Spring 2011, Vol. 9, No. 1

Qualitative & Multi-Method Research

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

Contents

Letter from the Newsletter Editor

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

This is my last issue as editor of the section newsletter. Robert Adcock will take over the job for the fall 2011 issue. We have been working together on this issue and I am certain he will provide new vision and ideas for the job. Having a new editor every few years is good policy. Any given person has his or her own interests, networks, and expertise. Robert has wide-ranging interests that cover a large part of what the section does. Inevitably he will place his own stamp on the newsletter, but that too is a good outcome.

I would like to thank all of the people who have contributed to the newsletter over my period of editorship. Without exception contributors got their pieces done on time (in the fuzzy logic sense of on time) and I never had any problems. Also, huge thanks go to Joshua Yesnowitz, who has been the production person from the start. The quality look of the newsletter is his work. He too has been a pleasure to work with and has helped to get the newsletter out on time.

As per custom, this issue contains the panels (provisional, as always) for APSA 2011. Rose McDermott has done an excellent job, and as usual we have an interesting set of panels. Robert Adcock has gathered together some panels from related groups that are of particular interest in terms of qualitative methods and we list them, as well.

Let me also take this opportunity to draw your attention to two award announcements. The David Collier Mid-Career Achievement Award recognizes distinction in methodological publications, innovative application of qualitative and multi-method approaches in substantive research, and/or institutional contributions to this area of methodology. The inaugural presentation was made in 2010, to Lisa Wedeen, University of Chicago, and James Mahoney, Northwestern University. The second award is the newly established Qualitative Submissions to *APSR* Award, for the best qualitative manuscript submitted to the *American Political Science Review* in the calendar year. Details on eligibility and nominations procedures can be found on page 43 of this issue of the Newsletter.

A Sea Change in Political Methodology*

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

Henry E. Brady
University of California, Berkeley
hbrady@berkeley.edu

Jason Seawright
Northwestern University
j-seawright@northwestern.edu

Shifting debates on what constitutes "science" reveal competing claims about methodology.[1] Of course, in its origin the term "science" means "knowledge," and researchers obviously hold a wide spectrum of positions on how to produce viable knowledge. Within this spectrum, we compare two alternative meanings of science, advanced by scholars who seek to legitimate sharply contrasting views of qualitative methods. This comparison points to a sea change in political science methodology.[2]

One meaning of "science" is advanced in King, Keohane, and Verba's 1994 book, *Designing Social Inquiry* (hereafter KKV), which proposes a bold methodological agenda for scholars who work in the qualitative tradition. The book's subtitle directly summarizes the agenda: "*scientific* inference in qualitative research" (emphasis added). To its credit, the book is explicit in its definition of science. It draws on what we and many others have viewed as a "quantitative template," which serves as the foundation for the desired scientific form of qualitative methods. In KKV's view, standard research procedures of qualitative analysis are routinely problematic, and ideas drawn from conventional quantitative methods are offered as guideposts to help qualitative researchers be scientific.

A starkly different position has been emerging, forcefully expressed by the statistician David A. Freedman (2010a). He reviews the central role of qualitative analysis in six major breakthroughs from the history of epidemiology—a field highly relevant to political science because it faces many of the same challenges of causal inference based on observational data. He argues that in epidemiology and the social sciences, qualitative analysis is indeed a "type of *scientific* inquiry" (emphasis added), within the framework of recognizing multiple types. In characterizing this form of qualitative analysis, Freedman emphasizes the contribution of "causal-process observations" (CPOs). In Freedman's view, these strategically selected pieces of qualitative evidence play an essential role in disciplined causal inference. He comments pointedly that in the effort to advance knowledge,

> Progress depends on refuting conventional ideas if they are wrong, developing new ideas that are better, and testing the new ideas as well as the old ones. The examples show that qualitative methods can play a key role in all three tasks . . . . (Freedman 2010a: 232)

Relatedly, Freedman underscores the fragility of much quantitative analysis.

> . . . far-reaching claims have been made for the superiority of a quantitative template that depends on modeling—by those who manage to ignore the far-reaching assumptions behind the models. However, the assumptions often turn out to be unsupported by the data…. If so, the rigor of advanced quantitative methods is a matter of appearance rather than substance. (Freedman 2010a: 232)

Against the backdrop of these starkly contrasting views of appropriate methods, in the second edition of *Rethinking Social Inquiry* (Brady and Collier 2010) we have added 125 pages of new text that focus on: (1) ongoing controversy regarding KKV's legacy; (2) continuing criticism of the standard quantitative template, including regression modeling and estimates of uncertainty; and (3) emerging arguments about both qualitative and quantitative methods that hold the promise of strengthening tools for causal inference.[3] The following pages provide an overview of these themes.

Ongoing Controversy over King, Keohane, and Verba

The methodological positions adopted by KKV continue to be of great importance in political science and well beyond. The book has an exceptionally high level of citations, and year after year it receives impressive sales rankings with online book sellers.

In the period since the publication of our first edition in 2004, quantitative and qualitative methodologists alike have underscored KKV's contribution. Philip A. Schrodt, a quantitative methodologist, argues that it has been the "canonical text of the orthodox camp" among political methodologists. In many graduate programs, it is considered "the complete and unquestionable truth from on high" (Schrodt 2006: 335). On the qualitative side, James Mahoney notes the book's striking importance and remarkable impact in political science (2010: 120).

Ironically, achieving "doctrinal status was not necessarily the intention of KKV's authors" (Schrodt 2006: 336), and their perspectives have doubtless evolved in the intervening years. Yet notably, in 2002—eight years after the book's original publication—King published an extended, programmatic statement on methodology, nearly the length of a short book, entitled "The Rules of Inference" (Epstein and King 2002). This publication departs little from the arguments of KKV.

KKV is controversial, as well as influential, and its continuing importance is of great concern to scholars disturbed by its narrow message. Our first edition already contained strong criticisms, and new commentaries—some extremely skeptical—have continued to appear. These more recent arguments merit close examination.

* Note: This article draws on the Introductions to Parts I and II of Brady and Collier, *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 2nd ed. (Lanham, MD: Rowman & Littlefield, 2010).

Schrodt (2006) presents a bruising critique:

> KKV establishes as the sole legitimate form of social science a set of rather idiosyncratic and at times downright counterintuitive frequentist statistical methodologies that came together…to solve problems quite distant from those encountered by most political scientists…. (2006: 336)

Schrodt views KKV as promoting "a statistical monoculture" that is "not even logically consistent" (2006: 336). Indeed, he is convinced that

> one of the reasons our students have so much difficulty making sense of [KKV] is that in fact it does not make sense. (2006: 336)

Mahoney, in his comprehensive essay "After KKV: The New Methodology of Qualitative Research," argues that KKV has "hindered progress in political science" by "controversially and perhaps unproductively promoting a singular quantitative approach" (2010: 121).

