Increasing Inferential Leverage in the Comparative Method: Placebo Tests in Small-n Research[*]

Adam N. Glynn[†] Nahomi Ichino[‡]

This Draft: October 20, 2012

Abstract

We explicitly delineate the underlying homogeneity assumption, procedural variants, and implications of the comparative method [Lijphart, 1975] and distinguish this from Mill's method of difference [1872]. We demonstrate that additional units can provide "placebo" tests for the comparative method even if the scope of inference is limited to the two units under comparison. Moreover, such tests may be available even when these units are the most similar pair of units on the control variables with differing values of the independent variable. Small-n analyses using this method should therefore, at a minimum, clearly define the dependent, independent, and control variables so they may be measured for additional units, and specify how the control variables are weighted in defining similarity between units. When these tasks are too difficult, process tracing of a single unit may be a more appropriate method. We illustrate these points with applications to Epstein [1964] and Moore [1966].

[*] We thank Derek Beach, Kevin Quinn, Rich Nielsen, Noah Nathan, Dustin Tingley, Torben Iversen, John Gerring, Gary King, Michael Hiscox, Kimuli Kasara, Julie Faller, and Anna Grzymala-Busse for helpful comments. Earlier versions of this paper were presented at the 2011 Annual Meetings of the American Political Science Association in Seattle and the Comparative Politics Workshop at Harvard University in March 2011.

[†] Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. aglynn@fas.harvard.edu

[‡] Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. nichino@gov.harvard.edu

1 Introduction

The use of comparison to determine whether an explanatory factor X affects some outcome Y in a pair or small number of cases features prominently in the social sciences [Collier, 1993]. These studies generally select cases to be as similar as possible on the important control variables and have different values of the key explanatory variable. They then examine whether the outcome differs across the cases, appealling to Mill's method of difference [Mill, 1872], the most similar or most similar systems design [Przeworski and Teune, 1970], the comparable-cases strategy or the comparative method [Lijphart, 1971, 1975]. But these methods are problematic in several ways. As Lieberson [1991, 1995] and Sekhon [2004] point out, Mill's methods for small-n comparisons require determinism, absence of measurement error, preclusion of other possible causes of the effect of interest, and lack of interaction effects for valid inferences. Because most social and political processes are unlikely to meet these stringent conditions, Sekhon [2004] suggests moving to a probabilistic framework and to a large-n study in which statistical methods can be applied. This is similar to King et al. [1994]'s prescription for adapting a small-n analysis into a large-n analysis at the subunit level.

But this solution may not be desirable or possible. For a variety of reasons, one may be interested in particular cases or even a specific case, and a move to large-n typically changes the study's goal from a causal effect for a particular unit to an *average* causal effect for many units. Similarly, the move to a subunit analysis typically changes the study's goal from a causal effect for that unit to an *average* causal effect for many subunits. The move to large-n may also not be directly feasible, since the cases available for a large-n study with differing values of the explanatory variable may not be very similar on the control variables [Brady, 2004, 53, Munck, 2004, 113]. After reducing the analysis to those matching cases that are sufficiently similar, we may only have enough cases for a small-n study. When can the comparative method in small-n studies provide leverage for inferring the causal effect of X on Y for these cases, and how should we select our cases for comparison in order to exploit this method?

Confidence in conclusions drawn using any method depends upon the validity of the assumptions underlying the method. We therefore make explicit the homogeneity assumption required for the comparative method, distinguish it from Mill's method of difference, and delineate its three procedural variants. In other words, we lay out the methodology of the comparative method [Lijphart, 1975]. With this framework, we demonstrate that *even if our goal is inference only for the two units being compared and not a larger population*, additional units can provide what is sometimes known as a "placebo test" [Abadie et al., 2010, 2011]. In the most general usage of the term, a placebo test is a secondary test that uses the same logic and preconditions as the primary test, but where it is known that there is no effect. Hence, the placebo test should not detect an effect. If the placebo test detects a non-existent effect, then we should be skeptical of the evidence from the primary test. As we discuss in this paper, the use of the homogeneity assumption in the comparative method means that these placebo tests may be available even when the two units being compared are the most similar pair of units on the control variables with differing values of the explanatory variable. Furthermore, these additional units beyond the scope of inference may have the same value of the explanatory variable X. To our knowledge, this point has never been stated explicitly, although it is implicit in the placebo tests of Abadie et al. [2010] and the placebo tests and p-values suggested in Abadie et al. [2011] and implied by analogous large-n placebo tests.[1] We elaborate on the exact methods for using these additional units through examples and a formal discussion in the following sections.

This surprising source of inferential leverage suggests several changes to current practice for small-n studies using the comparative method. First, studies that employ the comparative method should explicitly utilize additional units beyond those being compared, since comparing one selectively chosen unit to a primary case of interest does very little to strengthen inference. If it is too difficult or costly to assess additional cases beyond the comparison case, then the most appropriate form of analysis may be careful process-tracing [Gerring, 2007, ch. 7], or other methods for establishing the effect of X on Y in that main case, and researchers should not include a second case that adds little to the study.

At a minimum, researchers using the comparative method should provide the details necessary such that future studies on these additional units could provide this supplementary leverage, since the comparisons add little to our confidence in the analysis without these details. Researchers need not conduct numerous additional case studies themselves. But they should provide a roadmap so that other scholars may assess the validity of these comparisons – a new way in which scholars who have deep expertise in cases and regions of the world beyond a study's scope of inference can build on and contribute to that analysis in a cumulative, scientific manner. This means that small-n analyses using the comparative method must first clearly define the dependent, independent, and control variables such that they may be measured for additional units and also specify how the control variables are weighted in defining similarity between units. To this end, we propose a "list, measure, scale, and weight" standard.

Second, the strong assumptions necessary for simultaneously generating and testing theory are unlikely to be satisfied for most social science problems. This means that without specifying the theory prior to the analysis, a contrast with a second case using the comparative method cannot add confidence to our inference for the main case. Therefore, theory testing should be more clearly separated from theory generation than is the usual practice for studies employing the comparative method. The theory may be generated from the intensive study of the primary case of interest, but without inspection of other cases that may be used for testing.

Note that not all small-n comparisons use the comparative method for causal inference, and other small-n methods that may be employed for other goals like theory generation are beyond the

[1] See Imbens [2004] for a summary.

scope of this paper [Collier, 1993]. Other types of small-n comparative studies include the "parallel demonstration of history," in which a set of case studies substantiates the applicability of a general theory in different contexts, and the "contrast of contexts," in which the case studies focus on how unique features of each case affect the way in which general social processes transpire [Skocpol and Somers, 1980]. Our discussion also does not apply to comparisons of causal effects that have been established by other means; it applies only to comparisons that are intended to establish or help establish causal effects.[2]

The paper proceeds as follows. Section 2 describes two methods for small-n comparisons – the method of difference [Mill, 1872] and the comparative method [Lijphart, 1975] – and shows that there are three variants of the comparative method: most similar pair, most similar contrasting case, and sufficiently similar contrasting case. Section 3 elaborates the main points of the paper in a stylized example using the most similar pair design. Section 4 then presents a formal discussion of the comparative method as described by Lijphart [1975], including its assumptions, the implications of these assumptions for inferential leverage, and the relationship between this presentation and previous methodological discussions of Mill's method of difference [Lieberson, 1991, 1994, Sekhon, 2004]. Section 5 demonstrates the inferential leverage gained from additional units when the comparative method is used to find a "most similar" contrasting case to a specific case of interest by revisiting the Canada-United States comparison in Epstein [1964]. Section 6 applies our proposed standards and these methods when there is a "sufficiently similar" contrasting case to a specific case of interest by re-examining the comparison of China with Japan in Moore's *Social Origins of Dictatorship and Democracy* [1966].

2 Method of Difference and Comparative Method

One of the most important uses of comparison in both large-n and small-n studies is to establish causal effects. While having a large number of units may alleviate some of the difficulties of causal inference [King et al., 1994], particularly if a treatment is randomly assigned, there may be circumstances in which comparing a small number of units is necessary or even advantageous [Lijphart, 1975, Skocpol and Somers, 1980, Collier, 1993]. A number of methods have been proposed for such small-n comparisons, and a large portion of these methods involve the comparison of two units with different values of an independent variable. These methods are variously known as the Method of Difference [Mill, 1872], Most Similar Design [Przeworski and Teune, 1970], the Comparable-Cases Strategy [Lijphart, 1975], and the Comparative Method [Lijphart, 1971], among others.

