Case Selection, Case Studies, and Causal Inference: A Symposium

David Collier
University of California, Berkeley

Fall 2008, Vol. 6, No. 2

Qualitative & Multi-Method Research

Newsletter of the American Political Science Association Organized Section for Qualitative and Multi-Method Research

Contents

Symposium: Case Selection, Case Studies, and Causal Inference

Introduction
David Collier 2
Do the N's Justify the Means?
David A. Freedman 4
Response to David Freedman
James D. Fearon and David D. Laitin 6
Techniques for Case Selection: A Response to Freedman
John Gerring 7
Choosing Cases for Case Studies: A Qualitative Logic
Gary Goertz 11
Rejoinder
David Freedman 14

Case Selection, Case Studies, and Causal Inference: A Symposium

David Freedman, author of the opening and closing contributions to this symposium, passed away on October 17, 2008. A Professor of Statistics at the University of California, Berkeley, Freedman strongly believed that case knowledge and qualitative evidence are crucial to causal inference. An important statement of his view, noted elsewhere in this issue of the newsletter, is found in Freedman, "On Types of Scientific Enquiry: The Role of Quantitative Reasoning" (*Oxford Handbook of Political Methodolgy,* 2008).

Introduction

David Collier
University of California, Berkeley
dcollier@berkeley.edu

For scholars concerned with causal inference, how should cases be selected in case study research?

This symposium builds on previously published arguments by James Fearon and David Laitin (2008), who favor random sampling in case study analysis, and by John Gerring (2008), who favors purposive selection. The statistician David Freedman-long an advocate of case studies as an important research tool-comments on these published arguments; responses are offered by Fearon-Laitin and by Gerring; Gary Goertz adds a commentary of his own; and then Freedmanoffers concluding remarks.

In Fearon and Laitin's (2008) discussion, the goal is to draw insights about causal mechanisms from case studies so as to illuminate the findings from a large-N, regression-type analysis. The idea of random sampling is of course central to the broad literature on statistical inference, and for Fearon and Laitin a key advantage of this approach is to prevent scholars from deliberately selecting cases favorable to their preferred hypotheses, thus engaging in "cherry picking."

By contrast, in advocating purposive selection Gerring (2008) draws on the tradition that reaches back at least to understandings of case studies offered by Lijphart (1971), Eckstein (1975), and George (1979). Gerring's approach employs a large-*N* framework, which he uses to identify cases that are seen as typical, diverse, extreme, deviant, influential, crucial, pathway, most similar, and most different.

Yet another perspective, introduced in this symposium by Gary Goertz, likewise advocates purposive selection for case-study research aimed at causal inference. Goertz is primarily interested in the case studies in their own right, rather than their role in statistical analysis involving a large *N*. Goertz's point of departure is the cross-tabulation of two dichotomies (the outcome to be explained and the potential explanation), and his discussion of case selection focuses on choices among the cells in the resulting 2 x 2 table. This approach connects with the wider tradition of analyzing matching and contrasting cases, identified in different ways with the methods of agreement and difference of J. S. Mill (1974 [1843]), most-similar and most-different designs of Przeworski and Teune (1970), and Qualitative Comparative Analysis (Ragin 1987; see also 2000).

Freedman extends, refines, and in some ways departs from the above approaches. His overall position is to prefer purposive selection. For case-study analysis that is concerned with checking models employed in large-*N* research, he recommends a focus on cases consistent with predictions of the model, cases not consistent with its predictions, and influential cases that appear to have an especially strong effect on findings derived from the model.[1]

Among the issues discussed in this symposium, I find three to be of special interest. First, the idea of random sampling from a well-defined population is a gold standard for descriptive inference, and quite properly so. However Freedman suggests that this standard is less frequently—and less effectively—met than is often believed. Observational studies in the social sciences often involve some variant of a convenience sample. This certainly would appear true in macro-comparative research, as with a focus on the OECD countries. In that instance, one may have a convenience sample (driven in part by the availability of excellent data), but the presumed population of interest may never be clearly defined. Even with a random sample, Freedman points out that major problems of missing data can weaken inferences from sample to population. A statistical model may be necessary in correcting for potential bias due to missing data, yet this model may well introduce more bias than it removes. Freedman argues that as a consequence, a random sample can pose just as many uncertainties about generalization as a convenience sample.

In my view, the better part of wisdom may be to recognize that, under some (possibly many) circumstances, we should drop the pretense that we engage in random sampling from a defined population. Being realistic, departures from this (obviously useful) gold standard occur frequently. These considerations should, at least some of the time, lead scholars to be more cautious about undertaking generalization, and the expression "external validity" may sometimes raise higher expectations for achieving valid generalization than are warranted or appropriate. It is often more productive to pursue contingent generalizations by seeking to map findings from a particular set of cases onto carefully specified additional cases (possibly including, in international studies, additional world regions).

Second, Freedman agrees that cherry-picking should be avoided. However, he notes that until the scholar has actually done the case study research, it is often hard to know how cases will come out. This uncertainty makes it less likely that the researcher can intentionally select cases that support a preferred hypothesis. I am reminded of Donald Campbell's (1975) argument that the findings of case studies routinely go in a different direction than the researcher expects before starting the investigation. Cherry-picking may thus not be as grave a problem as the vivid metaphor suggests.

These comments about our weak prior knowledge of how particular cases will actually come out are certainly relevant to Goertz's focus on selecting cases from particular cells within his 2 x 2 table. How does one know in which cell the cases will be located? One solution is suggested by Gerring's approach. He uses large-*N* regression-type analysis—based on what is doubtless a more preliminary and imprecise coding of cases—in initially situating the cases; this is subsequently to be followed by the fine-grained coding that researchers can achieve, based on their close case knowledge.

Third, Freedman has long argued that descriptive findings are too often interpreted as causal relationships, with far too little attention to the fragility of causal inference. Correspondingly, descriptive findings may be given an importance that, taken by themselves, they do not deserve.

This perspective leads Freedman to note with concern the opening statement in Fearon-Laitin (2008), where they say that "almost by definition, a single case study is a poor method for establishing whether or what empirical regularities exist across cases. To ascertain whether some interesting pattern, or relationship between variables, obtains, the best approach is normally to identify the largest feasible sample of cases relevant to the hypothesis…."

Freedman comments that if these empirical regularities and relationships among variables are of interest because they contribute to causal inference, then this approach is too rigid, inappropriately devalues case studies, and fails to recognize the very different paths that can be followed in inferring causation. Freedman sees case studies as making diverse contributions: they can "overturn prior hypotheses, generate new lines of inquiry, or confirm causal claims." The empirical regularities that emerge in case studies may lack the presumed generality of those derived from large-*N* analysis. Yet they may contribute just as much because they rest on what may be the considerably greater power of insight derived from close case knowledge.

In conclusion, as Freedman puts it in his final remarks, this debate "has a happy ending." Any apparent disagreement with Fearon and Laitin over random selection and case studies is resolved through the exchange in this symposium. More broadly, there would doubtless be a consensus among the contributors that, as Freedman puts it,

> (i) There are many ways to do good science. (ii) In particular, neither cluster of methods has a general advantage over the other. (iii) Therefore, there are many fruitful ways for qualitative and quantitative researchers to interact.

Standards can and should be applied in evaluating alternative causal claims, but there is certainly no single method through which this analytic task should be accomplished.

Notes

[1] Here and elsewhere in this symposium, "model" refers to a statistical model, and should not be confused with a game-theoretic model. A statistical model is understood as a set of one or more mathematical equations—commonly regression equations—used in the analysis of empirical data. Among many purposes, a model may be employed in descriptive inference, as in an inference from a sample to a population, and in causal inference. As Freedman emphasizes in this symposium, descriptive inference faces numerous challenges, and he has argued in many publications that causal inference based on statistical models is very fragile indeed. Both of these points are crucial to the present discussion.

References

Box-Steffensmeier, Janet M., Henry E. Brady, and David Collier, eds. 2008. *Oxford Handbook of Political Methodology*. Oxford: Oxford University Press.

Campbell, Donald T. 1975. "'Degrees of Freedom' and the Case Study." *Comparative Political Studies* 8:2 (July), 178–93.

Eckstein, Harry. 1975. "Case Study and Theory in Political Science." In Fred I. Greenstein and Nelson W. Polsby, eds., *Handbook of Political Science*, Vol. 7. Reading, MA: Addison-Wesley.

Fearon, James D. and David D. Laitin. 2008. "Integrating Qualitative and Quantitative Methods." In *The Oxford Handbook of Political Methodology.* Janet Box-Steffensmeier, Henry E. Brady, and David Collier, eds. (New York: Oxford University Press), 756–76.

Freedman, David A. 2008. "On Types of Scientific Enquiry: The Role of Qualitative Reasoning." In *The Oxford Handbook of Political Methodology.* Janet Box-Steffensmeier, Henry E. Brady, and David Collier, eds. (New York: Oxford University Press), 300–18.