Weyland (2005), with obvious annoyance, suggests that the authors of KKV "offered to help out their inferentially challenged qualitative brethren," proposing that their work should be "as similar as possible to quantitative studies." The book in effect makes claims of "quantitative superiority" that "rest on problematic assumptions" (2005: 392), thereby reinforcing the mindset in which "qualitative research was often seen as lacking precision and rigor and therefore undeserving of the 'methods' label" (2005: 392).

In discussing the first edition of our own book, *Rethinking Social Inquiry*, Schrodt suggests that in this polarized context, "adherents of the [methodological] orthodoxy consider the heresies proposed therein to be a distraction at best; a slippery slope…at worst" (2006: 335). To take one example, what we would view as one of the orthodox commentaries is found in Nathaniel Beck (2006, 2010), who treats the idea of causal-process observations as an "oxymoron"—thereby essentially dismissing a basic concept in our book. He repeatedly acknowledges that scholars should "understand their cases" (e.g. 2006: 350) and that qualitative evidence contributes to this background knowledge, but he questions the idea that causal-process observations meet acceptable standards for causal inference (352).

Schrodt views elements of the response to *Rethinking Social Inquiry* among mainstream quantitative researchers as reflecting an unfortunate, defensive reaction. He argues that

> many in the statistical community have taken criticism of any elements of the orthodox approach as a criticism of all elements and circled the wagons rather than considering seriously the need for some reform. (2006: 338)

He also notes that when the editor of the methodology journal *Political Analysis* announced at the 2005 summer methodology meetings that the journal planned a symposium on *Rethinking Social Inquiry*, the room responded as if to express concern that "there are traitors in our midst!" (2006: 338). Schrodt comments that this resistance reflects "a worrisome contentment with the status quo" among quantitative methodologists (2006: 338).

Based on this discussion, it seems clear that major controversies over methods stand behind these criticisms. We now explore two of these controversies.

Criticism of the Standard Quantitative Template

Statistical Modeling and Regression Analysis

In the past few years, the standard quantitative template centered on regression analysis has come under even heavier criticism. This development has two implications here. First, given KKV's reliance on this template, it further sharpens concern about the book's influence. Second, looking ahead, this development greatly expands the spectrum of methodological approaches that should be—and in fact are being—discussed and applied, among both methodologists and users of alternative methods.

Much of this discussion centers on the enterprise of statistical modeling that stands behind regression analysis. In important respects, the precariousness of work with regression derives from the extreme complexity of statistical models. A statistical model may be understood as "a set of equations that relate observable data to underlying parameters" (Collier, Sekhon, and Stark 2010: xi). The values of these parameters are intended to reflect descriptive and causal patterns in the real world.

Constructing a statistical model requires assumptions, which often are not only untested, but untestable. These assumptions come into play "in choosing which parameters to include, the functional relationship between the data and the parameters, and how chance enters the model" (Collier, Sekhon, and Stark 2010: xi). Thus, debates on problems of regression analysis are simultaneously debates on the precariousness of statistical models. It is unfortunate that more than a few quantitative researchers believe that when the model is estimated with quantitative data and results emerge that appear interpretable, it validates the model. This is not the case.

We agree instead with the political scientist Christopher H. Achen, who argues that with more than two or three independent variables, statistical models will "wrap themselves around any dataset, typically by distorting what is going on" (2002: 443). Thus, what we might call a "kitchen sink" approach—one that incorporates numerous variables—can routinely appear to explain a large part of the variance without yielding meaningful causal inference. Relatedly, Schrodt states that with just small modifications in the statistical model, estimates of coefficients can

> bounce around like a box of gerbils on methamphetamines. This is great for generating large bodies of statistical literature…but not so great at ever coming to a conclusion. (2006: 337)

The econometrician James J. Heckman emphasizes that "causality is a property of a model," not of the data, and "many models may explain the same data" (2000: 89). He observes that "the information in any body of data is usually too weak to eliminate competing causal explanations of the same phenomenon" (91).[4]

Sociologists have expressed related concerns, and Richard A. Berk concisely presents key arguments:

> Credible causal inferences cannot be made from a regression analysis alone…. A good overall fit does not demonstrate that a causal model is correct….There are no regression diagnostics through which causal effects can be demonstrated. There are no specification tests through which causal effects can be demonstrated. (2004: 224)

Berk amusingly summarizes his views in section headings within the final chapter of his book on regression analysis: "Three Cheers for Description," "Two Cheers for Statistical Inference," and "One Cheer for Causal Inference" (2004: chap. 11).[5]

Mathematical statisticians have likewise confronted these issues. Freedman's skepticism about regression and statistical modeling has been noted above, and his incisive critiques of diverse quantitative methods have now been brought together in an integrated volume that ranges across a broad spectrum of methodological tools (Freedman 2010b).

Also from the side of mathematical statistics, Persi Diaconis argues that "large statistical models seem to have reached epidemic proportions" (1998: 797), and laments the harm they are causing. He states that "there is such a wealth of modeling in the theoretical and applied arenas that I feel a sense of alarm" (804). Given these problems, methodologists should take more responsibility for the epidemic of statistical models by advocating "defensive statistics" (1998: 805). Thus, it should be a professional obligation to proactively warn scholars about the host of methodological problems summarized here.

In sum, many authors are now expressing grave concern about methods that have long been a mainstay of political and social science, and that are foundational for KKV.

Estimating Uncertainty

Standard practices in mainstream quantitative methods for estimating the uncertainty of research findings have also been challenged. The quest to estimate uncertainty is quite properly a high priority, prized as a key feature of good research methods. KKV views understanding and estimating uncertainty as one of four basic features of scientific research (1994: 9). In its discussion of "defining scientific research in the social sciences," the book states that "without a reasonable estimate of uncertainty, a description of the real world or an inference about a causal effect in the real world is uninterpretable" (9). The received wisdom on these issues is central to mainstream quantitative methods.

Unfortunately, KKV presumes too much about how readily uncertainty can be measured. In the original debate over King, Keohane, and Verba, for example, Larry M. Bartels (2010) argues that these authors greatly overestimate the value of the standard insight that error in an independent as opposed to dependent variable affects findings in different ways. Bartels demonstrates that this would-be insight is incorrect.

A more pervasive problem involves significance tests. Any scholar acquainted with conventional practice in reporting regression results is well aware of the standard regression table with "tabular asterisks" scattered throughout.[6] The asterisks indicate levels of statistical significance, calculated on the basis of the standard errors of the coefficients in the table. Too often, when researchers report their causal inferences they simply identify the coefficients that reach a specified level of statistical significance. This is a dubious research practice.