[2] Large-n comparisons can also have alternative goals, and the clearest example may be a multi-site study in which separate randomized experiments are run in two different cities. For this type of study, the randomization and the comparison between treatment and control units within each city establishes the city-specific average causal effects, but we may also want to compare the city-specific effects. The comparison of city-specific effects may be of interest, but it does not establish the causal effect of the cities.

Although these methods are often discussed as equivalent [Lijphart, 1975, Przeworski and Teune, 1970, Sekhon, 2004, Gerring, 2007], there is some ambiguity regarding the exact procedures for each method, which is problematic because different procedures require different assumptions. To start, take John Stuart Mill's canonical statement of the method of difference [Mill, 1872, 452]:

> If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.

As noted in Sekhon [2004], this statement specifies a "process of elimination" procedure, by which theory generation and theory testing take place simultaneously. This statement implies that the two cases might be selected on the basis of the dependent variable and that the cause can be found by searching among a number of potential independent variables. Mill appears to change the implicit procedure on the next page, discussing the method of difference as "a method of artificial experiment" [Mill, 1872, 453], and this ambiguity is well documented in Adcock [2008]. For the purposes of this paper, we understand the method of difference in terms of the canonical quote above.[3] The assumptions necessary for this method have been discussed comprehensively in Geddes [1990], Lieberson [1991, 1994], and Sekhon [2004].

Contrast this with the comparative method, most directly stated by Lijphart [1975, 164]:

> ... the amount of variance of the dependent variables should not be a consideration in the choice of cases because this would prejudge the empirical question. The comparative method can now be defined as the *method of testing hypothesized empirical relationships among variables on the basis of the same logic that guides the statistical method, but in which the cases are selected in such a way as to maximize the variance of the independent variables and to minimize the variance of the control variables.* [emphasis in original]

Lijphart's method proscribes selecting cases on the dependent variable. It also requires that the researcher specify and distinguish between the explanatory variable and the control/matching variables prior to any analysis, and it is therefore clearly a theory testing procedure. We demonstrate in this paper that the assumptions needed for this procedure are weaker than the assumptions needed for the method of difference.

[3] Similarly, Przeworski and Teune [1970, 32] present "Most Similar Systems" Design as theory-generating in a way that seems to imply that the potential cause need not be chosen prior to the comparison: "Such studies are based on the belief that systems as similar as possible with respect to as many features as possible constitute the optimal samples for comparative inquiry." However, they discuss Almond and Verba [1963]'s study of civic culture as an example of this design on the following page, and write, "Almond and Verba chose countries that have a 'democratic political system' but differ with regard to their level of development." Again, this seems to indicate at least the possibility that a study with Most Similar Systems Design might be conducted for causal inference in a manner approximating experiments. Gerring [2007] summarizes the distinction as Most Similar Design for Theory Generation and Most Similar Design for Theory Testing.

Furthermore, Lijphart's statement is an ideal and does not capture the full range of practice with the comparative method. In some applications, one of the cases is pre-selected because that case is of particular interest, and the comparative method is used to select the most similar contrasting case [Nielsen, 2011]. The two cases may not be the pair that "minimizes the variance of the control variables" for such applications. In other applications, the control variables may be difficult to measure for all possible contrasting cases. A contrasting case then cannot be known to be "most similar," and consequently, the implicit claim is that the contrasting case is sufficiently similar for a meaningful comparison. Therefore, Lijphart's comparative method in fact encompasses three different case selection methods: most similar pair, most similar contrasting case, and sufficiently similar contrasting case.

3 A Stylized Example

We now illustrate placebo tests for the comparative method through a hypothetical study of landlocked African countries on the effect of British versus French colonial history on economic growth. In our example, we measure growth as the change in GDP per capita (PPP) for the first and last year of data availability in the Bates Africa data set and dichotomize this to positive or negative growth.[4] We denote the identity of the colonial power with X ($X = 1$ for British and $X = 0$ for French) and economic growth with Y ($Y = 1$ for positive growth and $Y = 0$ for negative growth). For this example, we use only two hypothetical control/matching variables: the percentage of the population that is Muslim (M_1) and the log of average rainfall (M_2). These are presented in a scatterplot in Figure 1(a) so we can visually inspect our data for countries with different values of X that are similar on the control/matching variables. Countries with British colonial history ($X = 1$) are represented by circles while countries with French colonial history ($X = 0$) are represented by triangles.

[Figure 1 about here]

Note that a number of the elements of this example are quite stylized. First, we would certainly want to control for additional variables. Second, with the exception of logging rainfall, the variables have not been recoded to accurately reflect similarity between the cases. For example, a 20 percentage point increase in the share of the population that is Muslim may have different effects for countries with Muslim populations above and below 50%. Third, we have not adjusted the weighting of the variables which would be reflected in distances in the figure. For example, in this presentation, a change of .5 in the log of average rainfall (M_2) covers approximately the same distance as a 20 point change in the percentage of the population that is Muslim (M_1). They indicate approximately equivalent differences in similarity, which may not accurately depict the relative

[4] Data from the Africa Research Program is available at `http://africa.gov.harvard.edu` (accessed August 19, 2011).

importance of rainfall for economic growth. All of these choices were made in order to simplify the presentation and discussion. A proper application would likely make more nuanced decisions, but nevertheless need to make analogous decisions, and these decisions must be explained and defended. The analysis in this section assumes that the researcher has already made and successfully supported such decisions.

Let us stipulate to the accuracy of Figure 1(a) in representing similarity/dissimilarity between cases. First note that there are almost no cases of former British colonies that are similar to former French colonies in the values of the control/matching variables. The only former British and former French colonies that are near one another are Uganda (British, $X = 1$) and the Central African Republic (French, $X = 0$), so that the only comparison that might reasonably be made "controlling" for the share of the population that is Muslim and average rainfall is between these two countries. In the terminology of matching, there is almost no overlap between the former British colonies and the former French colonies. Hence, this example presents the prototypical scenario for an $n = 2$ study with the comparative method. Proceeding with this approach, in Figure 1(b) we have removed all the other countries and indicated values of the outcome variable on the plot, with open symbols representing positive growth and solid symbols representing negative growth. We see that economic growth is positive for Uganda ($Y = 1$) and negative for the Central African Republic ($Y = 0$). Therefore, due to the similarity of these countries on the control variables, we might conclude that the effect of British colonial history was (or would have been) positive for these two countries ($Y_{Uganda} - Y_{CAR} = 1$).

However, a closer look at Figure 1(a) reveals that Lesotho and Zambia, which have been added to Figure 1(c), can provide inferential leverage. Although both have British colonial history ($X = 1$, represented with circles) and neither is very close to any country with French colonial history ($X = 0$), these cases should not have been discarded since they are almost exactly as close together as are Uganda and the Central African Republic. Lesotho and Zambia, which are similar on the control variables, have the same value of X but different outcomes ($Y_{Lesotho} = 1$, $Y_{Zambia} = 0$). Therefore, we become more hesitant to claim that X caused the difference in outcome between Uganda and the Central African Republic, which was also based on a claim of similarity on the control/matching variables.[5] In this sense, Lesotho and Zambia provide what is sometimes known as a placebo test – a test for which our theory predicts no effect or differences in outcome – for the comparison of Uganda and the Central African Republic [Abadie et al., 2010].[6]

[5] One objection may be that similarity should be judged differently in different parts of the plot. However, a "flip" test provides a quick check of this position. The question is whether one would have judged the similarity between Uganda and the Central African Republic to be sufficient if this pair of countries had been located where Lesotho and Zambia are located.