George, Alexander L. 1979. "Case Studies and Theory Development: The Method of Structured, Focused Comparison." In Paul Gordon Lauren, ed., *Diplomacy: New Approaches in History, Theory and Policy*. New York: Free Press.

Gerring, John. 2008. "Case Selection for Case-Study Analysis: Qualitative and Quantitative Techniques." In *The Oxford Handbook of Political Methodology.* Janet Box-Steffensmeier, Henry E. Brady, and David Collier, eds. (New York: Oxford University Press), 645–84.

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

Mill, John Stuart. 1974b [1843]. "Of the Four Methods of Experi mental Inquiry." In Book 3, Chapter 8, *A System of Logic,Raciocinative and Inductive*. Toronto: University of Toronto Press.

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

Ragin, Charles C. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: University of

Do the N's Justify the Means?

David A. Freedman
University of California, Berkeley

Box-Steffensmeier, Brady, and Collier (2008) examine the craft of political science from a rich variety of perspectives. I will comment on two chapters, one by Fearon and Laitin, the other by Gerring. These chapters are well reasoned, but reach opposite conclusions about a basic issue—how should cases be chosen? Fearon and Laitin focus on large-*N* research, with a logit model for civil war to illustrate the argument. To see whether causal inferences from the model hold up under closer scrutiny, they choose a sample of cases for detailed investigation ("multi-method research").

Fearon and Laitin say, "An important but neglected problem for this research approach is the question of how to choose the cases for deeper investigation....We propose that choosing cases for closer study *at random* is a compelling complement in multi-method research to large-*N* statistical methods in its ability to assess regularities and specify causal mechanisms" (758).

By contrast, according to Gerring, "In order to isolate a sample of cases that both reproduces the relevant causal features of a larger universe (representativeness) and provides variation along the dimensions of theoretical interest (causal leverage), case selection for very small samples must employ purposive (non-random) selection procedures" (645).

In short, Fearon and Laitin recommend sampling cases at random, whereas Gerring recommends purposive selection. To be sure, Gerring's main interest is choosing cases for small-*N* research, but his reasoning applies equally well to the multi-method research discussed by Fearon and Laitin. I will not resolve the conflict here, although I will make some suggestions. The essays raise other important questions about research methodology, and I will also comment on those.

At the outset, Fearon and Laitin make three valuable points. (1) Scholars can be remarkably, let's say, innocent when describing research designs and case selection. (2) "Cherry-picking cases" (by which Fearon and Laitin mean picking cases that support a particular line of argument) is often a bad idea. (3) Random sampling precludes cherry-picking.

An emphasis on choosing cases purely at random, however, may be misplaced. By now, fitting models to data is routine, and there are any number of well-intentioned software packages that automate large parts of the activity. With a totally random sample of cases, the likely finding is that the sample follows the trends predicted by the model.[1] After all, large-*N* scholars choose models that do a good job of tracking the data. (And if the first model they try doesn't work, they might go for a second model—or a third.)

When Fearon and Laitin get down to business, they choose a stratified random sample—stratified not only by explanatory variables (region) but also by the outcome variable (presence or absence of civil war). So the methodological advice amounts to this: within strata, choose your cases at random.

The advice is excellent, if you have a lot of cases in each stratum, and can afford a sample of reasonable size. But how does it help someone who does qualitative research where the number of cases is strictly limited? On the other hand, for model-checking in large-*N* research, I would recommend taking (i) some cases that are consistent with the predictions of the model, (ii) some that are inconsistent, and (iii) some from strata of special interest. Cases that markedly influence results should be considered too. Finally, random sampling is good and cherry-picking is bad—unless, of course, you want to make an existence proof or an argument a fortiori.

Fearon and Laitin conclude that studying the sample cases is a useful extension of the statistical modeling and suggests "a natural way that qualitative work might be integrated into a research program as a complement to rather than as a rival or substitute for quantitative analysis" (774–75). Folding qualitative work into a quantitative research program is an idea, but a general recommendation seems premature—especially when the evidence consists of a case study with *N* = 1, namely, their own investigation of civil war.

Fearon and Laitin also make an interesting comparison between small-*N* and large-*N* research methods: "Almost by definition, a single case study is a poor method for establishing whether or what empirical regularities exist across cases. To ascertain whether some interesting pattern, or relationship between variables, obtains, the best approach is normally to identify the largest feasible sample of cases relevant to the hypothesis or research question, then to code cases on the variables of interest, and then to assess whether and what sort of patterns or associations appear in the data" (757, footnote omitted).

The claimed superiority of large-*N* methods is obviously right if "empirical regularities" are statistical measures of association, like regression coefficients. The thesis is less obvious if "empirical regularities" are defined more broadly, so as to include (for example) causal relationships. Then Fearon and Laitin's "best approach" seems too rigid. In fact, a lot of good science gets done rather differently (Freedman 2008a).

Therefore, I suggest taking a more liberal view of the relationship between qualitative and quantitative research. Qualitative methods can overturn prior hypotheses, generate new lines of inquiry, or confirm causal claims. Indeed, large-*N* research is often done to confirm insights generated by case studies (Freedman 2008a). It should be common ground, however, that the best research programs combine qualitative and quantitative methods.

Looking beneath the surface of a statistical model is hard work, and requires intellectual fortitude for that reason among others. Fearon and Laitin looked, using an elegant and systematic technique, and reported what they saw. This might be an example worth following.

I turn now to Gerring. His Table 1 lists a variety of methods for case selection. Most of the suggestions are helpful, as is the accompanying discussion. However, some entries in the table are puzzling. For example, the table recommends the hat matrix and Cook's distance, which measure in different ways how each observation influences the regression outputs. Such measures might be helpful when selecting cases to probe a large-*N* model. For qualitative research, however, regression output seems irrelevant. The table also recommends discriminant analysis and factor analysis. But these are large-*N* techniques, pure and simple—or impure and madly complicated, depending on one's perspective.[2]

Gerring proceeds to make a strong claim about case selection: "The most useful statistical tool for identifying cases for in-depth analysis in a most-similar setting is probably some variety of matching strategy—e.g., exact matching, approximate matching, or propensity-score matching" (670, footnote omitted).

It is hard to see how techniques like propensity-score matching apply to small-*N* research.[3] Even for large-*N* research, the claim ignores abundant evidence on the fallibility of matching techniques. I agree that matching may have a role to play, but suggest that caution is in order.

Gerring raises broader issues that should be addressed too. For example, he says: "In large-sample research, the task of case selection is usually handled by some version of randomization" (645). I disagree. *Some* large-*N* research is based on randomized experiments or probability samples, but most is not. Convenience samples and observational studies are far more typical, with statistical models to address selection effects and confounding.

The difficulties with the modeling approach are well known (Berk 2004; Brady and Collier 2004; Freedman 2005; Mahoney and Rueschemeyer 2003). Of course, there will always be those who can ignore the difficulties. See, for instance, King, Keohane, and Verba (1994).

Gerring also has something noteworthy to say about experiments: "[i] in a randomized experiment…the researcher typically does not attempt to measure all the factors that might affect the causal relationship of interest. [ii] She assumes, rather, that these unknown factors have been neutralized across the treatment and control groups by randomization or by the choice of a sample that is internally homogeneous" (670).

Point (i) is correct. Point (ii) is off the mark. If the experiment is properly done, few assumptions are needed, because randomization guarantees that the treatment and control groups are balanced on the average. That is why experiments give unbiased estimates of causal effects.[4] Furthermore, there is no need to choose "a sample that is internally homogeneous"—which is all to the good, since that task is beyond our present capabilities.

When the computer actually prints out the random numbers that define the treatment and control groups, there will be minor imbalances due to the play of random chance. These imbalances are the source of random errors in estimates derived from the data. The impact of random errors is conventionally measured by standard errors and *P*-values. It is the randomization that justifies the conventional measures. Without the randomization, justification might be elusive.[5]

The *intention-to-treat principle* is to compare rates or averages for those assigned to treatment with those assigned to control. That is the tacit premise of the discussion. If regression adjustments are made to compensate for imbalances between groups, or to correct for crossover, matters become substantially more complicated (Freedman 2006, 2008b, 2008c, 2008d).

The logic of randomized controlled experiments is worth understanding, for two reasons at least. (i) Experiments are the gold standard for causal inference. (ii) The statistical methods used to analyze observational studies usually depend on the assumption that in some respect or another, the observational study at hand is like an experiment. The logic of randomized controlled experiments is therefore central, even for observational research.

How to choose cases? This question has intrigued scholars from John Stuart Mill onwards, perhaps because the answer depends on context. Any additional clarity is to be welcomed, and Gerring has provided more than a little. So have Fearon and Laitin.