A central problem here is that findings reported in regression tables are routinely culled from numerous alternative specifications of the regression model, which obviates the standard meaning and interpretation of the asterisks. Once again, Schrodt states the objection with particular clarity:

> The ubiquity of exploratory statistical research has rendered the traditional frequentist significance test all but meaningless. (2006: 337)

Freedman and Berk (2010: 24) underscore the dependence of significance tests on key assumptions. For descriptive inference (external validity), they assume a random sample, rather than the convenience sample common in political science. Even with a random sample, missing data—including the problem of non-respondents—can make it more like a convenience sample.[7] Another assumption requires a well-defined—rather than ill-defined or somewhat arbitrarily defined—population.

For causal inference (internal validity), avoiding data snooping is crucial if significance tests are to be meaningful. Here, the presumption is that the researcher has begun with a particular hypothesis and tested it only once against the data, rather than several times, adjusting the hypothesis and model specification in the search for results deemed interesting. This inductive approach is definitely a valuable component of creative research, but it muddies the meaning of significance tests. Against this backdrop, Freedman, Pisani, and Purves (2007) are blunt in their warnings:

> 1. "If a test of significance is based on a sample of convenience, watch out." (556)
>
> 2. "If a test of significance is based on data for the whole population, watch out." (556)
>
> 3. "Data-snooping makes P-values hard to interpret." (547)
>
> 4. "An 'important' difference may not be statistically significant if the N is small, and an unimportant difference can be significant if the N is large." (553)[8]

A key point should be added. In his various single-authored and co-authored critiques of significance tests, Freedman specifically does not turn to the alternative of Bayesian analysis. Rather, as in his other writings on methodology, he advocates common sense, awareness that statistical tools have major limitations, and substantive knowledge of cases as an essential foundation for causal inference.[9]

Where Do We Go from Here?

The practical importance of these problems is quickly seen in the fact that, to a worrisome degree, a great deal of quantitative research in political science has proceeded as if regression-based analysis, including accompanying measures of uncertainty, yields reliable causal inference. A vast number of journal articles have sought to make causal inferences by estimating perhaps half a dozen related (and commonly under-theorized) model specifications, picking and choosing among these specifications, and offering what is too often an ad hoc interpretation of a few selected coefficients—generally, quite inappropriately, on the basis of significance levels. These failings have been further exacerbated by the readily available statistical software that makes it easy for researchers with virtually no grasp of statistical theory to carry out complex quantitative analysis (Steiger 2001).

In the face of these serious challenges, we advocate two avenues of escape: refinement of qualitative tools, and further innovation in quantitative methods. These alternatives are explored in the new chapters of *Rethinking Social Inquiry*, second edition.[10]

Qualitative Methods

An important avenue is opened by further refinements in qualitative analysis. One familiar, traditional option here is the small-N comparative method, a strategy common both in cross-national comparisons and in comparisons of political units within nations—potentially involving regions, provinces or states, or metropolitan areas. Here, the analyst juxtaposes perhaps two, or four, or six units, selecting matching and contrasting cases so as to "control" for extraneous factors and allow a focus on the principal variables of concern. This approach is often identified with J. S. Mill's (1974 [1843]) methods of agreement and difference, and Przeworski and Teune's (1970) most similar and most different systems designs.

In our view, this small-N comparative approach is invaluable for concept-formation, pinning down indicators, and discovering and formulating explanatory ideas. However, it is much weaker as a basis for causal inference, given that it involves what is in effect a correlation analysis with a very small N. The matching and contrasting of cases employed cannot by itself succeed in controlling for variables the researcher considers extraneous to the analysis. Any presumption that this matching of cases creates a natural experiment or a quasi-experiment is misleading.

Rather, the key step is to juxtapose this comparative framing with carefully executed research carried out within the cases, quite often involving qualitative analysis. At this point, the tools of causal-process observations and process tracing become crucial. They certainly do not solve all problems of qualitative analysis, yet they make valuable contributions.

Regarding the first of these, a central element in our effort to place qualitative analysis on a more rigorous foundation (Collier, Brady, and Seawright 2010) has been our distinction between: (1) data-set observations (DSOs), which correspond to the familiar rectangular data set of quantitative researchers; and (2) causal-process observations (CPOs), i.e., pieces of data that provide information about context, process, or mechanism and contribute distinctive leverage for causal inference in qualitative research.[11]

Causal-process observations and the widely discussed procedure of "process tracing" are closely connected. When *process tracing* is used for causal inference,[12] the pieces of evidence on which the researcher focuses are specifically *causal-process observations*. Process tracing consists of procedures for singling out particular CPOs that are relevant for causal inference in a given context.

In our second edition, Andrew Bennett, David A. Freedman, and Henry E. Brady explore different facets of causal-process observations and process tracing.

Bennett's (2010) chapter provides a new introduction to process tracing, which is routinely invoked as a basis for causal inference in qualitative work. Unfortunately, the specific steps involved are often poorly understood—possibly one reason why many quantitative researchers are skeptical about causal inference in qualitative studies.

Bennett formulates a typology that places on two dimensions the alternative tests employed in process tracing. The tests are distinguished according to whether passing a particular test is *necessary* for inferring causation, and whether it is *sufficient*. This provides a new framework for thinking about the tests originally proposed by Van Evera (1997: 31–32): the straw-in-the wind, hoop, smoking-gun, and doubly decisive tests. Bennett illustrates his framework by applying the typology at the level of macro-politics, focusing on three well-known historical episodes in international relations.

Next, the statistician David Freedman's (2010a) chapter examines the role of causal-process observations in six major studies from the history of epidemiology, including John Snow's famous research on the causes of cholera. In Freedman's view, in both epidemiology and the social sciences, the use of qualitative evidence can be an important type of scientific inquiry in its own right, and it can potentially make a greater contribution than conventional quantitative approaches. More than a few quantitative researchers need to examine Freedman's views and reconsider their skepticism about inference in qualitative research.

Freedman is very specific about the contributions of qualitative evidence, arguing that it can play a valuable role in overturning prior hypotheses, as well as formulating and testing new hypotheses. In this sense, Freedman's position is quite different from that of Fearon and Laitin (2008: 756), who sharply subordinate qualitative vis-à-vis quantitative analysis. It also challenges Piore's (2006: 17) assertion that information from case studies "cannot be treated directly as empirical evidence."

Yet Freedman is also strongly committed to the careful juxtaposition of CPOs and DSOs. He is skeptical about much quantitative analysis, and he prefers quantitative research—for example, natural experiments—that is carried out jointly with careful qualitative analysis.

Henry Brady's (2010) chapter, focused on electoral behavior, shows how a sequence of CPOs can yield causal inference. This procedure gives crucial leverage in disputing the claim, advanced in a quantitative study by John Lott, that George Bush lost thousands of votes in the 2000 presidential election in Florida due to the media's early (and incorrect) call of the election outcome. Brady's working hypothesis is that the early call had little impact on the actual vote. He goes through a series of process-tracing steps to support his hypothesis, employing a sequence of vote counts and assumptions about voting behavior that identify the necessary conditions for Lott's hypothesis to be plausible.