[6] Note that this type of placebo test is subtly different from the analysis suggested in Sekhon [2004], whose approach implies that all of the units be similar on the "background conditions" [286] and that the scope of inference can be thought of as defined by all the units that satisfy those background conditions. In our stylized example, there are three pairs where the units in the pair appear to be similar to each other on the background conditions (i.e., control variables): Uganda/Central African Republic, Lesotho/Zambia, and Zimbabwe/Swaziland. However, unlike

This example has important implications for $n = 2$ comparisons: it demonstrates an additional source of inferential leverage that can be used to assess a comparison of two units, even if we are satisfied with the measurement of the control/matching variables, the coding and scaling of these variables, and how each of these variables are weighted in determining similarity. This is true even when there is only one reasonable matched pair with differing values of the explanatory variable and even when we confine the scope of inference to that matched pair. In other words, information from cases outside of our scope of inference can inform our inference from the comparison of our cases of interest. This implies that, even if we only care about the two particular cases under comparison and the two cases are the most similar pair with $X = 0$ and $X = 1$, we should also examine other cases. This point is discussed formally in the next section, before revisiting Epstein [1964] and Moore [1966] in Sections 5 and 6, respectively, to demonstrate how additional units add inferential leverage to studies that use the other variants of the comparative method.

4 The Comparative Method

Although the method of difference and the comparative method are usually thought to be equivalent, there are three key differences. First, as discussed in Section 2, the Lijphart [1975] procedure is much more clearly about theory testing, with its prohibition on the consideration of variance in the dependent variable and the designation of variables as independent and control prior to case selection. By contrast, Mill's statement of the method of difference appears to define a procedure for simultaneous theory generation and testing. Second, while Mill's statement requires the consideration of "every circumstance," the Lijphart [1975] statement implicitly relaxes this requirement by not asserting that "every" possible control variable must be considered. Third, Mill states that the two cases should have their circumstances "in common," while the Lijphart statement again relaxes this requirement by asserting only that the variance of the control variables be minimized.

The relaxation of Mill's standards is of course necessary for the comparative method to be used in practice. With the exception of precisely controlled laboratory experiments, one would never expect two cases to have "every circumstance in common." However, the relaxation of these standards requires a restatement of the assumptions required for the comparative method. We formalize this within the context of the example from the previous section.

the Skocpol [1979]/Geddes [1990] example used in Sekhon [2004], the pairs are not similar to one another other on the background conditions and lie in different parts of the matching space. Furthermore, only one of these pairs contains a country with $X = 0$, so that the scope of inference for our stylized example is defined by the Uganda/Central African Republic pair – we know nothing about the effect of X for the Lesotho/Zambia and Zimbabwe/Swaziland pairs. Nevertheless, the Lesotho/Zambia and Zimbabwe/Swaziland pairs can serve as placebo tests for our Uganda/CAR comparison.

4.1 The Homogeneity Assumption

Suppose as in Section 3 that we are willing to confine inference to the two cases under comparison, Central African Republic and Uganda in our stylized example. We typically want accurate estimates of the case-level causal effects – that is, the causal effect for each of these particular cases. This requires an assumption of unit homogeneity between the two cases [Holland, 1986, King et al., 1994].[7] Intuitively, we must assume that the most similar matched pair has unit i and unit j that are similar enough on the control/matching variables such that if they had the same value of X (i.e., if both had $X = 0$ or if both had $X = 1$), then they would both have had the same value of Y. Formally, if $X_i = 0$ and $X_j = 1$, then the observed outcomes can be written as $Y_i = Y_i(0)$ and $Y_j = Y_j(1)$, and the counterfactual outcomes can be written as $Y_i(1)$ and $Y_j(0)$. Unit homogeneity between the pair implies that the observed Y_i serves as a proxy for the counterfactual for case j ($Y_i = Y_j(0)$) and the observed Y_j serves as a proxy for the counterfactual for case i ($Y_j = Y_i(1)$).[8]

Using our stylized example, consider Table 1 which presents the observed variables, potential outcomes under both observed and counterfactual colonial histories, and the case-level causal effects. Without assuming homogeneity, the observed outcome for Uganda is also the potential outcome for Uganda with a treatment of British colonial history ($Y_{Uganda} = Y_{Uganda}(1)$), but we do not know the potential outcome for if Uganda had instead been a French colony ($X = 0$, $Y_{Uganda}(0) =?$). Therefore, we do not know the effect of X on Y for Uganda ($Y_{Uganda}(1) - Y_{Uganda}(0) =?$). Similarly, while the observed outcome for the Central African Republic is also the potential outcome for the Central African Republic with a treatment of French colonial history ($Y_{CAR} = Y_{CAR}(0)$), we do not know the potential outcome for if the Central African Republic had instead been a British colony ($X = 1$, $Y_{CAR}(1) =?$).

[Table 1 about here]

However, if we assume that Uganda and the Central African Republic are similar enough on values of the control/matching variables (M_1 and M_2) to be homogenous, then the observed Y value for Uganda serves as a proxy for the counterfactual Y value for if the Central African Republic had had $X = 1$ ($Y_{Uganda} = Y_{CAR}(1)$). Similarly, the observed Y value for the Central African Republic serves as a proxy for the counterfactual Y value for if Uganda had had $X = 0$ ($Y_{CAR} = Y_{Uganda}(0)$). This homogeneity assumption implies that the estimate of $Y_{Uganda} - Y_{CAR} = 1 - 0$ equals the effect of British colonial history for both Uganda ($Y_{Uganda}(1) - Y_{Uganda}(0)$) and the Central African Republic ($Y_{CAR}(1) - Y_{CAR}(0)$).

[7] The necessary assumptions change if we are willing to estimate only the average causal effect for the two cases. However, we have never seen the average causal effect for two cases cited as the effect of interest for a small-n study. One of the putative reasons for performing a small-n study is to avoid the necessity of having to estimate averages.

[8] King et al. [1994] actually utilizes probabilistic counterfactuals and an assumption of mean homogeneity. The assumptions of the comparative method can be weakened in this manner, but the conclusions will also be weakened. The placebo tests discussed in the next section will still provide evidence with regard to the validity of the comparison, but a formal presentation of such a framework is outside of the scope of this paper.

4.2 Implications of the Homogeneity Assumption

Using similarity to justify the homogeneity assumption for cases i and j has implications beyond units i and j. In particular, we must consider where else in the space of the control/matching variables we would have judged the i, j pair to be similar enough for homogeneity. This standard might change across the matching space; for example, we might not have considered Uganda and the Central African Republic to be similar enough if they had been 80% Muslim. However, such gradations should be decided prior to examining the data.

If we can define the region in which the similarity of the i, j pair would have been judged sufficient, then the homogeneity assumption has implications for other pairs of units in this region. If within this region we find a pair of units k and l that are as similar as the i, j pair, then the assumption of homogeneity for the i, j pair implies an assumption of homogeneity for the k, l pair. Furthermore, if the k, l pair both have the same value of X, then homogeneity is a testable assumption for this pair.[9]

Using our example, if we would have applied the same similarity standard to the Uganda and Central African Republic pair in the 6.5 to 7 region for rainfall, then the assumption of homogeneity for the Uganda/Central African Republic pair implies homogeneity for the Lesotho/Zambia pair. However, both Lesotho and Zambia have $X = 1$, and therefore homogeneity not only implies that the potential outcomes for Lesotho and Zambia should be the same ($Y_{Lesotho}(1) = Y_{Zambia}(1)$), but that the observed outcomes for Lesotho and Zambia should be the same ($Y_{Lesotho} = Y_{Zambia}$). This assumption can be tested. As we see in Figure 1(d), Lesotho has positive growth ($Y_{Lesotho} = 1$)

[9] While between-unit homogeneity implies the existence of placebo tests, note that this homogeneity does not necessarily imply some of the assumptions that are needed for the use of Mill's method of difference as a process of elimination of possible causal factors, without pre-selection of the causal factor of interest [Lieberson, 1991, 1994, Sekhon, 2004].

First, causal homogeneity between one or some pairs of cases does not necessarily imply causal homogeneity between all of the cases. We might initially assume that Uganda and the Central African Republic are homogenous and that Lesotho and Zambia are homogenous, but this does not imply that Zambia and the Central African Republic are homogenous. Therefore the homogeneity assumption does not imply causal determinism in the sense of Lieberson [1991, 1994], unless all of the cases have sufficiently similar values of the control variables.

Second, causal homogeneity does not imply a lack of interaction. For example, suppose that along with $X = 0$, case i also has $Y = 0$, and that along with $X = 1$, case j also has $Y = 1$. Therefore, if these cases are causally homogenous with respect to the effect of X on Y, then X is seen to have a positive effect on Y for these cases. Suppose further that both cases have $Z = 1$, but that if case i had had $Z = 0$, then it would have had $Y = 1$, and similarly, if case j had had $Z = 0$, it would have had $Y = 0$, and therefore if Z had been 0, the effect of X on Y would have been negative. This also assumes that the causal homogeneity with respect to the effect of X on Y would be maintained when $Z = 0$. There is an interaction in this scenario since the effect of X on Y is positive when $Z = 1$ and negative when $Z = 0$. We would be unable to discover this interaction without additional causally homogenous cases, but this does not preclude its existence.