On the other hand, some of the methods that Gerring proposes are ill-suited to qualitative research. Furthermore, he mistakes the role of random sampling and experimentation in large-*N* research, and fails to recognize the limits of other large-*N* techniques. Fearon and Laitin seem at times to imply that qualitative methods are useful only as checks on quantitative results. Such a perspective would undervalue contributions made by small-*N* methods. More generally, that kind of perspective ignores a crucial point: there are many ways to do good science.

Notes

[1] Fearon and Laitin show that close examination of typical cases (countries with no civil war and low probability of civil war according to the model) can be illuminating—a special and valuable feature of their research. Indeed, they use the cases to check the qualitative implications of their causal model.

[2] Seawright and Gerring (2008) give a clearer account of the matter, indicating that the relevant population must be large.

[3] The setting for propensity-score matching is usually an observational study where subjects self-select into one of two conditions; call these "treatment" and "control." The first step is usually to estimate the conditional probability that a subject winds up in treatment, given the covariates. Logit models are often used. This is not an activity to be undertaken with a small sample. For empirical evidence on the weaknesses of matching designs in large-*N* research, see for instance Arcenaux, Gerber, and Green (2006), Glazerman, Levy, and Myers (2003), Peikes, Moreno, and Orzol (2008), Wilde and Hollister (2007). For additional discussion pro and con, see *Review of Economics and Statistics* 86:1 (February 2004); *Journal of Econometrics* 125:1–2 (March-April 2005).

[4] Suppose, for instance, that we have an experimental population of 1,000 subjects, with 400 chosen at random and assigned to treatment; the remaining 600 are the controls. Each subject has two potential responses: one is observed if the subject is assigned to treatment, and the other if assigned to control. The average response of the 400 is an unbiased estimate of what the average would be if all 1,000 subjects were assigned to treatment. Likewise, the average response of the 600 is an unbiased estimate of the average response if all 1,000 subjects were controls. The general principle is this: with a simple random sample, the sample average is an unbiased estimate of the population average. For additional details, see Freedman (2006).

[5] For example, see Freedman, Pisani, and Purves (2007). Chapter 27 discusses experimental comparisons; technical detail is provided in A31–36. Chapter 29 explains what happens without randomization; also see Freedman (2008e).

References

Arceneaux, Kevin, Alan S. Gerber, and Donald P. Green. 2006. "Comparing Experimental and Matching Methods Using a Large-Scale Voter Mobilization Experiment." *Political Analysis* 14: 37–62.

Berk, Richard A. 2004. *Regression Analysis: A Constructive Critique*. Sage Publications.

Box-Steffensmeier, Janet M., Henry E. Brady, and David Collier. 2008. *The Oxford Handbook of Political Methodology*. New York: Oxford University Press.

Brady, Henry E. and David Collier. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman and Littlefield.

Fearon, James D. and David D. Laitin. 2008. "Integrating Qualitative and Quantitative Methods." In *The Oxford Handbook of Political Methodology.* Janet Box-Steffensmeier, Henry E. Brady, and David Collier, eds. (New York: Oxford University Press), 756–76.

Freedman, David A. 2005. *Statistical Models: Theory and Practice*. New York: Cambridge University Press.

Freedman, David A. 2006. "Statistical Models for Causation: What Inferential Leverage Do They Provide?" *Evaluation Review* 30: 691–713.

Freedman, David A. 2008a. "On Types of Scientific Enquiry: The Role of Qualitative Reasoning." In *The Oxford Handbook of Political Methodology.* Janet Box-Steffensmeier, Henry E. Brady, and David Collier, eds. (New York: Oxford University Press), 300–18.

Freedman, David A. 2008b. "On Regression Adjustments in Experiments with Several Treatments." *Annals of Applied Statistics* 2: 176–96.

Freedman, David A. 2008c. "Randomization Does Not Justify Logistic Regression." *Statistical Science* 23: 237–49.

Freedman, David A. 2008d. "Survival Analysis: A Primer." *The American Statistician* 62: 110–19.

Freedman, David A. 2008e. "Oasis or Mirage?" *Chance* 21: 59–61.

Freedman, David A., Robert Pisani, and Roger A. Purves. 2007. *Statistics*. 4th edition. New York: W. W. Norton & Company, Inc.

Gerring, John . 2008. "Case Selection for Case-Study Analysis: Qualitative and Quantitative Techniques." In *The Oxford Handbook of Political Methodology.* Janet Box-Steffensmeier, Henry E. Brady, and David Collier, eds. (New York: Oxford University Press), 645–84.

Glazerman, Steven, Dan M. Levy, and David Myers. 2003. "Nonexperimental versus Experimental Estimates of Earnings Impacts." *Annals of the American Academy of Political and Social Science* 589: 63–93.

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

Mahoney, James and Dietrich Rueschemeyer. 2003. *Comparative Historical Analysis in the Social Sciences*. New York: Cambridge University Press.

Peikes, Deborah N., Lorenzo Moreno, and Sean Michael Orzol. 2008. "Propensity Score Matching: A Note of Caution for Evaluators of Social Programs." *The American Statistician* 62: 222–31.

Seawright, Jason and John Gerring. 2008. "Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options." *Political Research Quarterly* 61: 294–308.

Wilde, Elizabeth T. and Robinson Hollister. 2007. "How Close Is Close Enough? Evaluating Propensity Score Matching Using Data from a Class Size Reduction Experiment." *Journal of Policy Analysis and Management* 26: 455–77.

Response to David Freedman

James D. Fearon
Stanford University
jfearon@stanford.edu

David D. Laitin
Stanford University
dlaitin@stanford.edu

As Jonathan Swift made mockingly clear, "modest proposals" that purport to solve previously unyielding problems can have horrible implications. Such proposals should be subjected to skeptical analysis. So we are pleased that our proposed random method of case selection for the qualitative component of multi-method research has attracted some skeptical commentary in the research community in political science.[1] And we are very grateful to David Freedman for providing a perspective on our approach. He is especially qualified to do so, as he is a leading statistician who has long worried about inflated claims for statistical methods in the social sciences, and has been a champion of approaches that are sensitive to the particularities of each datapoint.

We completely agree with Freedman's claim that there are many ways to do good social science. Indeed, as Freedman quotes us, we argued that the random narratives approach is "a compelling complement" to large-*N* research. This is more modest than Freedman's implication that we believe we have discovered the one true path for multi-method research. In fact, if everyone did random narratives, there would be no expert narratives for the research community to consult!

Furthermore, as Freedman points out, the method has been applied only to our work on civil war onsets. Perhaps it will not be the best approach for other questions that scholars want to use multiple methods to address. We agree, although one goal of our article was to argue that there are good *a priori* (or theoretical) reasons to think that the approach could be valuable for research designs on topics other than civil war onset. Multi-method and other social science research inevitably involves a process of going back and forth between theory and data (despite the pristine hypothesis-testing scenario assumed in statistics textbooks). The random narratives approach is a way to discipline and make more productive this back-and-forth process in a fairly typical social science setting, where one has cross-sectional or panel data with which to document empirical patterns, and historical materials available to investigate causal mechanisms in particular cases.

In the case of our work on civil war, we constructed a country/year dataset with civil war onset as the dependent variable.[2] We estimated a statistical model that identified several correlates of civil war onsets for which we proposed possible causal interpretations. The interpretations were based on a reading of the statistical results *and* our previous knowledge of a set of cases well known to us. To look at those same cases for qualitative support for a causal interpretation would have been intellectual double-dipping. The method of randomly selecting cases for analysis of causal mechanisms behind peace or war onset helped us to avoid or at least reduce this bias. But we certainly do not maintain that random selection of cases for detailed analysis would always be the most effective and efficient approach in a multi-method research project, independent of the subject matter or the stage of the research (in terms of "back and forth").

Freedman writes that our "claimed superiority of large-*N* methods is obviously right if 'empirical regularities' are statistical measures of association, like regression coefficients. The thesis is less obvious if 'empirical regularities' are defined more broadly, so as to include (for example) causal relationships." We agree here as well, although we were trying in the cited sentences precisely to distinguish empirical regularities in the sense of mere associations from causal relationships. We would not claim a generalized superiority of large-*N* methods for identifying causal relationships. Indeed, the main idea of the multi-method approach we are endorsing is to use case-specific evidence systematically to assess whether causal interpretations of the mere associations seen in a regression analysis are justified.

We do not therefore see how Freedman attributes to us the notions of the "superiority" of large-*N* methods or that "qualitative methods are useful only as checks on quantitative results." These claims may suggest incorrectly that we think causal relationships are easily read out of large-*N* statistical studies in social science. They also misread our view of the contributions made by small-*N* methods in the overall research process. In practice, as we noted above, there is a constant back-and-forth between data and theory in social science research, with case study evidence entering in more than one way. Knowledge of particular cases often helps to suggest causal mechanisms that may or may not be common and relevant in a larger sample of cases, and so may motivate and guide construction of a large-*N* study. A large-*N* study may in turn reveal new and different-from-expected patterns that stimulate new (or revised) theorizing about causal relationships, which may then be assessed by a return to case studies (chosen at random?). Those case studies may suggest new causal relationships that can subsequently be put to test with a newly constructed dataset. So it often happens in political science that case studies come into the scientific process at an early stage, motivating the research and the source of early conjectures, and then again at a later stage, after the regressions have been run.