Brady's analysis reminds us of the obvious but crucial point that process tracing can employ numerical data. The use of such data does not necessarily involve DSOs, i.e., a rectangular data set. Rather, isolated pieces of quantitative information are treated as CPOs.

Looking beyond these three chapters, we believe—notwithstanding extensive efforts to institutionalize graduate training in qualitative methods[13]—that this training does not adequately address process tracing and causal-process observations. This deficit has motivated us to prepare a set of exercises, available online, for teaching these analytic tools.[14]

Quantitative Methods

The final two chapters in the second edition of *Rethinking Social Inquiry* consider further the problems with quantitative methods and potential solutions to these problems. Jason Seawright's (2010) "Regression-Based Inference: A Case Study in Failed Causal Assessment" examines a number of unsuccessful attempts to employ quantitative, cross-national regression analysis to address a classic topic in comparative social science: the impact of political regime type on economic growth.

Seawright's substantive focus is of great importance, and his analysis has broad implications for the vast literature that uses the quantitative, cross-national method to study dozens of major topics in political and social science. Indeed, in discussing these failures of inference, he cites authors who call for an "obituary" for a significant part of quantitative cross-national literature. His arguments are also relevant for more general discussions of regression analysis. For example, many scholars believe they can improve regression-based inference by simply adding more control variables. Yet introducing further controls can potentially make inferences worse, rather than better. Seawright goes on to show that refinements in regression—such as designs employing matching and instrumental variables—are at best problematic in addressing these problems. In conclusion, he proposes that scaling down to a finer-grained focus, which may include substantial use of qualitative analysis, can come to the rescue in overcoming these failures.

The last chapter, by Thad Dunning (2010), evaluates efforts at "design-based inference." He focuses on the family of techniques known as natural experiments, including regression-discontinuity and instrumental-variables designs. These techniques seek to overcome the failures of regression explored by Seawright and discussed above. Dunning creates a typology that assesses these designs on three dimensions: (1) plausibility of the "*as-if* random assignment," an assumption central to natural experiments; (2) credibility of the statistical model; and (3) substantive importance of the key explanatory variable. This third point is especially significant because some of these studies, in the search for situations of *as-if* random assignment, drastically limit the substantive relevance of their investigations. Indeed, this shortcoming is a serious weakness of natural experiments.

By juxtaposing the three dimensions in his typology, Dunning evaluates whether particular natural experiments employ strong or weak research designs. He identifies a small number of unusually successful natural experiments, yet he also shows that many famous studies in this tradition do poorly in terms of his criteria. Hence, although Dunning views natural experiments as a promising tool, he argues that they should definitely not eclipse other methodologies.

Finally, Dunning underscores repeatedly the importance of qualitative data and qualitative insight in the design and execution of natural experiments. Without this foundation, researchers lack a basis for many judgments and decisions essential to the method—for example, the key assumption of *as-if* random assignment. Hence, this newly popularized methodology is strongly dependent on traditional tools—qualitative evidence, hard-won insight into the details of cases, and knowledge of context. It is a striking example of why a multi-method approach is invaluable in good research.

To summarize: We are unquestionably observing a sea change in political methodology. Conventional quantitative methods are now the focus of even sharper criticism. The tools of qualitative analysis are being further refined and legitimated in ways that address some of these failings. These qualitative tools certainly have their own problems and limitations, but we think real progress is being made. Natural experiments likewise show promise, though they are far from being a methodological Nirvana. Correspondingly, in our view an eclectic practice of methodology—involving the idea of "diverse tools" highlighted in the title of our book—is the most promising avenue to pursue.

Notes

[1] Morgan (1996) provides a broad overview of rival views of science, encompassing the natural, biological, and social sciences.

[2] For our own work, we share David Freedman's (2010a) view of plurality in scientific methods, and we also recognize social versus natural science as partially different enterprises. Yet the two can and should strive for careful formulation of hypotheses, intersubjective agreement on the facts being analyzed, precise use of data, and good research design (though obviously, there is wide disagreement as to how these desired outcomes should be achieved, and whether in any given context they are achieved). Within this big-tent understanding of science, we are happy to be included in the tent.

[3] A further initial point should be underscored. The focus in both editions of *Rethinking Social Inquiry* is on the study of causes and consequences—and specifically on causal inference. Of course, this focus represents just one facet of methodology. In our own work we have written extensively on conceptualization and measurement, and indeed, assessing causes and consequences emphatically calls for careful attention to concept-formation and operationalization. Yet the central concern here is with causal inference.

[4] From the standpoint of econometrics, see also Leamer (1983: 36–38).

[5] Related arguments of sociologists have been advanced by Morgan and Winship (2007: passim) and Hedström (2008: 324). Statements by psychometricians include Loehlin (2004: 230–234).

[6] Meehl (1978), cited in Freedman and Berk (2010: 24).

[7] See Freedman (2008: 15). Thus, starting with a random sample, in the face of problems such as resource constraints that limit tracking down respondents, the researcher can end up with what is in effect a type of convenience sample.

[8] I.e., if assumptions are not met, "significance" level depends on the sample size, without reflecting the real meaning of statistical significance.

[9] See, for example, Freedman (2008, 2010a).

[10] Four chapters are entirely new. Henry Brady's analysis of data-set observations versus causal-process observations was an appendix in the first edition.

[11] In quantitative research, the idea of an "observation" (as in DSO) has special status as a foundation for causal inference, and we deliberately incorporated this label in the idea of CPOs to underscore their relationship to causal inference. We are pleased that other scholars have also found the idea of CPOs to be useful. For example, see Mahoney (2010: 123–131), who uses it as one of the organizing principles in his synthesis of new trends in qualitative methods. The concept is also employed by Freedman (2010a).

[12] Mahoney (2010: 125–128) explains how it is also used for descriptive inference.

[13] Notable among these efforts, the APSA Organized Section for Qualitative and Multi-Method Research (which publishes this newsletter) sponsors short courses at the annual political science meetings, as well as numerous panels of great pedagogical value. The annual Institute for Qualitative and Multi-Method Research at Syracuse University also provides excellent instruction. These initiatives are discussed in Collier and Elman (2008).

[14] "Exercises for Teaching Process-Tracing" are found at the top of David Collier's UC Berkeley website: polisci.berkeley.edu/people/faculty.

References

Achen, Christopher H. 2002. "Toward a New Political Methodology: Microfoundations and ART." *Annual Review of Political Science* 5, 423–450.

Bartels, Larry M. 2010. "Some Unfulfilled Promises of Quantitative Imperialism." Chap. 6 in Brady and Collier, eds. *Rethinking Social Inquiry*, 2nd ed. (Lanham, MD: Rowman & Littlefield).