Third, this example of an interaction demonstrates that causal homogeneity does not preclude the existence of another cause of Y. Because case i which has $X = 0$ took the value $Y = 0$ when $Z = 1$ and took the value $Y = 1$ when $Z = 0$, we might say that Z has a negative effect on Y for case i when $X = 0$. Furthermore, case j which has $Z = 1$ took the value $Y = 1$ when $Z = 1$ and $Y = 0$ when $Z = 0$, so we might say that Z has a positive effect on Y when $X = 0$ for case j. Therefore, X and Z both have causal effects on Y. The fact that we will not be able to discover the effects of Z on Y from just the observation of cases i and j and the assumption of homogeneity with respect to the effect of X on Y for these cases does not preclude that Z potentially has an effect on Y.

while Zambia has negative growth ($Y_{Zambia} = 0$), so the assumption is falsified. Therefore, the Lesotho/Zambia pair provides a placebo test for the Uganda/Central African Republic pair.

5 "Most Similar" Contrasting Case

Applications of the comparative method often do not begin with the selection of the most similar pair of cases with differing values of the explanatory variable. More typically, researchers start with a case of interest and generate a theory about that case with an intensive within-case study. The theory is tested using within-case information that was not used for theory generation and is then further tested by selecting a contrasting case. What additional leverage does a contrasting case bring to this type of analysis?

In order for the comparative method with a most similar contrasting case to be valid and for the contrasting case to provide leverage, we must assume homogeneity of the potential outcomes, like for the comparative method with the most similar pair.[10] Even though these studies are often labeled "most similar" design, a most similar contrasting case may not be the most similar or perhaps not even very similar to the case of interest, and there are likely to be other cases that are more similar to one another on the control variables and have the same value of the independent variable. If such additional cases are available, placebo tests as in the stylized example in the previous sections can help assess this homogeneity assumption and extent of the leverage provided by the contrasting case.

We demonstrate this with Epstein [1964], which assesses the argument that the "uncohesive and nonresponsible character of American parties" ($Y_{US} = 0$) is due to the separation-of-powers system of government ($X_{US} = 0$). This argument has been made with a comparison to Britain with its parliamentary system ($X_{Br} = 1$) and cohesive parties ($Y_{Br} = 1$), but Epstein uses Canada as a contrasting case to the United States. Using Epstein's control/matching variables, we show that while Canada is likely the most similar case to the United States, Britain and perhaps Australia are more similar to Canada than Canada is to the United States. We further show how our confidence in Epstein's argument is strengthened by the inferential leverage provided by these additional units.

Epstein [1964] begins by discussing the similarity between the United States and Canada: both are socially and culturally diverse; cover large land areas; are federal systems in which the states or provinces have substantial powers; have similar social and economic class structures; and have single-member, simple plurality election systems [Epstein, 1964, 46–8]. By comparing the United States to Canada, which has a British-style parliamentary system with an executive responsible to a popularly elected legislature ($X_{Can} = 1$) [48] and cohesive legislative parties ($Y_{Can} = 1$) [52],

[10] If inference is confined to the causal effect for the case of interest, then only partial homogeneity is necessary (i.e., unit j provides the counterfactual for unit i but unit i need not provide the counterfactual for unit j). We have never seen this type of partial homogeneity assumption invoked, but it would require a specific argument as to why similarity on the control variables implies homogeneity for one value of the independent variable (e.g., $X = 1$), but not the other (e.g., $X = 0$).

Epstein concludes that the system of government has an effect on the cohesiveness of political parties ($Y_{Can}(1) - Y_{US}(0) = 1 - 0 = 1$). These variables are summarized in Table 2.

[Table 2 about here]

Although Epstein is commendably explicit in listing his control/matching variables, he does not comment on the weight that should be assigned to each control/matching variable. Such weights are unnecessary for his argument that Canada represents a better match for the United States than Britain does for the United States if the United States is more similar to Canada than to Britain on *all* of the control/matching variables. But they are necessary for checking the homogeneity assumption. Even if Canada is closer to the United States than is Britain, the Canada/Britain comparison would provide a placebo test for the Canada/United States comparison if, as depicted in Figure 2, Canada is as close or closer to Britain than it is to the United States.[11] Canada may be closer to Britain overall even if it is more similar only on one of the control/matching variables, if this variable is weighted sufficiently highly.

[Figure 2 about here]

Social and cultural diversity (M_1 in Table 2) is one variable on which Canada is arguably more similar to Britain than to the United States. Epstein [1964] likens the presence of the Francophone minority population in Canada to the African-American population in the United States and describes it a "divisive force" [47], but the Francophones' position in Canadian politics and society appears closer to that of the Scottish and Welsh populations in Britain. In both Britain and Canada, these minority populations are linguistically and culturally distinct from the majority population and still dominate today in areas they occupied prior to the arrival of the majority population. Moreover, these minority populations in Britain and Canada do not have the legacy of being enslaved by the majority population. If this control/matching variable were weighted sufficiently heavily, Canada would still be more similar to the United States than is Britain, but more similar to Britain than to the United States.

With this weighting, the comparison of Britain and Canada provides a placebo test, since they must be causally homogeneous if the United States and Britain, which are less similar to each other, are assumed to be causally homogeneous. Because Britain and Canada have the same value of the explanatory variable (parliamentary system, $X = 1$) and the same value of the outcome variable (cohesive legislative parties, $Y = 1$), causal homogeneity between Britain and Canada is not invalidated. Consequently, Epstein [1964]'s implicit assumption that Canada and the United States are causally homogeneous is also not invalidated, bolstering our confidence in his argument.

This example highlights the importance of making explicit the list, measurement, scaling and weighting of the control/matching variables. Even though Canada appears to be the case that is

[11] We may also want to conduct this placebo test if Canada is almost as close to Britain as it is to the United States, but for simplicity we omit this from the discussion.

most similar to the United States on each of these variables, it may be that Canada is more similar to Britain or another country depending on how those control/matching variables are weighted for an overall assessment of similarity. Indeed, Canada may also be more similar to Australia than to the United States, so that the Australia/Canada comparison may also be a placebo test for Epstein's argument for the United States.[12] But to make such a claim and conduct a placebo test, as we did using Britain, it is necessary for the list, measurement, scaling and weighting of the control/matching variables to be clearly specified. It is not possible to gain inferential leverage from additional units without this information.[13]

The existence of these placebo tests implies that a rigorous approach to the comparative method must consider the similarity between many more pairs of units than are implied by Lijphart [1975]. The most direct way to accomplish this is to provide a method for assessing similarity that can be replicated by other researchers, and the "list, measure, scale and weight" standard we discuss here and is implied by large-n matching procedures is one option.[14] In the next section, we demonstrate that when this standard is not satisfied, it is impossible to assess the inferential leverage provided by the contrasting case.

6 "Sufficiently Similar" Contrasting Case

While the Epstein [1964] study is a straightforward example of the comparative method for theory testing, it is often difficult to tell whether small-n comparative studies use comparisons for theory generation, theory testing, or both. For example, Barrington Moore uses numerous comparisons in *Social Origins of Dictatorship and Democracy* [1966] in presenting his argument for how the landed upper classes and the peasantry affected the development of modern democratic, communist, and fascist regimes by the mid-twentieth century. He justifies his use of comparisons by writing that "a comparative perspective can lead to asking very useful and sometimes new questions. There are further advantages. Comparisons can serve as a rough negative check on

[12] Although Australia is less socially and culturally diverse than Canada or the United States and it uses preferential voting (instant run-off voting) rather than simple plurality rule for most of its elections, it is covers a large land mass and has a similar social and economic class structure to Canada. Moreover, as Epstein notes himself, "having ten rather than 50 regional governments puts Canada in the same category with most federal systems outside the Western Hemisphere," among which is Australia with six states and two territories [47]. If Epstein had proposed weighting this control/matching variable very heavily, Australia may be more similar overall to Canada than Canada to the United States.

