comparison to selection algorithms that consider just the variable of interest or just confounding factors - even if this means that one no longer chooses the cases that match most closely on confounding factors. These algorithms are relatively best-performing, no matter what the underlying data-generating process (of those we have analyzed). This is a crucial lesson because applied qualitative researchers might not have much knowledge about the kind of data-generating process they are dealing with. Fourth, correlation between the variable of interest and confounding factors renders the selection algorithms less reliable. The same holds if the analyzed effect is weak. This reinforces existing views that qualitative case comparison is most suitable for studying strong and deterministic causal relationships (Dion 1998; Seawright 2002). Fifth, the reliability of case-selection rules depends on the variation in the dependent variable scholars can analyze. Accordingly, throwing away information by dichotomizing the dependent variable is a bad idea. A continuous dependent variable allows for more valid inferences; a dichotomous dependent variable should only be used if there is no alternative. Sixth, employing basic functions for aggregating information from more than

one variable (such as maximizing the difference between variation of *x* and variation of *z*) does not reduce by much the *ex ante* reliability of case-selection compared to more complicated aggregation functions (such as maximizing the ratio or the variance-weighted difference). The only exceptions occur if *x* and *z* are highly correlated and the effect of *x* on *y* is relatively small compared to the effect of *z* on *y*. As a general rule, applied researchers will not lose much by opting for the most basic aggregation function.

In conclusion, our Monte Carlo study is broadly consistent with the views of qualitative methodologists. After all, the best- or nearly best-performing algorithms in our analysis of alternative selection algorithms appear to be a variant of Gerring and Seawright's (2007) most similar design, which in turn draws on Przeworski and Teune's (1970), or a variant of Lijphart's (1975) suggestion for case-selection. However, we are the first to provide systematic evidence that upholds existing recommendations in the presence of stochastic error processes. In addition, we demonstrated that simple functions for linking variation of the explanatory variable with variation of the confounding variables perform relatively well in general. There is little reason to resort to more advanced functions unless the explanatory variable has a weak effect and is strongly correlated with the confounding variables.

References

Abell, Peter, 2001. Causality and Low-Frequency Complex Events. The Role of Comparative Narratives. Sociological Methods & Research, 30:1, 57-80.

Bartels, Larry M. 2004, The Unfulfilled Promises of Quantitative Imperialism, in Henry Brady and David Collier (eds.): Rethinking Social Inquiry: Diverse Tools, Shared Standards, Rowman and Littlefield, Lanham, 69-83.

Bennett, Andrew, 2004. Case Study Methods: Design, Use, and Comparative Advantages. Models, Numbers, and Cases: Methods for Studying International Relations, 19-55.

Brady, Henry E. 2004, Doing good and doing better: How far does the Quantitative Template get us? in: Henry Brady and David Collier (eds.): Rethinking Social Inquiry: Diverse Tools, Shared Standards, Rowman and Littlefield, Lanham, 53-83.

Bryman, Alan, 1984. The Debate about Quantitative and Qualitative Research: A Question of Method or Epistemology? British Journal of Sociology, 75-92.

Caren, Neal and Panofsky, Aaron, 2005. TQCA A Technique for Adding Temporality to Qualitative Comparative Analysis. Sociological Methods & Research, 34:2, 147-172.

Cartwright, Nancy, 2007. Are RCTs the Gold Standard? BioSocieties, 2:1, 11-20.

Collier, David 1995, Translating Quantitative Methods for Qualitative Researchers. The Case of Selection Bias, American Political Science Review 89, 461-466.

Collier, David and James Mahoney 1996, Insights and Pitfalls. Selection Bias in Qualitative Research, World Politics 49, 56-91.

Collier, David, James Mahoney and Jason Seawright 2004, Claiming too much: Warnings about Selection Bias, in Henry Brady and David Collier (eds.): Rethinking Social Inquiry: Diverse Tools, Shared Standards, Rowman and Littlefield, Lanham, 85-102.

Dion, Douglas 1998, Evidence and Inference in the Comparative Case Study, Comparative Politics 30, 127-145.

Eckstein, Harry, 1975. Case Study and Theory in Political Science. Greenstein, E and N. Polsby, eds. Handbook of Political Science (Vol. 7), London: Addison-Wesley.

Fisher, Ronald A., 1925. Statistical methods for research workers. Genesis Publishing Pvt Ltd.

Flyvbjerg, Bent, 2006. Five Misunderstandings about Case-study Research. Qualitative Inquiry, 12:2, 219-245.

Geddes, Barbara, 1990, How the Cases you Choose affect the Answers you get. Selection Bias in Comparative Politics, Political Analysis 2, 131-152.