Researchers in comparative politics invariably go back-and-forth between theory and data, and quite often they go back and forth between cases and broad patterns. Our modest proposal is an attempt to make progress on the question of by what principles to choose the cases in the context of the back and forth.

Notes

[1] See for example Evan Lieberman's critique of our proposed method, "Nested Analysis as a Mixed-Method Strategy for Comparative Research." *American Political Science Review* 99:3 (August 2005), 435–52.

[2] James D. Fearon and David D. Laitin, "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97:1 (February 2003), 75–90.

Techniques for Case Selection: A Response to David Freedman

John Gerring
Boston University
jgerring@bu.edu

Recognition of the problem posed by case selection in case study research stretches back, arguably, to the very beginnings of the genre, e.g., to early work by Frederic Le Play (1806–1882) and Florian Znaniecki (1882–1958). Harry Eckstein's (1975) classic study, a point of departure for political scientists today, appeared over three decades ago. Clearly, the field has been struggling with this issue for some time.[1]

The objective of my chapter for the *Oxford Handbook* was to summarize extant approaches to case selection, to add some new techniques to this battery of strategies (or at least provide a moniker for and a formal treatment of techniques that are already widely practiced), and to show how quantitative techniques might be brought to bear on these matters. ("Statistical" and "quantitative" will be employed synonymously in this discussion.) With respect to the latter, my argument was that most case-selection techniques could be practiced either qualitatively or quantitatively, given the right circumstances. (One technique, the "crucial case," can be practiced only qualitatively.) Nine purposive case-selection procedures were reviewed, each associated with a distinct case-study type: *typical*, *diverse*, *extreme*, *deviant*, *influential*, *crucial*, *pathway*, *most-similar*, and *most-different*.

I am very grateful to David Freedman—whose work has served as a touchstone for many of us—for offering his thoughts on this long-running issue and for offering me an opportunity to clarify and refine my initial statement. I believe that a good deal of common ground can be located here.

I should say, at the outset, that my chapter was based on prior work which presented these issues at greater length (Gerring 2007a, 2007b; Seawright and Gerring 2008). The chapter in the *Oxford Handbook* was not intended to provide a comprehensive treatment of this very large subject. Indeed, some considerations important to case selection were omitted entirely from the chapter, for lack of space (e.g., Gerring and McDermott 2007). Quite possibly, some of the points of apparent disagreement are a product of the condensed format demanded by the *Handbook*. Others are doubtless due to my own failure of communication. There may also be one or two points of genuine disagreement. In any case, I am anxious to explore what these might be, in the hopes that by doing so we can move the field forward.

Let me begin with what I take to be a point of agreement. Large-sample work aims for some version of random sampling. This is the textbook method of case selection. Freedman corrects my overly optimistic assessment, pointing out that large-*N* research often does not achieve this aim. Although I have not studied the matter, I would assume that random sampling is often achieved when the units of analysis are individuals and when these responses are drawn from survey research (e.g., in public opinion studies). I would assume that random samples are generally not achieved when the individuals composing a sample are being subjected to an experimental protocol or when the units of theoretical interest are larger entities such as countries.

By contrast, research based on very small samples (*N*=1 or several) cannot employ this time-honored technique of case selection for two critical reasons. First, there is a high likelihood that the (randomly) chosen cases will be wildly unrepresentative of the population (note that where *N*=5, sampling variance is much higher than where *N*=50 or 100). Second, a randomly chosen sample is unlikely to provide adequate leverage for the research question under investigation (note that where *N* is large, the resulting sample is likely to contain variation on theoretically relevant dimensions, but this is not the case where *N* is very small). This was not an issue addressed in my chapter for the *Handbook*, but it is an important assumption underlying the chapter (Gerring 2007a; Seawright and Gerring 2008), and one that may be worth expatiating upon in the present context.

To clarify, where chosen samples are medium to large, as in Fearon and Laitin's sample of 25 countries, it is reasonable to employ random sampling or stratified random sampling, as they do. (Note, however, that because one of Fearon/Laitin's sampling criteria is the existence of a civil war—the dependent variable of interest—it does not fit the usual understanding of stratified random sampling. Still, there are often good reasons for selecting on the dependent variable, in the tradition of case-control studies.)

Naturally, the existence of 25 country-cases imposes a considerable burden on the analysts, necessitating long, in-depth studies for each case (which the authors are in the process of completing). Typically, when the number of chosen cases is this extensive, the amount of detail and original research devoted to each case is limited. Depth and breadth tend to vary inversely. This recalls a point of definition. If a "case study" refers to an in-depth analysis of a single case (with the objective of saying something about a broader population of cases), then the case study format becomes more diffuse as *N* increases. However, there is no hard and fast boundary between a case study and a cross-case study. One flows into the other. That is one of the many ambiguities of the term (see below).

Now, let us suppose counterfactually that Fearon and Laitin proposed to conduct a study of a single case (e.g., Algeria), or a very small sample composed of three or four cases, while maintaining their theoretical interest in generating insights about a global population of nation-states. Here, there are lots of reasons to be suspicious of random or stratified random sampling approaches to case-selection. This, of course, is not what the authors intend: the case of Algeria is offered as an example of a much larger sample of cases—25 in all. But, for heuristic purposes, let us discuss a few of these potential difficulties.

Begin by stratifying the total population of potential country cases ($N \approx 180$) across three dimensions that are deemed to be theoretically relevant, e.g., *socioeconomic status* (rich/poor), *civil war* (yes/no), and various *regions* including Africa, the Americas, Eurasia, and the Pacific. The intersection of these dimensions sub-divides the universe of country-cases into sixteen sub-strata. It should be obvious that a fairly large sample will be necessary in order to represent, in a plausible fashion, the full range of cases in the population—at a *minimum*, sixteen (one from each sub-stratum). However, this presumes a very high degree of homogeneity within each sub-stratum such that all cases within a sub-stratum yield virtually equivalent results (with respect to whatever causal proposition is being explored). In this setting, which we must imagine is extremely rare in the social sciences, it hardly matters how one chooses cases—random or purposive. Every selection procedure building on the aforesaid stratification will achieve the same results. If, on the other hand, there is some theoretically relevant variation across cases within a substratum, then a much higher number of cases will be necessary to produce a sample that can claim to be representative (probabilistically) of the broader population of nation-states. Clearly, we are in large-*N* territory.

Now, if the researcher sacrifices the goal of representativeness she is free to choose among particular substrata, ignoring others. (Note that this is not what survey researchers mean by *over-sampling* since it is no longer possible to reconstruct, by weighting, a truly representative sample.) Suppose she decides, on some basis, that she will choose cases only from the substratum of cases that are poor, civil-war prone, and Asian. In principle, this could be handled by random draw, removing the researcher from the decision and any potential bias that might result. Yet, there are several reasons to think twice about this procedure.

First, there is the problem of sampling variance within each substratum—presumably less than one would find in the population at large (since the substratum is chosen with an eye to creating greater homogeneity), but still perhaps enough to give pause. China and Burma, both of whom qualify as members of the identified substratum, are very different places. Thus, representativeness of the substratum (let alone of the larger population) is unlikely to be achieved in an *N*=1 sample chosen in this manner. Arguably, representativeness (in this limited sense) is more likely to be achieved when the researcher chooses the case purposefully than when she chooses randomly, for she can incorporate background knowledge of the cases (factors not included in the stratification) and exercise judgment about which case lies nearest to the mean along relevant dimensions.

The second obstacle concerns the leverage (for causal inference) that a given case is likely to provide. Recall that a good sample is not only representative but also insightful. This can mean lots of different things, but to simplify let's say that some cases look more like natural experiments than others. Suppose that the civil war in China is accompanied by all sorts of theoretically extraneous factors (connected with the communist insurgency or foreign intervention) that have little to do with the theoretical hypotheses at hand. Burma, by contrast, has few of these potential confounders. Under the circumstance, Burma is clearly a better choice (all other things being equal). This, too, validates a purposive (non-random) approach to case-selection.

Third, there are practical factors like the language capacities of the researcher, the availability of documents and other sources, and the ability to gain entry to a country for field research. Again, one may find strong reasons to favor Burma over China (or China over Burma)—reasons that would be washed away if the case were chosen randomly.

To be sure, there is no limit, in principle, to the number of features that can be included in a stratification procedure. Every "purposive" feature mentioned above can be incorporated into the formal stratification, removing it from possible investigator bias. This also enhances the clarity and explicitness of the case-selection procedure. Yet, one wonders whether all of these factors can really be measured, ex ante, across the entire population. And if they could, the resulting stratification would include so many dimensions that—with a fixed and moderate-sized population—each sub-stratum would be miniscule, allowing little scope for random draws.