Beck, Nathaniel. 2006. "Is Causal-Process Observation an Oxymoron?" *Political Analysis* 14:3 (Summer), 347–352.

Beck, Nathaniel. 2010. "Casual-Process 'Observation': Oxymoron or (Fine) Old Wine." *Political Analysis* 18:4 (Autumn), 499–505.

Bennett, Andrew. 2010. "Process-Tracing and Causal Inference." Chap. 10 in Brady and Collier, eds. *Rethinking Social Inquiry*, 2nd ed. (Lanham, MD: Rowman & Littlefield).

Berk, Richard A. 2004. *Regression Analysis: A Constructive Critique*. Thousand Oaks, CA: Sage.

Brady, Henry E. 2010. "Data-Set Observations versus Causal-Process Observations: the 2000 U.S. Presidential Election." Chap. 12 in Brady and Collier, eds. *Rethinking Social Inquiry*, 2nd ed. (Lanham, MD: Rowman & Littlefield).

Brady, Henry E., and David Collier, eds. 2004/2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 1st and 2nd eds. Lanham, MD: Rowman and Littlefield.

Collier, David, and Colin Elman. 2008. "Qualitative and Multi-Method Research: Organizations, Publication, and Reflections on Integration." In *Oxford Handbook of Political Methodology*. Janet Box-Steffensmeier, Henry E. Brady, and David Collier, eds. (Oxford: Oxford University Press).

Collier, David, Henry E. Brady, and Jason Seawright. 2010. "Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology." Chap. 9 in Brady and Collier, eds. *Rethinking Social Inquiry*, 2nd ed. (Lanham, MD: Rowman & Littlefield).

Collier, David, Jasjeet Sekhon, and Philip Stark. 2010. "Editors' Introduction: Inference and Shoe Leather." In David A. Freedman, *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. David Collier, Jasjeet Sekhon, and Philip B. Stark, eds. (Cambridge: Cambridge University Press).

Diaconis, Persi. 1998. "A Place for Philosophy? The Rise of Modeling in Statistical Science." *Quarterly of Applied Mathematics* 56:4 (December), 797–805.

Dunning, Thad. 2010. "Design-Based Inference: Beyond the Pitfalls of Regression Analysis?" In Brady and Collier, eds. *Rethinking Social Inquiry*, 2nd ed. (Lanham, MD: Rowman & Littlefield).

Epstein, Lee, and Gary King. 2002. "The Rules of Inference." *University of Chicago Law Review* 69:1 (Winter), 1–133.

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

Freedman, David A. 2008. "Rejoinder." *Qualitative and Multi-Method Research* 6:2 (Fall), 14–16.

Freedman, David A. 2010a. "On Types of Scientific Inquiry: The Role of Scientific Reasoning." In Brady and Collier, eds. *Rethinking Social Inquiry*, 2nd ed. (Lanham, MD: Rowman & Littlefield).

Freedman, David A. 2010b. *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. David Collier, Jasjeet Sekhon, and Philip B. Stark, eds. Cambridge: Cambridge University Press.

Freedman, David A., Robert Pisani, and Roger Purves. 2007. *Statistics*, 4th ed. New York: Norton.

Freedman, David A., and Richard A. Berk. 2010. "Statistical Assumptions as Empirical Commitments." In David A. Freedman, *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. David Collier, Jasjeet S. Sekhon, and Philip B. Stark, eds. (New York: Cambridge University Press).

Heckman, James J. 2000. "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective." *Quarterly Journal of Economics* 115:1 (February), 45–97.

Hedström, Peter. 2008. "Studying Mechanisms to Strengthen Causal Inferences in Quantitative Research." In *Oxford Handbook of Political Methodology*. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, eds. (Oxford: Oxford University Press).

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

Leamer, Edward E. 1983. "Let's Take the Con out of Econometrics." *American Economic Review* 73:1 (March), 31–43.

Loehlin, John C. 2004. *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Mahoney, James. 2010. "After KKV: The New Methodology of Qualitative Research." *World Politics* 62:1 (January), 120–147.

Meehl, Paul E. 1978. "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology." *Journal of Consulting and Clinical Psychology* 46:4 (August), 806–834.

Mill, John Stuart. 1974 [1843]. "Of the Four Methods of Experimental Inquiry." *A System of Logic*. Toronto: University of Toronto Press.

Morgan, Pamela S. 1996. "'What Is Science?' and the Cognitive Theory of Contested Concepts." Department of Linguistics, University of California, Berkeley.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals* *and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

Piore, Michael J. 2006. "Qualitative Research: Does it Fit in Economics?" *European Management Review* 3:1 (Spring), 17–23.

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

Seawright, Jason. 2010. "Regression-Based Inference: A Case Study in Failed Causal Assessment." In Brady and Collier, eds. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 2nd ed. (Lanham, MD: Rowman & Littlefield).

Schrodt, Philip A. 2006. "Beyond the Linear Frequentist Orthodoxy." *Political Analysis* 14:3 (Summer), 335–339.

Steiger, James H. 2001. "Driving Fast in Reverse." *Journal of the American Statistical Association* 96:453 (March), 331–338.

Van Evera, Stephen. 1997. *Guide to Methods for Students of Political Science*. Ithaca, NY: Cornell University Press.

Weyland, Kurt. 2005. "Book Review." *Perspectives on Politics* 3:2 (June), 392–393.

Symposium: Perfecting Methodology, or Methodological Perfectionism?

Large-N Observational Data Analysis (aka Messy Data): A Modest Defense

John Gerring
Boston University
jgerring@bu.edu

A specter is haunting political science. It is the specter of methodological perfectionism.[1] This dogma places methods before substance and imposes a narrow spectrum of acceptable methods on the discipline.[2]

There are of course many varieties of methodological perfectionism, and political science has experienced more than a few over the past century. Methodologists from the "statistical" side of the discipline (aka PolMeth, quantitative methods) generally prize causal knowledge over descriptive knowledge, theory appraisal over theory discovery, micro-theory (aka micro-mechanisms) over macro-theory, and internal validity over external validity.

Methodologists from the "qualitative" (aka case study, QCA) side of the discipline share most of these preferences, with the notable exception of theory appraisal. (They value work that is exploratory rather than confirmatory.) I will not address qualitative work in the interpretive tradition, as this tradition—while it certainly has its own version of perfection—is not as influential at the present time.

So defined, both the quant and qual side of the methodological divide agree on the nature of the problem we are facing today. Broadly stated, the most common method of drawing causal inferences in social science—based on large samples with no pretense of a randomized treatment ("regression")—does not work. Consequently, we need to re-think the traditional approach to causality.