Gelman, Andrew, 2011. Review Essay: Causality and Statistical Learning. American Journal of Sociology, 117:3, 955-966.

George, Alexander and Bennett, Andrew, 1979. Case Studies and Theory Development. Free Press.

George, Alexander L. and Bennett, Andrew, 2005. Case studies and theory development in the social sciences. MIT Press.

Gerring, John 2004, What is a Case Study and what is it good for? American Political Science Review 98, 341-354.

Gerring, John and Jason Seawright, 2007, Techniques for Choosing Cases, in: John Gerring, Case Study Research. Principles and Practices, Cambridge University Press, Cambridge, 86-150.

Gerring, John and Rose McDermott 1997, An Experimental Template for Case Study Research, American Journal of Political Science 51, 688-701.

Gerring, John, 2007, Case Study Research. Principles and Practices, Cambridge University Press, Cambridge.

Gobo, Giampietro, 2004. Sampling, Representativeness and Generalizability. In C. Seale, G. Gobo, J.F. Gubrium, D. Silverman D. (Eds.), Qualitative Research Practice. London: Sage, 435-456.

Goffman, Erving, 1961. Asylums: Essays on the Social Situation of Mental Patients and other Inmates. New York: Anchor Books.

Hempel, Carl G. and Oppenheim, Paul, 1948. Studies in the Logic of Explanation. Philosophy of Science 15:2, 135-175.

Hidalgo, F.D. and Sekhon, Jaseet S., 2011. Causality. International Encyclopedia of Political Science. Thousand Oaks, CA: SAGE Publications, Inc, 204-211.

Holland, Paul, 1986, Statistics and Causal Inference, Journal of the American Statistical Association 81, 945-960.

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

Hume, David, 1748, Philosophical Essays Concerning Human Understanding. London: A. Millar.

Kaptchuk, Ted J., 2001. The Double-blind, Randomized, Placebo-controlled Trial: Gold Standard or Golden Calf? Journal of Clinical Epidemiology, 54:6, 541-549.

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

Lieberson, Stanley, 1991. Small N's and Big Conclusions: An Examination of the Reasoning in Comparative Studies based on a Small Number of Cases. Social Forces, 70:2, 307-320.

Lieberson, Stanley, 1994. More on the Uneasy Case for Using Mill-type methods in Small-N Comparative Studies. Social Forces, 72:4, 1225-1237.

Light, R.J., Singer, J.D. and Willett, J.B., 1990. By Design. Cambridge, MA: Harvard University.

Lijphart, Arend 1971, Comparative Politics and the Comparative Method, American Political Science Review 65, 682- 693.

Lijphart, Arend 1975, Comparable Cases Strategy in Comparative Research, Comparative Political Studies 8, 158-177.

Mahoney, James 2007, Qualitative Methodology and Comparative Politics. Comparative Political Studies 40, 122-144.

McGinnis, Robert, 1958. Randomization and Inference in Sociological Research. American Sociological Review 23:4, 408-414.

McKeown, Timothy 2004, Case Studies and the Limits of the Quantitative Worldview, in Henry Brady and David Collier (eds.): Rethinking Social Inquiry: Diverse Tools, Shared Standards, Rowman and Littlefield, Lanham, 139-167.

Meckstroth, Theodore 1975, "Most Different Systems" and "Most Similar Systems": A Study in the Logic of Comparative Inquiry. Comparative Political Studies 8, 133-177.

Mill, John Stuart. 1843. A System of Logic. London: Parker.

Morgan, Stephen L. and Harding, David J., 2006. Matching Estimators of Causal Effects Prospects and Pitfalls in Theory and Practice. Sociological Methods & Research, 35:1, 3-60.

Morgan, Stephen L., and Christopher Winship, 2014. Counterfactuals and Causal Inference. Cambridge University Press.

Neumayer, Eric and Plümper, Thomas, 2016. Robustness Tests: Causal Inference with Behavioral Data, unp. manuscript, LSE and Vienna University of Economics.

Neyman, Jerzy. 1923/1990. On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Translated and edited by D. M. Dabrowska and T. P. Speed. Reprinted in Statistical Science 5, 463–80.

Pearl, Judea. 2010. The Foundations of Causal Inference. Sociological Methodology 40, 75-149.

Pearl, Judea, 2015. Causes of Effects and Effects of Causes, Sociological Methods & Research 44, 149-164.

Popper, Karl R. 1959. The Logic of Scientific Discovery. London: Hutchinson. First published in German by Springer in 1934.

Przeworski, Adam and Henry Teune 1970, The Logic of Comparative Social Inquiry. New York: Wiley.