[13] Note that this contrasts with Gerring [2007], who in discussing most similar design with the Epstein [1964] example, argues that "it is not usually necessary to *measure* control variables (at least not with a high degree of precision) in order to control for them. If two countries can be assumed to have similar cultural heritages, one needn't worry about constructing variables to measure that heritage. One can simply assert that, whatever they are, they are more or less constant across the two cases... This can be a huge advantage over large-N cross-case methods, where each case must be assigned a specific score on all relevant control variables – often a highly questionable procedure" (emphasis in the original) [133].

[14] Nielsen [2011] provides a blueprint for using large-n matching procedures which satisfy these criteria to find the most similar contrasting case for a small-n study.

accepted historical explanations" [xix]. Moreover, Skocpol and Somers [1980] characterize Moore's use of comparative history as "primarily for the purpose of making causal inferences about macro-level structures and processes" [181] and present his use of "negative cases" as an application of Mill's method of difference [183]. As discussed in Section 4, Mill's method of difference requires a number of assumptions beyond homogeneity for simultaneous theory generation and theory testing. If homogeneity does not hold, then neither the comparative method nor Mill's method of difference will produce a meaningful comparison.

We consider Moore's use of Japan as a contrasting case to China and for illustrative purposes proceed as if the comparative method were being used for theory testing. But, as we will elaborate, Japan may not be the most similar contrasting case for China, and since Moore makes no explicit claims to this effect, his implicit argument is that Japan is a "sufficiently similar" contrasting case for China. It is rare in practice to be able to make the case that a contrasting case is "most similar" as in Epstein [1964], since evaluating the control/matching variables for all possible contrasting cases is usually prohibitively costly and requires expertise in many areas. We are often only able to examine a subset of cases and/or variables, so that implicit arguments that the cases being compared are "sufficiently similar" are quite common.

In Moore's use of Japan as a contrasting case to China, the fundamental question is: to what extent should the comparison with Japan increase our confidence in Moore's argument about China? Unfortunately, we are unable to definitively answer this question, because Moore does not provide the roadmap that would be necessary to determine (1) whether there are any more similar contrasting cases that would produce a different conclusion, or (2) whether there are any more similar pairs with the same value of the independent variable that could function as placebo tests.

The major difficulty is that Moore [1966] does not clearly list all of his explanatory variables or distinguish between his independent variables and his control/matching variables. Nor are guidelines provided for how the variables should be measured or how the control/matching variables should be scaled or weighted for an overall assessment of similarity between these cases. For this analysis, we rely and build upon Skocpol [1973]'s distillation of three "explanatory variable clusters" from *Social Origins*. We also infer how the variables should be measured for other possible cases and incorporate our own ideas on how the variables might be scaled and weighted in establishing overall similarity from our reading of Moore [1966].

With the variables from Skocpol [1973] and our delineation of independent and control/matching variables, we argue below that Choson (today's Korea) is more similar to China than is Japan to China and show that Choson provides inferential leverage by casting doubt on the validity of Moore's China/Japan comparison. We further demonstrate that Choson need not be more similar to China than is Japan to invalidate this comparison, by considering foreign threats in the nineteenth century as an additional control/matching variable. The overall conclusion is that Moore's contrast with Japan adds little confidence to his argument for China beyond his case study of China.

6.1 Similarity with Moore's Variables

Following his case studies, Moore describes his complex argument for the emergence of democratic, fascist, and communist regimes in three thematic chapters in Part III of *Social Origins*. In a critical review of this work, Skocpol [1973] organizes Moore's overall argument into three explanatory variable clusters – bourgeois impulse, mode of commercial agriculture, and peasant revolutionary potential. Her summary of Moore's coding of the cases on these variables is partially reproduced in Table 3. These three factors shape which classes ally with each other as the economy begins to modernize, whether there is a peasant revolution, and ultimately what type of modern regime emerges.

[Table 3 about here]

To apply the comparative method to Moore's China case study and its contrast with Japan, we must first define and separate the three explanatory variable clusters into independent variables and control/matching variables. As coded by Skocpol [1973], mode of commercial agriculture is certainly one of the control/matching variables, because both China and Japan are coded as labor-repressive. Peasant revolutionary potential is certainly one of the independent variables because China is coded as high while Japan is coded as low. It is less clear whether bourgeois impulse, for which China is coded as weak while Japan is coded as medium-strength, is a secondary independent variable or a control/matching variable on which the two cases are not perfectly matched.

We treat bourgeois impulse as a control/matching variable for two reasons. First, it is difficult to distinguish between bourgeois impulse and mode of commercial agriculture or determine how to measure them. The mode of commercial agriculture generally refers to whether the upper classes rely on the market or use "political" and more traditional social means to supply labor to work its land holdings [Moore, 1966, 433]. Skocpol relies on "scattered remarks" to figure out how to measure the strength of bourgeois impulse and "wonder[s] if these implicit criteria were applied independently of results, or consistently" [Skocpol, 1973, 13]. We were similarly uncertain of Moore's definition of bourgeois impulse, although we determine that the strength of bourgeois impulse is associated with the growth of cities, the rise of an urban commercial and manufacturing class, and demand from commodity markets. We are also uncertain about when to measure these variables, which is problematic because these features of a country's political economy and society can be very different over the long run.

Both bourgeois impulse and mode of commercial agriculture are rooted in the relationship between the state, the landed upper class, and the overall socio-economic and political system, and consequently, these variables are hard to disentangle. In China, for example, upper class families invested in the classical Confucian education of a son to sit for exams to join the imperial bureaucracy, with the understanding that this investment would be "recouped" through the land and wealth to be acquired through his appointment [Moore, 1966, 165]. When growing commerce threatened the economic and social status of this scholar-landlord class, the imperial bureaucracy generally tried to "absorb and control commercial elements" [175] through taxation and the establishment of monopolies. The availability of this route to wealth through the state "deflected ambitious individuals away from commerce" [174], contributing to the weakness of bourgeois impulse. Moreover, the "labor repressive" mode of commercial agriculture, or "political methods... [that kept] the peasants at work" [180], were a crucial part of "making property pay" [181],[15] both attracting and supporting the landed upper class.

Second, Moore emphasizes peasant revolutionary potential when accounting for why no peasant revolution took place and consequently no communist regime emerged in Japan [254–5]. This peasant revolutionary potential is more theoretically distinguishable from bourgeois impulse and mode of commercial agriculture. While these two clusters mostly concern the relationship between the landed upper classes and the state, peasant revolutionary potential focuses on the relationship between the landed upper classes and the peasants, whose roles in the peasant revolution leading to communism Moore seeks to understand [xxiii]. Moore defines peasant revolutionary potential as "the weakness of the institutional links binding peasant society to the upper classes, together with the exploitative character of this relationship" [Moore, 1966, 478], but provides few guidelines for its measurement.

Having designated the independent and control/matching variables in this way, we can consider the Japan/China comparison. We denote peasant revolutionary potential as X and whether there is a communist regime as Y using notation from the previous sections. With Skocpol [1973]'s coding of peasant revolutionary potential as either high ($X = 1$) or low ($X = 0$), China can be characterized as $X_C = 1$, while Japan can be characterized as $X_J = 0$. The logic of the comparison here is that, despite not being perfectly matched on the control/matching variables, Japan is "sufficiently similar" to stand in for a China with low peasant revolutionary potential.

Choson (today's Korea) may provide evidence against the sufficiency of this similarity. As in China, the monarchy and the landed upper class (*yangban*) in Choson relied heavily on tenant farmers for income and tax revenues [Shin, 1998]. The landed upper class acquired and maintained its land holdings through having family members qualify for state administrative positions through examinations on Confucian scholarship as in China. Moreover, the "Yi Dynasty (1392–1910) consciously tried to form its administration according to a neo-Confucian interpretation of ancient Chinese works and had more continual and stronger Chinese influence on the shape of the state than did Japan" [Sorensen, 1984, 306].[16] The overall relationship of the landed upper classes to the state in Korea was more like that of the scholar-landlords in China than that of the non-landed

[15] Moore [1966] does not provide specific details on these political methods in China.