For those trained in statistics, the way forward is to be found in experimental or quasi-experimental evidence. Any imperfections in the assignment process or in the post-treatment period should be handled by appropriate statistical procedures—via matching, instrumental variables, regression-discontinuity models, and the like (Angrist and Pischke 2010; Gerber, Green, and Kaplan 2004; Imbens and Wooldridge 2009; Morgan and Winship 2007; Rubin 2005).

For those trained in qualitative methods, the way forward lies in case-based methods, e.g., (a) tracing a discrete process from the purported cause, through its various mechanisms, to an effect, (b) causal-process observations, (c) counterfactual thought-experiments, (d) well-matched cases that satisfy the strictures of a most-similar or most-different case design, or (e) comparative-historical work, which combines elements of the foregoing (Bennett 2010; Bennett and Elman 2006; Brady and Collier 2010; George and Bennett 2005; Gerring 2007; Mahoney 2000, 2010; Mahoney and Rueschemeyer 2003).

Many methodologists embrace both of these solutions (e.g., Freedman 2008; Seawright 2010), as do I.[3] Indeed, qualitative evidence is often quite important for conducting a strong experimental or quasi-experimental design (Cook et al. 2010; Dunning 2008b; Paluck 2010; Rosenbaum 2010: 323–24). I do not want to portray these two solutions as necessarily in conflict with one another. Nonetheless, they are quite distinct approaches to causal inference, and I shall treat them as such.

I want to propose that both the experimental and case-based approaches to causal inference are valid, but—here are the crucial caveats—only if understood as ideals rather than uniform thresholds of adequacy, and only if understood within the larger context of methodological objectives and tools that have traditionally been applied to questions of social science (and new tools that are now entering the lexicon, such as randomization inference and extreme bounds analysis). The costs of adopting stricter methodological standards must be reckoned along with the benefits. Tradeoffs—e.g., between causal and descriptive knowledge, theory appraisal and theory discovery, micro-theory and macro-theory, internal and external validity—are inescapable.

Let us consider these costs in somewhat more detail. (This ground has been covered many times, but some mention of these issues is important in order to properly frame the main argument.) Although experiments usually achieve a high degree of internal validity, there is often a sacrifice in external validity or in the type of problems that can be addressed.[4] Natural experiments are wonderful tools but are limited to circumstances of extraordinary serendipity, and are not always easy to generalize from (Dunning 2008a). The current raft of econometric tools are fairly easy to apply (given handy statistical software packages), but do not always rectify the problems they are designed to rectify. A prime example in point is the instrumental variables approach to causal inference, where one often finds a facile use of a technique without any real acknowledgment of its potential problems (e.g., violation of the exclusion restriction).

Likewise, case-based causal inference is easy to practice but hard to practice well. Confounders are generally legion, even in carefully matched cases (the most-similar form of comparison), and counterfactual thought-experiments are not always sufficient to eliminate them. Strong process tracing research usually depends on strong and specific theoretical predictions about the causal mechanisms at work. However, most social-science theories do not issue highly specific predictions about process and outcome, or they specify a number of possible mechanisms, none of which is necessary for Y (causal equifinality). Likewise, strong process-tracing is usually possible only with proximal causal relationships; distal causes, which compose a large share of social science theories, are difficult to process-trace. The question, to adopt the vocabulary employed by Bennett (2010), is how often Hoop tests, Smoking Gun tests, or Doubly Decisive tests can be applied to case study evidence. Is this a realistic standard of evidence for most case studies? (Bennett [2010: 219] specifies, "The evidence must strongly discriminate between alternative hypotheses.") Likewise, do case studies claiming to have applied these tests do so with integrity? Just as it is important to question dubious assertions about "natural experiments," it is important to question assertions of slam-dunk process-tracing. And even when internal validity can be established, it is often difficult to generalize from a case study.[5]

What worries me is that, insofar as our discipline's current experimental and case-based standards are taken seriously (i.e., interpreted strictly), a very high bar is being set for admission into the ranks of political science, a standard that only a small minority of work currently satisfies (and that much of my own work certainly does not). This sort of methodological perfectionism puts researchers in a situation where they may (a) pretend that they have attained methodological purity when they have not (in order to convince colleagues and reviewers), (b) privately feel morose about their work, and/or (c) abandon their substantive/theoretical goals in favor of things that might be studied in a way that satisfies current standards. None of these developments are especially healthy for the discipline—not to mention for the individuals involved.

Of course, one might argue that a surge of methodological extremism is precisely what is needed at the present time, as a tonic. If lots of what we currently do does not pass the more exacting standards that the discipline has adopted, it could be a sign that we need to buckle up and work harder, and learn some new tricks—not that we should relax our standards. And if current methods are not working very well because the nature of the data on certain subjects is recalcitrant, perhaps it makes sense to reallocate our energies towards problems that are more tractable. To this extent, methods may legitimately drive substance (i.e., in situations where little progress can be made on Subject A and lots of progress on Subject B).

The advent of experimental and quasi-experimental standards has had the salutary effect of raising methodological consciousness in the discipline (a good thing, from most perspectives) and has resulted in some superb work (a good thing, from all perspectives). And now that incentives of political scientists are strongly aligned to promote first-class research designs, the ingenuity of the discipline should be unleashed. At the end of a decade or so we should be able to answer the question that has dogged proponents of these methods all along: what is the possible purview of such methods (i.e., experiments, natural experiments, and strong case-based knowledge) for political science questions? And what is the value-added for the discipline of having these first-class studies? Do they cumulate into broader theoretical insights?

Perhaps the current evangelical spirit represents a temporary swing of the pendulum—an entirely appropriate reaction against the traditional, regression-based species of political science that became the hallmark of the behavioralist movement. In a few years the fever may pass and we will find some reasonable middle ground.

If this is what the future holds, we need to start thinking about what that middle ground might consist of. And if not, i.e., if the current methods juggernaut continues, we need to figure out what the end-result of our collective move toward higher quantitative and qualitative standards might be. Either way, our subject matter is highly consequential.

I must introduce one note of clarification before proceeding. The spirit of methodological perfectionism that I am describing is most noticeable among methodologists and among those who follow current developments in the methodological literature. Recent PhDs are likely to be more aware of these trends than older scholars. Those who attempt to publish in the very top mainstream journals (e.g., *APSR*, *AJPS*, *JOP*, *BJPS*, *WP*) are more likely to be aware of these trends than those who write for subfield journals or who publish books rather than articles. Research universities are more prone to these developments than liberal arts colleges.

For some scholars, this debate may seem like a debate from nowhere. However, insofar as the spirit of methodological perfectionism pervades the discipline's top journals and top departments, it affects the direction of political science at large. Consequently, those who feel distant from these currents may still be affected by the tide.