In short, it is difficult to justify selecting a sample of *one or several* cases in a purely random or stratified random fashion. One can see why case study researchers want to attach proper names to their potential cases before finalizing their selection. Case knowledge is often revelatory.

Thus, random sampling within pre-identified sub-strata is sometimes viable (it is indeed suggested in my chapter as part of the "diverse-case" procedure). Where it is, it offers a potential solution to problems of researcher bias that might otherwise influence the case-selection procedure, a concern raised by Fearon and Laitin (see also Gerring 2007a: 145). Where it is not, the researcher must fall back on purposive (non-random) procedures. Either way, the most critical question is probably this: *on what principles* should the case-selection criteria—stratified or not—occur? How, in the context of a stratification, should the strata and sub-strata be selected? This is the main topic addressed in my chapter. (Note that the selection of strata, and the choice of sub-strata to over-sample, presupposes a selection principle for the strata; this I consider to be the purposive, or intentional, element.)

I shall now proceed to address some of the criticisms raised by Freedman's commentary. Freedman's summary comment on my chapter is that "some of the methods that Gerring proposes are ill-suited to qualitative research." I gather that his hesitancy about applying statistical methods to case study research stems, in part, from his view of the inherent limits of quantitative techniques when faced with nonexperimental data. I share this skepticism. Even so, it seems clear that case study research is often compelled to labor with nonexperimental data. This being the case, the relevant question is whether problems of causal attribution endemic to observational studies are made any worse by the application of statistical models to aid in the process of case selection. Here, I am agnostic.

Before continuing, I should clarify that I am emphatically not proposing that statistical analysis be applied to very small samples. This, combined with the ubiquitous assignment problem posed by observational data, is a recipe for disaster. What I am proposing is that quantitative techniques might find employment, at least on certain occasions, in the selection of cases for in-depth research focused on one or several cases. (I would also argue that there is no reason to preclude the statistical analysis of large-sample *within-case* evidence—though this issue was not raised in the chapter.)

I am assuming, of course, that the population of interest is large enough to justify the employment of regression, matching, and other suggested techniques. I write: "In certain circumstances, the case-selection procedure may be structured by a quantitative analysis of the larger population" (646). Again, "Sometimes, these principles can be applied in a quantitative framework and sometimes they are limited to a qualitative framework. In either case, the logic of case selection remains quite similar, whether practiced in small-*N* or large-*N* contexts" (ibid.).

Note that all of these techniques are introduced as techniques of case *selection* (as signaled in the title of the chapter), not case *analysis*. The idea is that, in some instances, the relevant requirements for a statistical analysis may be met, and in these instances it makes sense to formalize the case-selection strategies that would otherwise be carried out in a qualitative manner. Thus, with respect to the "deviant-case" strategy, rather than choosing a case that appears unusual with respect to some informal model of causal relationships, one might actually test these assumptions formally in a statistical model, choosing a case (or cases) with a high residual(s).

Occasionally, the goal of a case study is to confirm/disconfirm a statistical model. Here, an appropriate strategy of case selection might be an "influential-case" analysis—where relevant cases are identified by examining hat matrix and Cook's distance statistics for individual cases (developed initially in Seawright and Gerring 2008). Similarly, other quantitative techniques such as cross-tabulations may be helpful in the context of a "diverse-case" analysis. To say that a case-selection procedure is purposive does not, therefore, imply that it must be small-*N* or qualitative.

Of course, it is always an open question how much confidence one ought to place in large-*N* statistical models. Yet, an article on case selection in case study analysis did not seem the appropriate venue to expatiate on a point that so many others (notably David Freedman) have persuasively argued. I stated explicitly that "relevant data must be available for [a] population...on key variables, and the researcher must feel reasonably confident in the accuracy and conceptual validity of these variables. [Further,] all the standard assumptions of statistical research (e.g., identification, specification, robustness) must be carefully considered, and wherever possible, tested." I then warned against "the unreflective use of statistical techniques." Doubtless, some people will continue to use statistical techniques even where they are not warranted. But this potentiality should not prevent us from discussing instances in which the employment of statistical techniques *is* warranted.

At this point, it may be helpful to formally define the key term, "case study." Sometimes, the case study method is equated with qualitative methods. My understanding of the concept is different (Gerring 2007a, 2007b). A case study, as I see it, is most usefully defined as the intensive study of a single case, or several cases, where the purpose is to shed light on a broader population of cases. It should not be equated with qualitative methods, since we already have a term for this concept. To be sure, any *cross-case* analysis—i.e., in a Millean-style comparative study—would have to be qualitative, for the sample is extremely small (by definition). But case selection and within-case analysis may be qualitative and/or quantitative. (The resulting two-by-two matrix produces four cells, and all are occupied.)

I turn now to the comparison of case study techniques and experimental techniques—a minor point of the chapter but a central objective in previous work (Gerring 2007a: chapter 6; Gerring and McDermott 2007). I concur with Freedman that "The logic of randomized controlled experiments is... central, even for observational research." Thus, in thecourse of discussion of Mill's most-similar method (aka the "method of difference") I make several analogies to experimental methods.

For example, in order to reduce background noise, experimentalists often stratify a sample into relatively homogeneous ("most-similar") sub-strata prior to treatment. The treatment is then randomized *within* each sub-strata (or block). If the sub-strata are very small, e.g., blocks of two, the procedure may be described as an iterated most-similar comparison with randomized treatment. In this respect, I think it is fair to say that both Millean and experimental studies often attempt to identify samples (or sub-samples) that are as internally homogeneous as possible. I regret that I did not offer some explication of this point—which is not self-evident—in the chapter.

A great deal of ground has been covered in this all-too-brief review. The common thread running through the narrative is that our understanding of case study research is enhanced when we make comparisons and contrasts with other sorts of research—e.g., with large-N cross-case analysis with observational data or with experimental research. Sometimes, techniques not usually associated with the case study can be utilized in the selection of a few cases for intensive analysis. Sometimes, on the other hand, these techniques are inappropriate. I trust that we are moving closer to a consensus on these matters.

In any case, there are many hopeful signs that the gulf that has traditionally separated case study and non-case study methods is narrowing. This colloquy is an excellent example of that propitious development. Let me close by thanking David Freedman, James Fearon, and David Laitin for facilitating this exchange. I only wish I had the benefit of their counsel when crafting earlier projects.

Note

[1] All direct quotations are from my chapter in the Oxford volume, unless otherwise noted.

References

Eckstein, Harry. 1975. "Case Studies and Theory in Political Science." In *Handbook of Political Science, vol. 7. Political Science: Scope and Theory.* Fred I. Greenstein and Nelson W. Polsby, eds. (Reading, MA: Addison-Wesley), 94–137.

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

Gerring, John. 2007b. "The Case Study: What it is and What it Does." In *Oxford Handbook of Comparative Politics.* Carles Boix and Susan Stokes, eds. (New York: Oxford University Press), 90–122.

Gerring, John and Rose McDermott. 2007. "An Experimental Template for Case-Study Research." *American Journal of Political Science* 51:3 (July), 688–701.

Seawright, Jason and John Gerring. 2008. "Case-Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options." *Political Research Quarterly* 61:2 (June), 294–308.

Choosing Cases for Case Studies: A Qualitative Logic

Gary Goertz
University of Arizona
ggoertz@u.arizona.edu

David Freedman in his essay has called into question some of the advice given by Fearon-Laitin and Gerring in their chapters in the *Oxford Handbook of Political Methodology*. I would like to second those criticisms and extend them in various ways. In particular, Freedman does not address in much detail the regression or logit model which underlies, explicitly or implicitly, both these chapters—and more generally Gerring's book on case study methodology (2007). This "regression approach" to case studies (to give it a name) informs much discussion about case studies and qualitative methods, going back to King, Keohane, and Verba (1994) and more recent works such as Lieberman (2005). In these few pages I can but sketch a rationale for choosing cases following a different logic of research. In the first part of the essay I address the choice of case studies from a qualitative logic of research. In the second part, I briefly describe a "descriptive–causal" approach to case study selection which is different from the regression logic of Fearon, Laitin, and Gerring.

Case Studies Selection and Research Agendas

Many, if not most, research projects start with a decision to explain some phenomenon, $Y = 1$ for short. This can be war, democracy, voting, or whatever. As such one will naturally select cases of $Y = 1$ for intensive scrutiny, because explaining $Y = 1$ is exactly the overall goal of the research project.

With so much focus in methodology classes on the problems with "selecting on the dependent variable" it is easy to lose sight of the key methodological principle that one should select some individual cases of $Y = 1$ for intense examination. For example, when a new disease occurs medical researchers first focus on people with the disease in order to understand it. When AIDS was first discovered there was intense concentration on those who had the disease. Ragin is one of the few who has constantly stressed the importance of focusing on the $Y = 1$ cases (1987, 2000, 2008). A first principle of case selection of case studies is:

> Principle 1: One should choose, *diversely*, among the *good* instances of $Y = 1$ for case studies.