Rescher, N., 1964. Hypothetical Reasoning. Amsterdam: North Holland.

Rohlfing, Ingo. 2014. Comparative Hypothesis Testing Via Process Tracing, Sociological Methods & Research 43, 606-642.

Rosenthal, James A., 1996. Qualitative Descriptors of Strength of Association and Effect Size. Journal of Social Service Research, 21:4, 37-59.

Rubin, D.B., 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. Journal of Educational Psychology, 66:5, 688-701.

Ruse, Michael, 1982. Creation Science Is Not Science. Science, Technology, & Human Values, 7:40, 72-78.

Salmon, Wesley C. 1989. Four Decades of Scientific Explanation. In Kitcher&Salmon (Eds.), Scientific Explanation, Minnesota studies in the philosophy of science (Vol XIII, pp. 3–219). Minnesota: University of Minnesota Press.

Särndal, Carl-Erik., Swensson, Bengt and Wretman, Jan, 2003. Model assisted survey sampling. Springer Science & Business Media.

Savolainen, Jukka, 1994. The Rationality of Drawing Big Conclusions Based on Small Samples: In Defense of Mill's Methods. Social Forces, 72:4, 1217-1224.

Sayer, Andrew, 2000. System, Lifeworld and Gender: Associational versus Counterfactual Thinking. Sociology, 34:4, 707-725.

Seawright, Jason 2002, Testing for Necessary and/or Sufficient Causation: Which Cases are Relevant? Political Analysis 10, 178-193.

Seawright, Jason and Gerring, John, 2008. Case-selection Techniques in Case Study Research a menu of Qualitative and Quantitative Options. Political Research Quarterly, 61:2, 294-308.

Sekhon, Jasjeet S. 2004, Quality Meets Quantity: Case Studies, Conditional Probability, and Counterfactuals. Perspectives on Politics 2, 281-293.

Sekhon, Jasjeet S., 2008. The Neyman-Rubin model of causal inference and estimation via matching methods. The Oxford Handbook of Political Methodology, 271-299.

Sekhon, Jasjeet S., 2009. Opiates for the Matches: Matching Methods for Causal Inference. Annual Review of Political Science, 12, 487-508.

Vose, David, 1996. Quantitative Risk Analysis: A Guide to Monte Carlo Simulation Modelling. John Wiley & Sons.

Zhou, Xiang and Xie, Yu, 2016. Propensity Score–based Methods Versus MTE-based Methods in Causal Inference: Identification, Estimation, and Application. Sociological Methods & Research 45, 3-40.

Appendix Table 1: MC Results Continuous Outcome SD(x)=0.3, N=100, SD(z)=1.0, Varying Correlation (x,z)

	Algorithm	corr=-0.9	corr=-0.7	corr=-0.3	corr=0	corr=0.3	corr=0.7	corr=0.9
1	random	20.411	26.946	25.505	20.796	43.789	49.871	31.853
2	max(y)	57.993	99.264	270.517	105.183	136.001	21.068	16.749
3	max(x)	3.021	2.402	1.572	1.390	1.610	2.524	2.908
4	min(z)	69.408	27.372	32.238	41.889	20.927	373.118	68.821
5	max(y)max(x)	21.081	49.380	83.554	56.402	32.309	13.379	11.375
6	max(y)min(z)	342.961	2005.662	105.559	125.917	197.299	86.462	99.143
7	max(x)min(z)	3.015	1.862	1.380	1.291	1.343	1.819	3.100
8	max(y)max(x)min(z)	146.883	80.848	43.529	95.349	122.412	35.729	40.032
9	lijphart	11.206	7.000	4.714	4.842	5.319	7.246	11.404
10	augmented lijphart	2.773	1.871	1.377	1.293	1.340	1.800	2.851
11	weighted max(x)min(z)	2.319	1.668	1.277	1.242	1.207	1.611	2.350

Note: The table displays the root mean squared error. Smaller numbers indicate higher reliability.

Appendix Table 2: MC Results Binary Outcome SD(x)=1.0, SD(z)=1.0, corr(x,z)=0, Varying Sample Size N

	Algorithm	N=20	N=40	N=60	N=80	N=100
1	random	3.625	2.637	1.906	2.927	1.638
2	max(y)	3.044	3.130	3.361	10.140	7.093
3	max(x)	0.818	0.825	0.829	0.836	0.831
4	min(z)	3.039	1.745	6.143	4.306	2.881
5	max(y)max(x)	0.767	0.785	0.794	0.804	0.803
6	max(y)min(z)	3.914	3.184	7.053	5.224	5.022
7	max(x)min(z)	0.813	0.805	0.823	0.824	0.826
8	max(y)max(x)min(z)	0.732	0.756	0.778	0.783	0.787
9	lijphart	0.927	0.945	0.909	0.908	0.911
10	augmented lijphart	0.811	0.803	0.819	0.823	0.819
11	weighted max(x)min(z)	0.812	0.807	0.823	0.823	0.825

Note: The table displays the root mean squared error. Smaller numbers indicate higher reliability.