Messy Data

Despite frequent espousals of pluralism, methodologists seem to agree that there is one category of endeavor that does not deserve forbearance. This excluded category may be described as *large-N observational research where the treatment bears no resemblance to randomized assignment, where the assignment principle is not known, where there are no good instruments, and where there is no opportunity for convincing process-tracing*.

This mouthful is sometimes subsumed under the rubric of "regression." However, regression methods are often used for the analysis of experimental data and quasi-experimental data. Moreover, other (non-regression) estimators—e.g., difference of means tests, matching, randomization inference—can be used to analyze large-N samples of the sort described above.

In truth, the blighted patch of desert described by the passage in italics is a residual category. It includes any research design that doesn't pass either the quantitative (experimental/quasi-experimental) or qualitative (case-based) thresholds of adequacy, and whose purpose is causal inference (not merely description). I shall refer to this nebulous area henceforth as *large-N observational data*, aka *messy data*. Assume by this phrase that violations of stable unit treatment value assumption (SUTVA) and ignorability are serious concerns; confounders cannot be ruled out. Therefore, the internal validity of the study is seriously challenged. Perhaps most worrisome of all, the researcher has no way of determining whether the assumptions necessary for causal inference have been met *unless* a parallel experiment has been carried out. Thus, the only time we can be pretty sure that messy large-N observational research is accurate is when it is redundant.[6]

In the face of such damning conclusions, it is easy to see why there has been a flight by methodologists from large-N observational data to other approaches. David Freedman (1997: 114; emphasis added) states baldly, "I see no cases in which regression equations, let alone the more complex methods, have succeeded as engines for discovering causal relationships." While this may be viewed as somewhat extreme, Freedman articulates a widely-held skepticism toward causal analyses based on ex post statistical adjustments. Of late, regression has become the whipping-boy of methodologists, both quant and qual. It is perhaps the only area of agreement one finds today that stretches across the entire methodological spectrum—from positivism (so-called) to interpretivism (so-called).[7] A sign of this new consensus is that it has become a sign of sophistication to scoff at "correlations" and to describe them as "descriptive" or as "stylized facts."

Let me be clear. I have no argument with these critiques. However, I want to argue that Freedman's conclusion is wrong. More important, I want to argue that it misframes the question we ought to be asking.

Let me now utter a few platitudes that I assume most readers will be willing to entertain, at least as a point of departure.

1. We learn about the world in myriad ways—including common sense, personal experience, secondary research (conducted by others), theoretical suppositions, deductive logic, exploratory data analysis, and (last and perhaps least) formally devised research designs. Of these, some conform to the current template of acceptable research, while most do not.

2. Different problems demand different approaches, and political science encompasses an extraordinarily broad set of theoretical frameworks and empirical data.

3. Where others have been to some extent determines where it might be fruitful to go. A good research design is understandable only in the context of a particular research tradition, where triangulation on a common problem is often useful.

If readers are willing to accept these truisms, it follows that no single methodological standard is likely to be applicable to all political science work (unless that standard is extremely abstract, i.e., on the level of philosophy of science). It also suggests a more flexible methodological standard for work in political science, one that might include large-N observational data analysis.

Note that many practicing political scientists continue to employ messy observational data as their empirical workhorse. Thus, my argument may be understood as a qualified defense of the status quo—qualified, in the following manner.

I want to argue that the question we ought to be asking is not whether method A or B is adequate—by some absolute standard of adequacy—but rather (a) whether it adds to our knowledge of a subject, (b) whether it is the best method (or one of several equally good methods) for the job, and (c) whether an accurate assessment of overall uncertainty (not simply statistical uncertainty) is attached to the conclusions. If these criteria are met, then the study ought to be considered methodologically adequate, even if far from ideal. It follows that a study may be extremely shaky but still adequate, so long as it allows us to update our priors, beats the alternatives, and presents a plausible uncertainty estimate. The slogan is *best-possible*—rather than *best* (Gerring 2011a).

The foregoing statement pertains to methodology considered in its narrowest sense, i.e., pertaining to internal validity and precision. I would also argue that we need to find a way to incorporate other goals into our understanding of methodology. This includes, most importantly, the theoretical contribution of a study (its breadth of application and commensurability with other work). I don't imagine that there will be much dissent from this argument; I raise it only so that readers can keep it in view, and because it sometimes militates toward large-N observational data and away from some of the more rarified methods, which are often limited in external validity and/or theoretical fecundity.

Messy Data at Work: Democracy as a Dependent Variable

The way to prove my point—and to prove that Freedman is wrong—is not to engage the question at the abstract level of philosophy or methodological theory, where complex issues are usually difficult to resolve. Rather, following Freedman (who was fond of dissecting published work), I must show that for some questions of importance to political science, large-N observational data provides the best (or a best) approach to the problem, there is some value-added to our understanding of a subject, and reasonable estimates of uncertainty can be arrived at (through statistical procedures such as Bayesian inference, randomization inference, or extreme bounds and/or through qualitative reasoning).

I shall begin with a research area that provides a tough test for my argument: crossnational regressions with institutions on the left and right side of the model. It is a tough test because such analyses are characterized by many of the features that lead methodologists to despair of ever reaching causal inferences based on observational data. Institutions are broad, abstract phenomena—democracy, development, rule of law, property rights, veto points—that are difficult to conceptualize, much less to measure. Even where they can be measured, they are difficult to interpret causally because they are usually not manipulable (perhaps not even in principle, although this may be debated). Institutions tend to be slow-moving and thus provide little change over observable periods, usually limited to the past half-century. Institutions are highly correlated—good things go together, as do bad things—so it is difficult to tease apart the signal from the background noise and from potential confounders. One rarely finds quasi-experimental assignment to treatment; endogeneity is the norm. Good instruments (ones that are strongly correlated with the treatment and do not affect the outcome) are rare. And the units of interest—nation-states—are remarkably heterogeneous. (Some might even wonder if they belong in the same sample.)

Recall, however, the burden of the argument: not that good inferences result from such bad data but that better inferences may result from large-N observational data analysis than from other types of analyses. Indeed, many of the characteristics of the typical crossnational regression also pose problems for experimental or quasi-experimental analysis or for convincing case-based analysis (Coppedge forthcoming: ch 5; Gerring and Thacker 2008: ch 7; Seawright 2010). It is not clear that there is a good alternative.

Now, let us turn to a specific example. I shall focus on an area well known to most readers (and certainly to all comparativists) by virtue of the volume of work that has been devoted to it and the prominence of the theory—*modernization*. This refers to the relationship between development and democracy, a topic first tackled in a serious way by Seymour Martin Lipset (1959), and since explored by myriad studies, generally in a regression framework (e.g., Boix and Stokes 2003; Casper and Tufis 2003; Epstein et al. 2006; Przeworski and Limongi 1997).