According to this principle you want to choose a *diverse*, not random, set of cases because you do not want to miss an important causal path to *Y*. Fearon and Laitin stress the value of random selection. Both they and Freedman recognize the pitfalls of "cherry-picking" and that random selection can help avoid this problem. At the same time Fearon-Laitin want to use random selection to choose "representative" or "typical" cases. Because they start with a regression model, they work from the presumption that there is a β_i that is the typical causal effect of X_i. A qualitative research logic is much more likely to start with an INUS model of causation where there are multiple paths to Y, such as $Y = x * A * B + X * B * C$ (lower case means absence of factor). If you start with an INUS view of the world you do not necessarily believe there is one representative causal effect of X, since, depending on the path, the presence of X or its absence x may be a cause of Y. So one looks for diverse cases in order to not miss causal paths. In short, whenever Gerring, Fearon, and Laitin use the idea "random" I would suggest replacing it with "diverse."[1]

Principle 1 also stresses that one should choose among "good," i.e., unambiguous, cases of $Y = 1$. In the case of civil wars this means one should not choose randomly among all cases of $Y = 1$. There is certainly a good percentage of cases of civil war that are marginal, or "gray" cases of civil war. Definitions of phenomena draw boundaries; almost inevitably there are cases near those boundaries that may not fit the concept very well. If one is beginning to study AIDS, one would not choose cases that may or may not be AIDS or cases that seem atypical of AIDS.

After one has expanded some significant shoe leather understanding deeply some $Y = 1$ cases one might begin to have some ideas about the causes of *Y*. The concerns with cherry-picking are real and I think they are most often tied to selecting cases based on *X*, particularly $X = 1$. It is natural to focus on cases where the author's theory works. Hence from a qualitative point of view the risks of bias are in some sense larger when selecting on *X* than selecting on *Y*.

These potential causes, $X = 1$, then lead to choosing cases that allow one to see how plausible (to use Eckstein's terminology) *X* is as a potential cause of *Y*. Thus we will choose cases where *X* is present to see if we can figure out the causal mechanism linking *X* to *Y*. Typically, we will focus our attention on the (1,1) cases, i.e., $Y = 1$ and $X = 1$ cases, because these allow for causal process tracing (George and Bennett 2005). Here we have a second principle for choosing instances for case studies:

> Principle 2: Select, *diversely*, cases of $X = 1$.

For example, researchers trying to explain lung disease noticed that this seemed to be common among smokers. This naturally led then to efforts such as experiments on rats where they were forced to smoke a lot. The rationale for diversity in Principle 2 is the same as in Principle 1: one wants to detect multiple causal paths to $Y = 1$.

It cannot be stressed enough that part of case study methodology involves counterfactual analysis (Tetlock and Belkin 1996; see Levy 2008 for an excellent discussion). Counterfactual methodology turns $X = 1$ into $X = 0$. This is in part why we can focus on the $X = 1$ cases, because we later turn them into $X = 0$ via counterfactual analysis.

For some *X*s it may not be possible, or may be very difficult, to do a good counterfactual analysis. If we are exploring the impact of wealth or GDP/capita on civil war, it would be hard to say what is the likelihood of civil war if, for example, Sierra Leone were a wealthy, developed country.[2] The fact that this is a difficult counterfactual (see Ragin 2008 for this concept) means that selecting a case study of $X = 0$ based on wealth will be problematic as well. This is related to the well-known "minimum rewrite rule" for counterfactuals which recommends only modest changes in X for counterfactuals. This is so important that I think it needs to be elevated to a principle:

> Principle 3: If the counterfactual $X = 0$ is problematic in individual cases of $Y = 1$, $X = 1$, then choosing actual instances of $X = 0$ for case study analysis is problematic as well.

Principles 1 and 2 stress that we choose case studies based on sampling on $Y = 1$ or $X = 1$. This leads to a principle which is a corollary of the first two:

> Principle 4: Cases of $Y = 0$ and $X = 0$ are often not very useful for intensive case study examination.

Jim Mahoney and I have made the argument elsewhere (2004) that the (0,0) cases are problematic for qualitative researchers. Often there is a large number of these cases. Random selection among this (often very large) number is unlikely to produce cases that will be useful for a case study. These cases might be very important in a large-*N* statistical analysis, but much less useful for a case study. For example, it is crucial to the large-*N* study of the linkage between smoking and lung cancer to include (0,0) cases.

I suggest that these principles for choosing instances for case study research are what researchers often naturally do, and at the same time good practice. Underlying the belief that multi-method research is good lies the notion that different methods give us different information and different views of the phenomenon. If one bases case study research on the regression model we lose the distinctive, and I think complementary, advantages of case study methods vis-à-vis regression methods.

The Descriptive–Causal Approach to Case Study Selection

Fearon-Laitin and Gerring explicitly link the selection of case studies to regression or logit models. In this section I propose that we can and should use our knowledge of the cases and of patterns in the cases to restrict our attention to quite limited regions of the data (or absence of data) determined by our empirical and theoretical interests. This I call, for reasons that should become clear, the "descriptive–causal" approach to case study selection.

For purposes of contrast I will use Gerring's (2007) example of the relationship between wealth, aka GDP/capita, and democracy, which is a central example in his core Chapter 5, and more generally in the literature on the social, economic, and political requisites and correlates of democracy.

The descriptive–causal approach takes advantage of the fact that we have often accumulated some basic descriptive knowledge about the cases and about some general patterns in the data. Selection of case studies then relies on case knowledge. I use the term descriptive–causal to apply to descriptive statistics that have causal implications. For example, the democratic peace can be formulated as "Democracies do not fight wars with each other." This is *descriptive* in the sense that it gives the frequency of occurrence of a phenomenon. It is *causal* in the sense that it makes a link between a potential cause, democracy, and a potential effect, war.

Lipset's American Sociological Association presidential address (1994) is in fact a review of the literature that he was so instrumental in launching several decades earlier. This review is full of descriptive–causal patterns. For example, "every country with a population of at least 1 million that has emerged from colonial rule and has had a continuous democratic experience is a former British colony" (Lipset 1994: 5). Dahl (1971) says that "all highest-level [developed] countries are polyarchies" (cited in Diamond 1992: 97). Such examples can easily be multiplied.

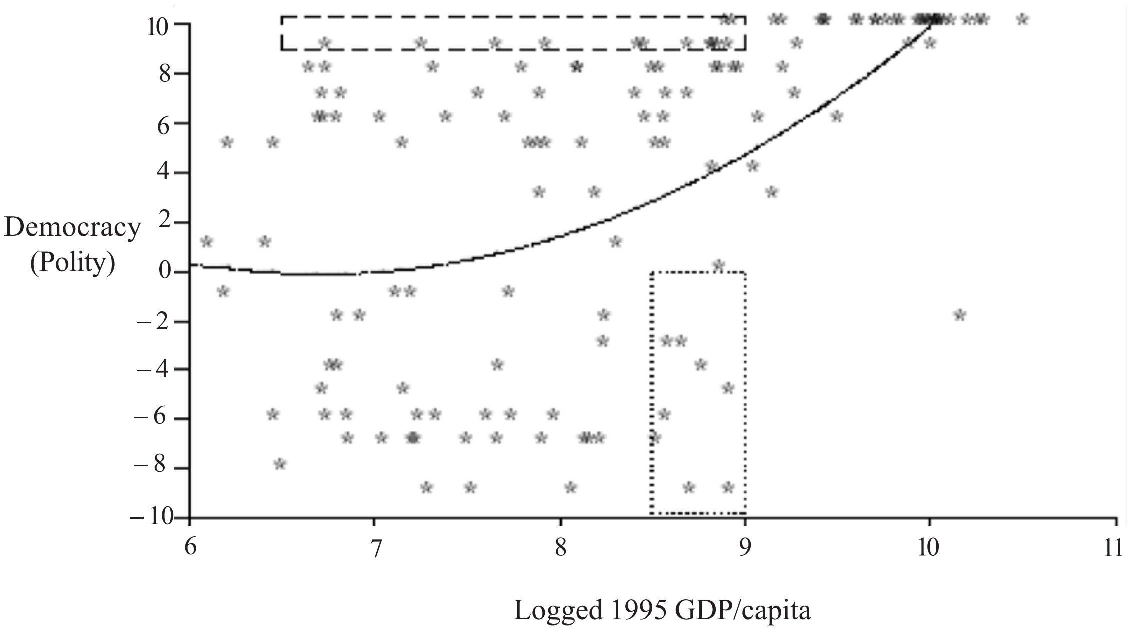
The key thing about these descriptive–causal statistics is that they point to regions of the data where we might want to focus our attention in the form of an intensive case study. Figure 1 (adapted from Gerring 2007) illustrates the descriptive–causal approach and how it differs from a regression one. As a point of reference I have drawn a regression curve (OLS) through these data (not in the original Gerring figure). Take for example Przeworski et al.'s well-known analysis (2000) of the wealth–democracy relationship which depends on, roughly, noting that (1) most highly developed countries are democracies and (2) such countries rarely fall back into authoritarian systems. If these descriptive–causal patterns are correct then over time we should see very few cases in the lower-right hand corner of figure 1. Thus the interesting thing is not the regression curve through the middle of the data, but rather the lower-right-hand region. Given our interest in this region we might explore the one outlier in the figure (Singapore). It seems that this pattern really starts at a logged GDP/capita of about 9. For countries that are poorer the pattern does not seem to hold. Hence we might choose a case study in the region delimited by the dotted lines.