[16] Note that one way in which Japan was more similar to China than was Choson to China is that both China and Japan had emperors, while the Korean king was officially subordinate to the Chinese emperor. The Korean king did not have all the symbols of legitimacy and power available to the Japanese or Chinese emperor [Palais, 1975, 10]. But this factor is irrelevant for Moore's argument, so it does not contribute to our assessment of similarity or dissimilarity between the cases.

warrior aristocracy to the centralized feudal state in Japan. Therefore, although we do not have a good understanding of the coding of bourgeois impulse, we assign Korea a value of moderately-weak in contrast to Japan's value of medium. This assessment of the similarity of the cases is depicted in Figure 3.[17]

[Figure 3 about here]

However, despite the similarities between China and Korea on the control/matching variables, Korea appears to be more similar to Japan on the independent variable, peasant revolutionary potential. In China, peasant revolutionary potential is high ($X_C = 1$) because, "[t]he government and the upper classes performed no function that the peasants regarded as essential for their way of life" [205]. Institutional links between peasant society and the landed upper classes were too weak to absorb the pressures from its exploitative nature [478]. At the same time, "solidary arrangements ... constitute[d] focal points for the creation of a distinct peasant society in opposition to the dominant class and [served] as the basis for popular conceptions of justice and injustice that clash[ed] with those of the rulers" [479].

In Japan, peasant revolutionary potential is low ($X_J = 0$) because there was a "close link between the peasant community and the feudal overlord, and his historical successor the landlord" [254]. Moore reports that the gentry provided some relief in times of poor harvests and that "the Japanese peasant community provided a strong system of social control that incorporated those with actual and potential grievances into the *status quo*" [254]. Moreover, irrigation and rice planting required cooperation among villagers [263–4], and the tax system created the "tightly knit character of the Japanese village... [which] tied the peasants closely to their rulers" [258–9].

In Choson/Korea, peasant revolutionary potential is also low ($X_K = 0$) because village organization connected landlords and tenants and funneled tenant grievances through established channels. Rural Korean society had various institutions led by the rural elite such as village compacts (*tongyak*), credit rotating systems (*kye*), labor reciprocating systems (*pumashi*), and lineage associations (*munjung*). The rural elite also "maintained close ties with [the cultivators] because they...[lived] in the countryside with other economically less fortunate residents" [Kim, 2007, 997]. The Yi dynasty had a system of village organization like the Chinese *pao-chia*, but they emphasized the sub-village neighborhood unit rather than the larger village unit. Their responsibilities were more limited, so that they "counterbalance[d] whatever strength the villages might have developed" [Eikemeier, 1976, 108].

Since Choson/Korea is more similar on the control/matching variables to China and has the same value of the independent variable as Japan, it appears to be a more similar contrasting case to China than is Japan. We code the outcome variable for Korea as $Y_K = 0/1$, but without clear

[17] Furthermore, although peasants were tied to the land in Japan, such that the mode of commercial agriculture may be coded as "labor repressive," the tax regimes left Japanese peasants with more of the agricultural surplus than their counterparts in China and stimulated increased agricultural production [254].

guidance from Moore on when this variable should be measured, it is difficult to code the outcome variable for Korea which becomes communist only in its northern half in the second half of the twentieth century. This coding for after the Korean War (1950–53) is compatible with the time period used by Moore to code China as communist ($Y_C = 1$), even though Moore's discussion of China's peasant revolution only extends to the 1930s, since the Chinese Communists did not consolidate their control over the country until the Nationalists were driven from the mainland in 1949. This is also consistent with a coding of the outcome in Japan as not communist ($Y_J = 0$), although inconsistent with Moore's coding of the outcome in Japan as fascism, since by 1949, the fascist regime in Japan had been defeated and replaced by American military occupation. Japan's industrialization and modernization through a "revolution from above" also began decades earlier. However, it is clear that when restricting the analysis to the variable clusters presented in Skocpol [1973], comparing China ($Y_C = 1$) with Choson/Korea ($Y_K = 0/1$) leads to a less certain conclusion about the effect of peasant revolutionary potential than does comparing China with Japan ($Y_J = 0$).

6.2 Similarity including Foreign Threat

Our assessment of similarity between the cases and hence the inferences drawn on the basis of similarity may change if we consider additional control/matching variables. This section considers one such additional factor: the extent of foreign threat in the nineteenth century (labeled M_3).[18] China, Japan, and Choson all faced threats from foreign powers that triggered major social and political changes. The First Opium War (1839–42) made clear to the Qing dynasty the backwardness of the Chinese military, and Chinese losses led to the recognition of extraterritoriality and the transfer of Hong Kong to the British Empire. The arrival of American ships on Japanese shores in 1853 forcefully ended Japan's isolationist foreign policy and was the start of unwanted engagement with several western powers demanding the opening up of trade. Similarly, Choson had cut off almost all international relations as the "hermit kingdom" until the 1860s when the Russians, French, British, Americans, and Japanese opened up trade and established extraterritoriality through military invasions and gunboat diplomacy. On this control/matching variable, Choson is more like Japan than China in that these foreign threats were sudden, major disruptions to centuries of isolation. By contrast, China had been the predominant power in East Asia with active military campaigns to extend its empire and tributary state system. If we rank the three countries on this variable, we would find China with the lowest foreign threat, Japan with moderately high foreign threat, and Choson with high foreign threat.

[Figure 4 about here]

[18] We have restricted M_3 to foreign threats in the nineteenth century and not considered foreign threats or intervention in later periods. Uncertainty about the timing of the independent variable means that later foreign threats may be part of the causal pathway from independent variable to outcome variable. The inclusion of such a "post-treatment" variable will complicate inference about the effect of our independent variable.

Figure 4 presents the matching space, with Moore's control/matching variables ($w_1M_1+w_2M_2$) on the x-axis, as in Figure 3. We have not explicitly chosen the weights w_1 and w_2, because all three countries are coded the same on M_2 (labor-repressive mode of commercial agriculture) and hence the weights do not affect the scaling on the x-axis.[19] Japan, Korea, and China keep the same location on the x-axis in Figures 4 from Figure 3, but differ in the extent of nineteenth-century foreign threat (M_3) which is represented on the y-axis. With just our ordinal rankings of the countries on $w_1M_1 + w_2M_2$ and M_3, we can only speculate as to the exact locations of the three countries in Figure 4. Our guesses place Japan closer to China than Korea is to China in the matching space, so that Japan is a better match for China than is Korea. This can be seen in Figure 4, with Japan lying on the dotted arc centered at China while Korea lies beyond it. It also places Korea closer to Japan than Japan is to China, as shown by Korea's positioning inside the dashed arc centered at Japan and passing through China.

Even though Japan is more similar to China than is Korea in this arrangement, the Korea/Japan comparison still provides inferential leverage and casts doubt on the homogeneity assumption required for the China/Japan comparison. Korea and Japan, which are more similar than are China and Japan and have the same value of the independent variable X, do not have the same outcome Y. From this failed placebo test, we would conclude that Japan is likely not sufficiently similar to China for a meaningful comparison and that the contrast with Japan adds little to our confidence in Moore's argument for China.

This example highlights that without a clear roadmap from Moore, we cannot determine what placebo tests are possible or the extent to which the comparison with Japan should increase our confidence in Moore's argument about China. The analysis required us to make a number of assumptions about how to measure the dependent variable, what was the independent variable, what control variables should be included, how they should be measured and scaled, and what relative weights should be accorded them. The locations of the cases in Figure 4 reflect our assumptions, which allowed us to use Choson/Korea to evaluate whether Japan/China is a meaningful comparison.

7 Conclusion

Increasing the number of observations is often recommended to address limitations to small-n studies that use the comparative method for causal inference [King et al., 1994, 208, Lijphart, 1975, 163]. However, this is an impracticable or undesirable solution in some circumstances. One such circumstance is when there is only one pair of cases with $X = 1$ and $X = 0$ that are similar on the control/matching variables, as in the stylized example of landlocked African countries in Section

[19] However, these weights are constrained by the weight on M_3 relative to the other control/matching variables given implicitly by distances in our figure. All of these weights would matter if we were to consider additional cases that have a different value on M_2.

3. Increasing the number of pairs with differing values of the independent variable would only have reduced the validity of this analysis. Another is when we seek causal effects for individual cases, as for the United States in Epstein [1964] or for China in Moore [1966]. Without additional assumptions, increasing the number of pairs with differing values of the independent variable would provide more leverage only for average causal effects and not for case-specific causal effects.