Appendix Table 3: MC Results Binary Outcome SD(x)=1.0, N=100, SD(z)=1.0, Varying Correlation (x,z)

	Algorithm	corr=-0.9	corr=-0.7	corr=-0.3	corr=0	corr=0.3	corr=0.7	corr=0.9
1	random	8.114	2.672	4.221	2.600	1.754	1.308	1.983
2	max(y)	4.651	6.662	3.687	6.796	2.787	3.281	7.919
3	max(x)	0.974	0.924	0.861	0.833	0.813	0.802	0.799
4	min(z)	4.779	3.220	1.789	1.921	2.073	1.742	5.435
5	max(y)max(x)	0.967	0.894	0.830	0.810	0.798	0.799	0.798
6	max(y)min(z)	10.050	4.631	2.729	4.534	3.888	4.997	9.693
7	max(x)min(z)	0.806	0.802	0.820	0.830	0.823	0.818	0.800
8	max(y)max(x)min(z)	0.729	0.755	0.784	0.791	0.778	0.742	0.681
9	lijphart	1.319	1.025	0.942	0.928	0.875	0.965	1.309
10	augmented lijphart	0.768	0.785	0.816	0.823	0.821	0.818	0.801
11	weighted max(x)min(z)	0.801	0.804	0.820	0.830	0.824	0.816	0.795

Note: The table displays the root mean squared error. Smaller numbers indicate higher reliability.

Appendix Table 4: MC Results Binary Outcome SD(x)=1.5, N=100, SD(z)=1.0, Varying Correlation (x,z)

	Algorithm	corr=-0.9	corr=-0.7	corr=-0.3	corr=0	corr=0.3	corr=0.7	corr=0.9
1	random	1.411	1.445	1.794	1.420	2.306	1.835	1.224
2	max(y)	4.220	5.267	3.399	4.789	4.503	3.064	1.634
3	max(x)	0.911	0.901	0.881	0.871	0.866	0.865	0.865
4	min(z)	2.891	1.568	1.556	1.273	1.441	4.080	2.394
5	max(y)max(x)	0.890	0.878	0.869	0.865	0.864	0.865	0.865
6	max(y)min(z)	5.553	15.612	3.498	1.863	1.623	5.266	6.293
7	max(x)min(z)	0.871	0.851	0.865	0.867	0.865	0.854	0.843
8	max(y)max(x)min(z)	0.842	0.837	0.856	0.859	0.855	0.836	0.831
9	lijphart	0.997	0.858	0.858	0.843	0.862	0.929	1.018
10	augmented lijphart	0.780	0.823	0.852	0.859	0.858	0.848	0.839
11	weighted max(x)min(z)	0.812	0.834	0.861	0.865	0.864	0.854	0.833

Note: The table displays the root mean squared error. Smaller numbers indicate higher reliability.

Appendix Table 5: MC Results Binary Outcome SD(x)=0.3, N=100, SD(z)=1.0, Varying Correlation (x,z)

	Algorithm	corr=-0.9	corr=-0.7	corr=-0.3	corr=0	corr=0.3	corr=0.7	corr=0.9
1	random	8.995	8.163	80.732	6.335	7.917	12.190	11.440
2	max(y)	106.287	11.848	19.486	20.691	17.370	10.001	8.574
3	max(x)	1.446	1.297	0.992	0.776	0.589	0.412	0.383
4	min(z)	131.176	12.420	5.636	9.436	6.493	6.991	9.614
5	max(y)max(x)	1.554	1.393	1.003	0.692	0.469	0.348	0.340
6	max(y)min(z)	148.288	21.529	16.576	15.011	11.756	17.360	23.556
7	max(x)min(z)	1.233	0.838	0.777	0.766	0.793	0.838	1.202
8	max(y)max(x)min(z)	1.677	0.790	0.659	0.569	0.632	0.763	1.690
9	lijphart	4.187	2.538	1.642	1.772	2.002	2.222	3.240
10	augmented lijphart	1.199	0.829	0.778	0.764	0.790	0.827	1.106
11	weighted max(x)min(z)	1.129	0.825	0.792	0.769	0.753	0.723	0.876

Note: The table displays the root mean squared error. Smaller numbers indicate higher reliability.