Another well-known descriptive–causal pattern in the literature is that it seems very difficult for poor countries to become high-quality stable democracies. Much of the early work focused on this particular pattern. This pattern directs our attention to the upper left and center part of figure 1, where we should see poor democracies if they exist. If we consider 10 on the polity scale to be "high-quality democracy" then we notice that this particular causal effect does not kick in until a logged GDP/capita of about 8.5. Here too we want to choose some case studies in the zone bounded by the dashed lines to explore more closely this potential causal relationship.

As figure 1 makes clear, if we are interested in these descriptive–causal patterns, working from a regression or probit line is of little use. If we randomly choose cases or choose cases based on the regression, the likelihood that they would be informative for our causal purposes is very low.

This essay does not say that the regression approach to case studies is not useful. Rather it argues that there are alternative ways to think about selecting case studies. In a different empirical or theoretical setting the regression approach may just be what the doctor ordered. At the same time much of

Figure 1: Descriptive-Causal Patterns: Wealth and Democracy



Source: Adapted from Gerring 2007: 96.

empirical, case-oriented knowledge that is so important to qualitative scholars is expressed in descriptive–causal claims like those of Lipset, Diamond, and Dahl. It is not surprising that these claims lie at the core of Ragin's Boolean and fuzzy set methodologies. They are a means of formalization of many descriptive–causal claims.

Using Case Studies to Evaluate Scope Conditions

With the notable exception of Dul and Hak (2008), the literature on case studies has not seen them as useful in exploring scope conditions. As Freedman notes in his essay, most statistical data analyses are based on samples or on populations of convenience. Mahoney and I have argued (2006) that qualitative scholars are often more concerned with scope conditions because they are much more concerned with cases that do not fit the theory.

Figure 1 illustrates nicely how one can use case studies to evaluate scope conditions. In figure 1 Gerring and Seawright have left out a handful of cases, which in fact lie in the lower-right corner, and hence are of particular interest to the Przeworski et al. pattern, and can be potentially seen as problematic.

If we look at these countries a new pattern becomes very clear: they are all wealthy countries because of large oil reserves, such as Saudi Arabia. The comparative politics literature (e.g., Ulfelder 2007) has already noticed that these countries suffer from the "natural resource curse." These authoritarian governments can remain in power without taxing their subjects. Hence it might be very reasonable for Gerring to exclude these cases from consideration because they do not fit the causal mechanisms that we find in the rest of the world.

While these cases are outliers in the regression model, a random selection of outliers would never detect this pattern (or more precisely, would detect it with extremely low probability). There are lots of outliers from the regression curve. It is because we have particular theoretical and empirical interest in the lower right-hand corner that these outliers become important to us and an analysis of the individual cases can lead to scope restrictions.

The key point here is that cases are not representative of some given population but rather that the population is constructed via of knowledge of the cases, cases studies, and our causal analyses of them.

Conclusion

In this essay I have argued that we should, and usually do, have clear purposes for selecting instances for intensive case study. In general, we normally want to focus on the cases where $X = 1$ and $Y = 1$; we probably want to avoid cases of (0,0) unless we have a clear substantive rationale.

Qualitative scholars also select cases based on their knowledge of the cases and patterns in the cases. This descriptive–causal knowledge can point to particular regions of the data for selection of case studies.

One likewise uses case studies to construct and delimit populations. Instead of being given or taken by convenience, qualitative scholars construct populations using their knowledge of the cases.

I am a firm believer in the toolbox metaphor for methods. I see the descriptive–causal approach to case studies as another useful tool. Depending on the empirical and theoretical goals it might be more appropriate than the regression approach and in some circumstances it might be less.

Notes

[1] Of course one needs to define the dimensions of diversity; one option is region as in the Fearon-Laitin chapter.

[2] This leads to King and Zeng's (2007) counterfactual critique of this literature.

References

Diamond, Larry. 1992. "Economic Development and Democracy Reconsidered." In Gary Marks and Larry Diamond, eds. *Reexamining Democracy: Essays in Honor of Seymour Martin Lipset*. (Newbury Park, CA: Sage Publications).

Dul, Jan and Tony Hak. 2008. *Case Study Methodology in Business Research*. Amsterdam: Elsevier.

Fearon, James and David Laitin. 2008. "Integrating Qualitative and Quantitative Methods." In Janet Box-Steffensmier, Henry Brady, and David Collier, eds. *The Oxford Handbook of Political Methodology.* Oxford: Oxford University Press.

George, Alexander and Andrew Bennett. 2005. *Case Studies and Theory Development.* Cambridge: MIT Press.

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

Gerring, John. 2008. "Case Selection for Case Study Analysis: Qualitative and Quantitative Techniques." In Janet Box-Steffensmier, Henry Brady, and David Collier, eds. *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press.

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

King, Gary and Langche Zeng. 2007. "When Can History be our Guide? The Pitfalls of Counterfactual Inference." *International Studies Quarterly* 51:183–210.

Levy, Jack S. 2008. "Counterfactuals and Case Studies." In Janet Box-Steffensmier, Henry Brady, and David Collier, eds. *The Oxford Handbook of Political Methodology.* Oxford: Oxford University Press.

Lieberman, Evan. 2005. "Nested Analysis as a Mixed-Method Strategy for Comparative Research." *American Political Science Review* 99: 435–52.

Lipset, Seymour Martin. 1994. "Social Requisites of Democracy Revisited." *American Sociological Review* 59: 1–22.

Mahoney, James and Gary Goertz. 2004. "The Possibility Principle: Choosing Negative Cases in Comparative Research." *American Political Science Review* 98: 653–69.

Mahoney, James and Gary Goertz. 2006. "A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research." *Political Analysis* 14: 227–49.

Przeworski, Adam, et al. 2000. *Democracy and Development: Political Institutions and Well-being in the World, 1950–1990*. Cambridge: Cambridge University Press.

Ragin, Charles. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press.

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

Ragin, Charles. 2008. *Redesigning Social Inquiry: Fuzzy Sets and Beyond*. Chicago: University of Chicago Press.

Tetlock, Philip E. and Aaron Belkin, eds. 1996. *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*. Princeton: Princeton University Press.

Ulfelder, Jay. 2007. "Natural Resource Wealth and the Survival of Autocracy." *Comparative Political Studies* 40: 995–1018.

Rejoinder

David A. Freedman
University of California, Berkeley

I would like to begin by thanking James Fearon, David Laitin, John Gerring, and Gary Goertz for their comments, which help to clarify the issues on the table. Whatever differences remain, we can all agree that David Collier did a great job in organizing this discussion.

Fearon and Laitin

As I read the paper, Fearon and Laitin (2008) made a clear statement that case studies were a poor way to establish empirical regularities; large-*N* methods were to be preferred. The claim was justified using a narrow definition of "empirical regularities," which excluded pretty much everything except summary statistics (means, standard deviations, regression coefficients…). The paper segued to an implication that there was one natural way of integrating qualitative and quantitative research methods, with the former as ancillaries to the latter.

However, the story has a happy ending. Fearon and Laitin explain that my reading of the paper was not the intended reading. As it turns out, we agree on the following points. (i) There many ways to do good science. (ii) In particular, neither cluster of methods has a general advantage over the other. (iii) Therefore, there are many fruitful ways for qualitative and quantitative researchers to interact. (iv) When it comes to making causal inferences, case studies often have considerable power—although, to be sure, for statistical inference, bigger *N* is usually better.[1]

There was also some back-and-forth about sampling. On this topic too, there is now reasonable agreement. As with other choices to be made in research, much depends on context and on background knowledge. Fearon and Laitin raise a new point, contrasting the messy realities of research design with "the pristine hypothesis-testing scenario assumed in statistics textbooks." This shaft is well-aimed, although the environment is target-rich: there are econometrics textbooks, psychometrics textbooks....

What can be said about the substantive research in Fearon and Laitin (2008)? I believe that Fearon and Laitin used their logit model descriptively, to find patterns in the data that suggest one causal theory or contradict another theory. They did not rely on the model to make causal inferences. Instead, they used the case studies to do the heavy lifting. They chose cases using an elegant and impartial technique. These are useful ideas, which should find many applications.