By clarifying the assumptions necessary for the comparative method, we show that even in these situations, placebo tests may be available such that units beyond those under comparison can add inferential leverage to the analysis. Lesotho and Zambia, a pair of former British colonies, cast doubt on inference from the Uganda/Central African Republic comparison in our stylized example. The explicit comparison of Britain with Canada conferred greater confidence in the conclusions drawn from the comparison of the United States and Canada by Epstein [1964]. Finally, the comparison of countries with low peasant revolutionary potential, Korea and Japan, made us reconsider Moore [1966]'s use of Japan as a contrast to China.

As noted above, these placebo tests point to a previously unrecognized way in which scholars of comparative politics with regional or area expertise can contribute to the broader discipline. Such experts frequently offer proximity to causal mechanisms, better measurement, and sensitivity to the question of whether a general theory applies in particular contexts. But such scholars may also utilize their expertise in evaluating the validity of small-n comparisons that do not include a case from their own region, if similarity should be judged the same way in their own region as in a given study. That is, an Africanist might point out that the similarity argument being made to justify the comparison of two Latin American countries would imply that two sub-Saharan African countries that are at least as similar should have the same outcome – but perhaps do not. Such observations by a specialist on sub-Saharan Africa that falsify homogeneity would be useful even if the scope of inference of a given study is limited to these two Latin American countries.

The possibility of these tests also enjoins researchers to take two steps beyond the current practice for small-n studies of specifying the control/matching variables that justify the selection of the cases to be compared. First, they require that the dependent, independent, and control variables be clearly defined and delineated before the analysis, and that researchers provide enough information such that these variables can be measured on additional units. Second, they require that researchers explicitly state how the control/matching variables are weighted in making a claim of similarity. If these steps are too onerous, then the comparison case adds little inferential value and should not be included in the study. In other words, the study should focus on establishing the causal effect for the main case through other techniques and not include a second case just to be comparative. But wherever possible, a researcher using the comparative method should follow these steps to allow others to conduct placebo tests to build on his analysis. These steps will ultimately improve the credibility of conclusions drawn from small-n comparative studies.

References

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative Politics and the Synthetic Control Method. MIT Political Science Department Research Paper No. 2011–25, 2011.

Robert K. Adcock. The Curious Career of 'the Comparative Method': The Case of Mill's Methods. Presented at the Annual Meeting of the American Political Science Association, Boston, 2008.

Gabriel A. Almond and Sidney Verba. *The Civic Culture: Political Attitudes and Democracy in Five Nations*. Princeton University Press, Princeton, 1963.

Henry E. Brady. Doing Good and Doing Better: How Far Does the Quantitative Template Get Us. In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, pages 53–68. Rowman & Littlefield Publishers, 2004.

Henry E. Brady and David Collier, editors. *Rethinking social inquiry : diverse tools, shared standards*. Rowman & Littlefield Publishers, Lanham, MD, 2004.

David Collier. The comparative method. In Ada W. Finifter, editor, *Political Science: The State of the Discipline II*. American Political Science Association, Washington, D.C., 1993.

Dieter Eikemeier. Villages and Quarters as Instruments of Local Control in Yi Dynasty Korea. *T'oung Pao, Second Series*, 62(1/2):71–110, 1976.

Leon D. Epstein. A Comparative Study of Canadian Parties. *American Political Science Review*, 58(1):46–59, 1964.

Barbara Geddes. How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics. *Political Analysis*, 2:131–52, 1990.

John Gerring. *Case Study Research: Principles and Practices*. Cambridge University Press, New York, 2007.

Geoffrey Hawthorn. The United States in South Korea. In *Plausible worlds: possibility and understanding in history and the social sciences*, pages 81–122. Cambridge University Press, New York, 1991.

Paul W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–60, 1986.

Guido W. Imbens. Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

Sun Joo Kim. Taxes, the Local Elite, and the Rural Populace in the Chinju Uprising of 1862. *Journal of Asian Studies*, 66(4):993–1027, 2007.

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

Stanley Lieberson. 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, 1991.

Stanley Lieberson. More on the Uneasy Case for Using Mill-Type Methods in Small-N Comparative Studies. *Social Forces*, 72(4):1225–37, 1994.

Arend E. Lijphart. Comparative Politics and the Comparative Method. *American Political Science Review*, 65(3):682–93, 1971.

Arend E. Lijphart. II. The Comparable-Cases Strategy in Comparative Research. *Comparative Political Studies*, 8(2):158–77, 1975.

John Stuart Mill. *A system of logic, ratiocinative and inductive: being a connected view of the principles of evidence and the methods of scientific investigation.* Longmans, Green, Reader and Dyer, London, 8th edition, 1872.

Barrington Moore, Jr. *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World.* Beacon, Boston, 1966.

Gerardo Munck. Tools for Qualitative Research. In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, pages 105–22. Rowman & Littlefield Publishers, 2004.

Richard A. Nielsen. Case Selection via Matching. Working paper, Harvard University, 2011.

James B. Palais. *Politics and policy in traditional Korea.* Harvard University Press, Cambridge, MA, 1975.

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

Jasjeet S. Sekhon. Quality meets Quantity: Case Studies, Conditional Probability, and Counterfactuals. *Perspectives on Politics*, 2(2):281–93, 2004.

Gi-Wook Shin. Agrarian Conflict and the Origins of Korean Capitalism. *American Journal of Sociology*, 103(5):1309–51, 1998.

Theda Skocpol. A Critical Review of Barrington Moore's Social Origins of Dictatorship and Democracy. *Politics & Society*, 4(1):1–34, 1973.

Theda Skocpol. *States & Social Revolutions: A Comparative Analysis of France, Russia, & China.* Cambridge University Press, New York, 1979.

Theda Skocpol and Margaret Somers. The Uses of Comparative History in Macrosocial Inquiry. *Comparative Studies in Society and History*, 22(2):174–97, 1980.

Clark Sorensen. Farm Labor and Family Cycle in Traditional Korea and Japan. *Journal of Anthropological Research*, 40(2):306–23, 1984.

(a) Without Homogeneity Assumption:

	Observed		Potential Outcomes		Causal Effect
Case	X	Y	Y(1)	Y(0)	$Y(1) - Y(0)$
Uganda	1	1	1	?	1 - ?
CAR	0	0	?	0	? - 0

(b) With Homogeneity Assumption:

	Observed		Potential Outcomes		Causal Effect
Case	X	Y	Y(1)	Y(0)	$Y(1) - Y(0)$
Uganda	1	1	1	0	1 - 0
CAR	0	0	1	0	1 - 0

Table 1: Observed variables (X and Y), potential outcomes ($Y(1)$ and $Y(0)$), and causal effects ($Y(1) - Y(0)$) for Uganda and the Central African Republic, with and without an assumption of homogeneity. The estimated effect is $Y_{Uganda} - Y_{CAR} = 1 - 0$. However, for each case, we observe only one of the potential outcomes, so in order for the estimated effect to be accurate for the causal effect for Uganda, we must assume that $Y_{CAR} = Y_{Uganda}(0)$, and in order for the estimated effect to be accurate for the causal effect for the Central African Republic, we must assume that $Y_{Uganda} = Y_{CAR}(1)$.

	United States		Canada		Britain
Y: Cohesiveness of political parties	Low		High		High
T: System of government	Separation of powers		Parliamentary		Parliamentary
M_1: Social and cultural diversity	African-Americans and slavery legacy	$\sim\sim$	Francophones in Quebec	$\sim$	Scots, Welsh in home regions
M_2: Land area	Large	$\sim$	Large	$\sim\sim$	Small
M_3: Federal	Federal	$\sim$	Federal	$\sim\sim$	Unitary
M_4: Class structure	Democracy before industrialization	$\sim$	Democracy before industrialization	$\sim\sim$	Past feudalism, industrialization first
M_5: Electoral system	SMD, plurality	$\sim$	SMD, plurality	$\sim$	SMD, plurality

Table 2: Variables in Epstein [1964] Study.

	Japan	Choson/Korea	China
Y: Communist regime	No (0)	Partial (0/1)	Yes (1)
T: Peasant revolutionary potential	Low (0)	Low (0)	High (1)
M_1: Bourgeois impulse	Medium	Moderately Weak	Weak
M_2: Mode of commercial agriculture	Labor-repressive	Labor-repressive	Labor-repressive

Table 3: Summary of Moore [1966], based upon Skocpol [1973], including our own coding of Choson/Korea.

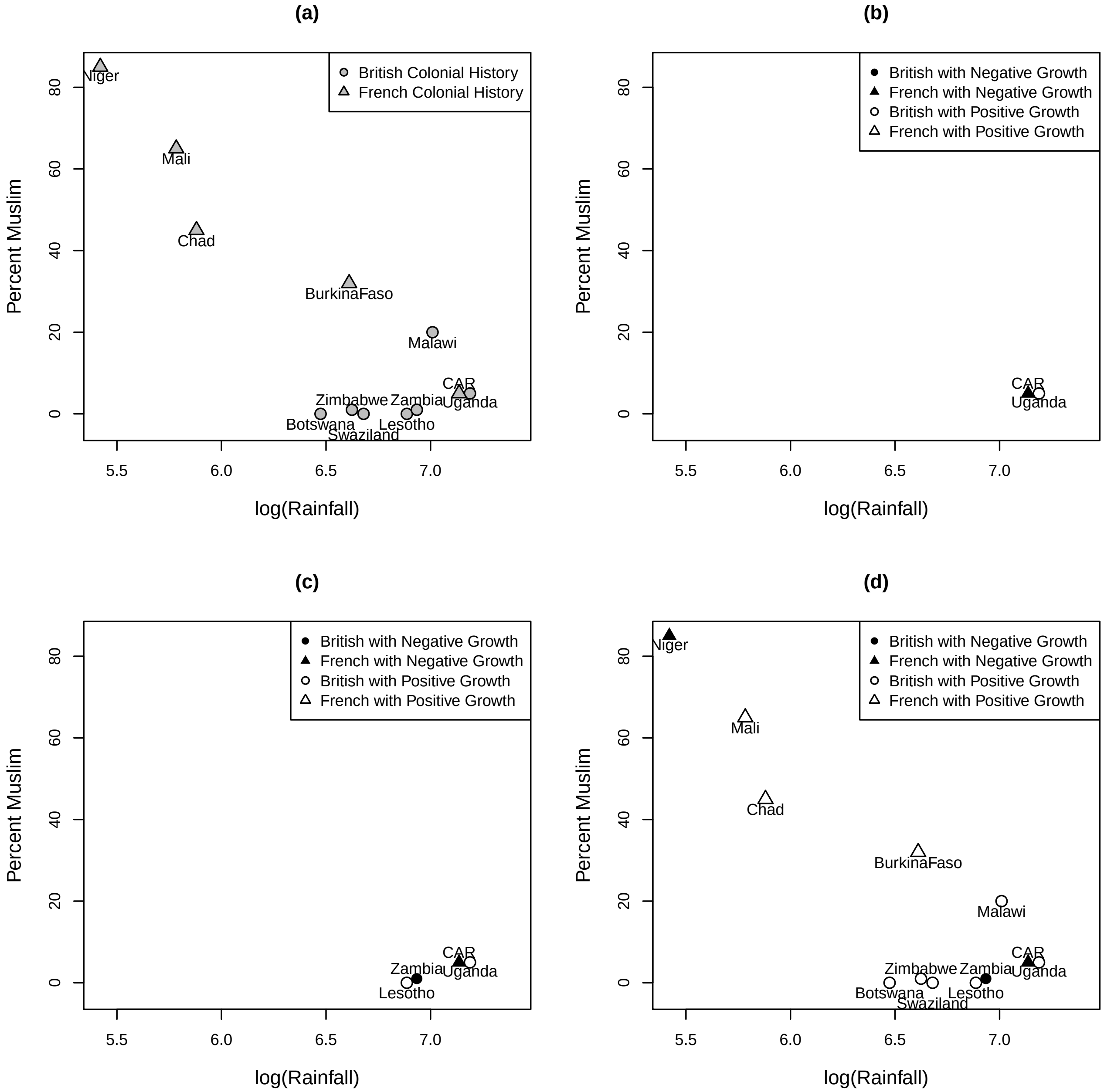

Figure 1: *A Stylized Example for the Effect of British or French Colonial History on Economic Growth*

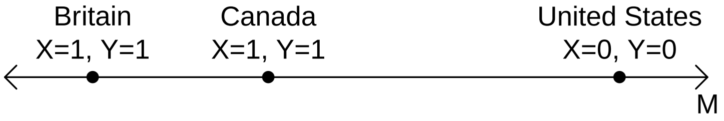

Figure 2: *Example of "Most Similar" Contrasting Case Design, Epstein [1964].* Epstein's matching variables are weighted such that Canada is more similar than Britain to the United States and is more similar to Britain than to the United States.

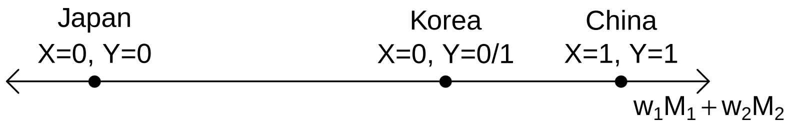

Figure 3: *Example of "Sufficiently Similar" Contrasting Case Design, Moore [1966].* Skocpol's summary of Moore's control/matching variables are weighted such that Korea is more similar to China than is Japan. The case of interest for this study is China.

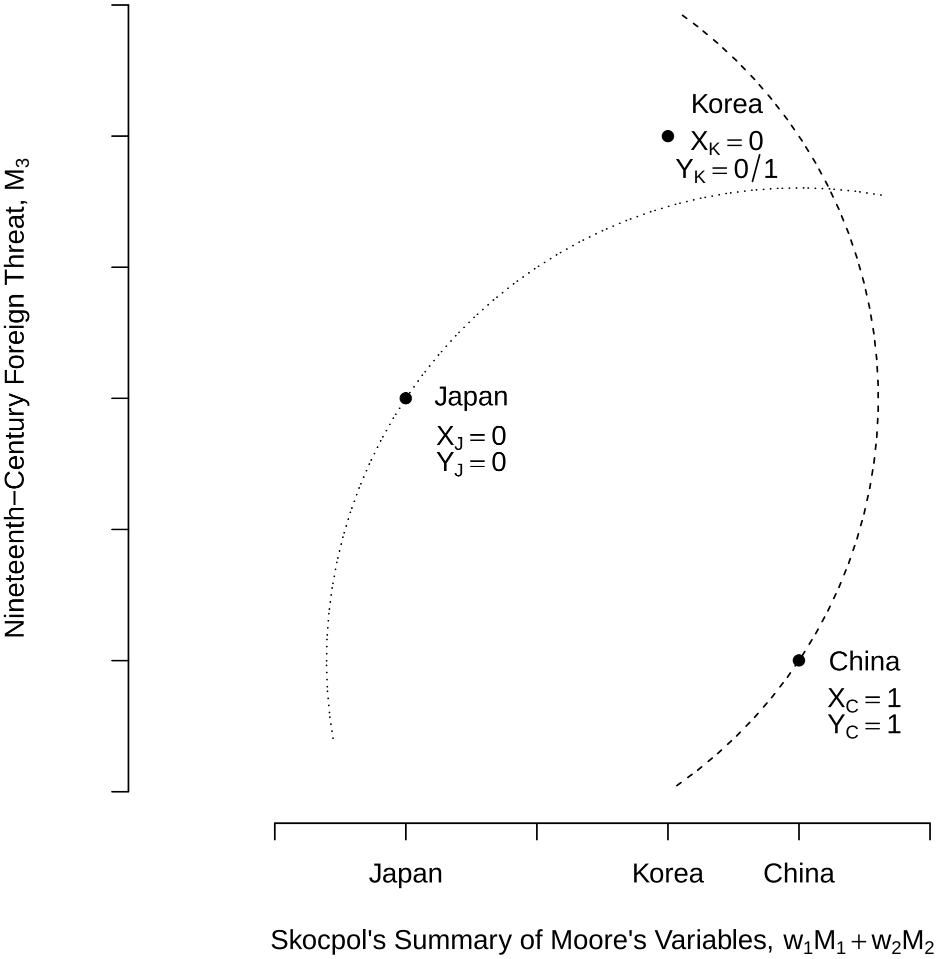

Figure 4: *Moore [1966], with Nineteenth-Century Foreign Threat on the y-axis.*