Gerring

I think there is agreement on the following points:

(i) Large-*N* researchers should use random samples, and often they do. Often, however, there is a divergence between the ideal and the real.

(ii) Modeling and matching are large-*N* techniques that make stringent assumptions about data-generating mechanisms. These techniques can help us choose a small number of cases for in-depth study when (a) we are choosing those cases from a big, well-defined population, (b) there are complete data for all the cases in the population, and (c) the assumptions behind the modeling and matching are viewed as reasonable for the population. These conditions are clear in Gerring's response, as they are in Seawright and Gerring (2008). They will seldom (if ever?) be satisfied in small-*N* research: even the first two might be problematic.[2]

(iii) If the experiment was done well, few assumptions are needed to analyze the data. Blocking subjects to achieve greater homogeneity may be a good idea, but that is something you do before randomization, not when you are analyzing the data.

(iv) Causal inferences are frequently based on observational studies rather than experiments, with elaborate modeling and matching to control for confounders. Assumptions play a large role in these proceedings, and the opportunity for error is correspondingly large. This makes a striking contrast with experiments. I would add, however, that in many cases our causal knowledge derives from well-designed observational studies, where the data do not require complex statistical analysis (Freedman 2005: Chapter 1); I think Gerring will agree.

Although this topic is only tangential to Gerring's work, more should be said about non-response. Non-response rates are high for many surveys, and the level is generally rising. Even if we start from a probability sample, the actual respondents going into the analysis can be a lot like a convenience sample, because non-respondents and respondents can be very different (Freedman et al. 2007: Chapter 19).

To illustrate the magnitude of the problem, I will use three of the papers reprinted in Freedman (2005).[3] These papers start with large probability samples. However, 50% to 75% of the data are missing, because subjects refused to cooperate with the survey, or declined to provide some of the data that were needed. This is especially poignant because the papers are grappling with the endogeneity of selection into treatments of one kind or another. However, endogeneity of selection into the sample is politely ignored. As Gerring (2008: 678) says, "Not all twists and turns on the meandering trail of truth can be anticipated."

Goertz

I disagree with half of what Goertz says, but will only respond to three things: (i) the interpretation of Fearon-Laitin and Gerring, (ii) the philosophy of case selection, and (iii) the advice to ignore a cell in the 2 x 2 table.

(i) According to Goertz, "Fearon-Laitin and Gerring explicitly link the selection of case studies to regression or logit models." I cannot see any explicit statement either in Fearon and Laitin (2008) or in Gerring (2008) to that effect. Of course, Fearon and Laitin are selecting cases in the context of a logit model. However, these scholars do not rely on the assumptions behind the model (Freedman 2005: Chapter 6) when selecting cases. Indeed, the principal recommendation on case selection is to use stratified random samples. This has little to do with models.

What about Gerring? To be sure, a few of his techniques for small-*N* case selection are linked to regression models. In my view, previously noted, these suggestions will rarely be helpful. By contrast, most of his discussion—for instance, of typical and diverse cases—is blessedly model-free and generally useful. (Is this causation or just association?)

(ii) Goertz says, "The key point here is that cases are not representative of some given population but rather that the population is constructed via knowledge of the cases, case studies, and our causal analyses of them." The statement comes perilously close to a recommendation that we should start with a theory, choose cases in conformity with that theory, and then conclude that the evidence supports the theory. No one is immune from this tendency, but it is a habit to be discouraged rather than encouraged.

(iii) Goertz considers a binary causal variable *X*, where *X* = 1 means the causal factor is present, while *X* = 0 means the factor is absent. There is a binary response variable *Y*. The data can be presented in a 2 x 2 table:

	X	
Y	1	0
1	A	B
0	C	D

Goertz recommends in favor of looking at cell A when doing qualitative research; he recommends against looking at cell D. Curiously, he adds that for large-*N* research, cell D "*might* be very important [emphasis supplied]."

Playing favorites with cells is a risky business. At least in my experience, it is often hard to see where the cases go until you study them. Moreover, despite Goertz's reservations, all four cells are important to large-*N* researchers. Indeed, consider data like the following:

	X	
Y	1	0
1	10	20
0	20	??

If the number of cases in cell D is above 40, there is positive association; if the number is below 40, there is negative association. Since there is a fundamental difference between "*X* causes *Y*" and "*X* prevents *Y*," cell D matters in large-*N* research, along with the other cells.[4]

The boundary between large *N* and small is salient in the present context. For the moment, let us set the boundary at *N* = 17. If you accept Goetz's position, cell D can be relevant when *N* is above 17; cell D cannot be relevant when *N* is below 17. This is not a tenable position, and moving the boundary will not solve the problem.

If all four cells are relevant, close inspection of cases in all four cells has to be a good idea, at least under some circumstances. For example, a critic might assert that cases in cell D exhibit causal heterogeneity. The most straightforward way to rule that out is to look at cases in cell D.

The present exchange offers a real example. Fearon and

Laitin found cell D to be informative. This contradicts Goertz's position. In qualitative research, to be sure, examining only one cell in the table may sometimes be a good idea.[5] However, advice that cell D should generally be ignored is, well, advice that should be ignored.

Conclusion

I will draw an empirical conclusion[6] from the discussion: there are few recipes for good research. (Cooking and scholarship depend on somewhat different skill sets.) Nearly 50 years ago, my friend Larry Jackson defined the scientific method as "guess and verify." The only improvement I can make is to emphasize part of the recommendation: guess and *verify*.

Notes

[1] Fearon and Laitin point out that you do not want to increase N by stretching concepts. From my perspective, increasing sample size should reduce sampling error; but the effect on non-sampling error is unpredictable. Moreover, large-N research is often needed to demonstrate causation. Epidemiologic studies on the health effects of smoking, mentioned by Goertz, illustrate the point. For a brief review, see Freedman (1999).

[2] Fearon and Laitin are drawing a sample from a large, well-defined population to which a model has been fitted, so the first two conditions are satisfied. However, as indicated above, far from relying on the model, Fearon and Laitin are using the sample cases to test the model. Gerring indicates that the hat matrix and Cook's distance may be helpful in such contexts. I agree, but this favorable conjunction of circumstances is rare in qualitative research; in multi-method research, the story may be different.

[3] In the fourth paper, the unit of analysis is the state, so the issues are a little different.

[4] This discussion ignores sampling error, which is reasonable if N is large. The "odds ratio" is used to summarize the data, as is standard in epidemiology. Let a denote the number of elements in cell A, and so forth. If there are cases in all four cells, the odds ratio is $(a/c)/(b/d) = (a/b)/(c/d) = (ad)/(bc)$. The association is positive when the odds ratio is above 1.0; the association is negative when the odds ratio is below 1.0. You need all four numbers to compute the odds ratio.

If l denotes the odds ratio, the causal interpretation is this: setting X to 1 rather than 0 multiplies the odds that $Y = 1$, by the factor l. Equivalently, if $Y = 1$ rather than 0, the odds that $X = 1$ are multiplied by the factor l. For additional information, see Gordis (2008).

[5] Great work can be done with one cell, or even one case. Isn't de Tocqueville's *Democracy in America* a classic example of within-case analysis?

[6] Of course, like others discussed earlier, this conclusion depends in part on context and background knowledge.

References

Box-Steffensmeier, Janet M., Henry E. Brady, and David Collier. 2008. *The Oxford Handbook of Political Methodology*. Oxford University Press.

Fearon, James and David Laitin. 2008. "Integrating Qualitative and Quantitative Methods." In Janet Box-Steffensmier, Henry Brady, and David Collier, eds. *The Oxford Handbook of Political Methodology* (Oxford: Oxford University Press), 756–76.

Freedman, David A. 1999. "From Association to Causation: Some Remarks on the History of Statistics." *Statistical Science* 14: 243–58. Reprinted in *Journal de la Société Française de Statistique* 40 (1999), 5–32 and in John Panaretos, ed. 2003. *Stochastic Musings: Pespectives from the Pioneers of the Late 20th Century*. Lawrence Erlbaum Associates, 45–71.

Freedman, David A. 2005. *Statistical Models: Theory and Practice*. Cambridge University Press.

Freedman, David A., Robert Pisani, and Roger A. Purves. 2007. *Statistics*. 4th edition. New York: W. W. Norton & Company, Inc.

Gerring, John. 2008. "Case Selection for Case Study Analysis: Qualitative and Quantitative Techniques." In Janet Box-Steffensmier, Henry Brady, and David Collier, eds. *The Oxford Handbook of Political Methodology* (Oxford: Oxford University Press), 645–84.

Gordis, Leon. 2008. *Epidemiology*. 4th edition. Philadelphia: Elsevier-Saunders.

Seawright, Jason and John Gerring. 2008. "Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options." *Political Research Quarterly* 61: 294–308.