Two findings have emerged from this body of work. The first is that development (as measured by GDP per capita) helps democracies consolidate. The richer a democracy, the less likely it is to relapse into authoritarian rule. The second is that development has a very small—perhaps even nonexistent—effect on democratization. Rich countries are only slightly more likely to democratize than similarly situated poor countries. This second finding is still contested, but the boundaries of the possible relationship seem clear. Development has a small or null effect on democratization.

Our questions of interest are whether regression (i.e., large-N observational data analysis) has added to our knowledge of the subject, whether it is the best method (or one of several equally good methods) for the job, and whether an accurate assessment of overall uncertainty has been attached to the conclusions of the cited studies. I think the answers are: yes, yes, and probably not. Thus, in two respects messy-data methods proved their worth, though perhaps not in the third.

However, the one failing is by no means irremediable (see Glynn, below). Indeed, it is partly a product of the methodological perfectionism that I am complaining about. Authors feel compelled to present their findings as if they arose from experimental or quasi-experimental evidence, even though they probably know (in their heart of hearts) that their t statistics do not encompass many threats to inference situated in the research design. If I am correct in these conclusions, then a modest case for large-N observational data has been established, at least in one instance.

Michael Coppedge's magisterial survey, *Approaching Democracy: Research Methods in Comparative Politics* (forthcoming), examines what we know about democratization, and how we know it. Here, a much broader survey of the methodological and substantive ground is covered. Because it bears directly on our question, I will paraphrase at some length.

Coppedge does not discuss any experimental or quasi-experimental work, suggesting that this method has yet to be applied (or has not been successfully applied) to the topic. Case study work on democratization is voluminous, and Coppedge is impressed by the rich narrative histories of specific cases as well as the number of potentially fruitful general hypotheses that might be garnered from this corpus. However, he also notes that there are many threats to inference, even with respect to the cases under study (problems of internal validity). Moreover, the broader hypotheses that might be gleaned from the case study literature have been difficult to generalize from because they are developed in the context of specific cases and are not always couched in ways that would lend themselves to broader application. Coppedge concludes that theory generation, not theory testing, is the province of case studies in the area of democratization. "Histories and case studies are great ways to develop ideas about things that *may* matter generally, but cannot show that they *do* matter generally" (Coppedge forthcoming: 18–19, chap. 5).

For theory testing, Coppedge concludes that crossnational statistical analysis is required—even though it may not always be sufficient, for all the reasons we have discussed (and which Coppedge discusses in much greater detail). Likewise, Coppedge criticizes crossnational analyses for their lack of theoretical integration: researchers settle on slightly different operationalizations, samples, and/or estimators and, as a result, their findings do not cumulate. Nonetheless, he hazards the following conclusions, based on his reading of the literature (and his own analyses). Factors that seem to have little impact on any democracy outcome include "land area, population, age of the country, the rule of law (as currently measured), colonial rule (without differentiating among colonial powers), and linguistic fragmentation." Factors that have some impact on at least one outcome measure of democracy, and have been confirmed by multiple studies and extensive robustness tests, are summarized as follows:

> Income (the log of per capita GDP) is associated with higher cross-national levels of democracy; income and economic growth are both associated with a higher probability of survival as a democracy and a lower probability of transition. Greater absolute changes in level and a higher probability of breakdown are found in the more unequal societies. Rentier states tend to be less democratic. And religiously fragmented societies are less stable: more likely to experience both transitions and breakdowns. (chap. 9: 56–57, draft version)

In addition, Coppedge identifies a number of factors whose effects are robust, though difficult to interpret. This includes "the core vs. periphery distinction, the proportion of democratic neighbors, the distinction between capitalist and communist economies, the number of past regime transitions, a

British colonial legacy, regional dummies, and the most opaque of all, the lagged dependent variable."

Coppedge concludes that crossnational analysis of democratization is still in its infancy. This might be interpreted as a pessimistic commentary for a field that has been running regressions for five decades, or as an optimistic view of things still to be discovered, depending upon one's point of view. In any case, even if we take as given only the conclusions warranted by extant research—as summarized by Coppedge's painstaking and comprehensive review—it is still something of an achievement, given the difficulties of the subject matter. More to the point, it is a lot more than we have learned from other approaches. To this extent, and with all the usual caveats, messy data analysis is vindicated.

Messy Data at Work: Democracy as an Independent Variable

Let us now consider a slightly different question, where democracy lies on the right side of a causal model. Seawright (2010) claims that regression has made no contribution to the question of democracy's relationship to growth. He says (echoing the common wisdom) that no hypothesis is robust to all plausible estimators and specifications and all "positive" (statistically significant) findings are subject to potential confounders.

My own view is that democracy shows a strongly positive relationship to growth if measured in a non-dichotomous and historical fashion (Gerring, Bond, Barndt, and Moreno 2005; Persson and Tabellini 2009). However, let us lay that argument aside, as it is not addressed by Seawright's otherwise admirably comprehensive review. I agree with Seawright's (and others') conclusions, so far as they go. As conventionally operationalized, there is no robust and plausible relationship between democracy and growth. Is this not useful knowledge? Have we not updated our priors? To put the point somewhat differently, do we know anything more about the theoretical question of interest after having looked at the crossnational data? I submit that we have learned quite a lot.

I want to bang on this drum a little while longer. Suppose a policymaker is interested to know whether there might be a relationship between democracy and growth. He comes to you, the resident comparativist, for recommendations of studies that s/he should take a look at. Seawright's advice would seem to be the following: "Read only case studies, avoid all large-N crossnational studies, and hope that someone, someday, figures out a way to study this in an experimental or quasi-experimental fashion." I doubt that this is sage advice.

Note that if there were a reasonably strong (and therefore practically and theoretically relevant) causal relationship between democracy and growth one would expect it to appear in crossnational empirical tests and to be at least somewhat stable across various (plausible) robustness tests. The fact that it does *not* (when democracy is operationalized in the conventional fashion) is informative. One is much less inclined to believe the thesis. Indeed, failing to disprove a null hypothesis is often just as useful as proving a hypothesis, even though it is difficult to do so in a definitive fashion (because proving a null hypothesis means, in effect, disproving any possible relationship between X and Y, an argument that must take into account all possible forms that a relationship might take).

Now, let us approach the question from another angle. The critique of large-N observational data would be trenchant *if* a viable alternative were available. However, Seawright equivocates on this point. He suggests scaling down our theoretical ambitions to examine causal mechanisms—factors predicted to lie in between democracy and growth. Yet, the only empirical example of this sort of work that is cited (Baum and Lake 2003) is also a regression analysis—a fact that Seawright notes, disapprovingly. Moreover, Seawright lays out a number of methodological difficulties that such mechanismic studies are likely to face. One is left to wonder whether or not there is a viable alternative to the crossnational regression in this particular instance.

It so happens that I have participated in a case-study endeavor to show causal mechanisms lying within the democracy (stock) and growth relationship (Gerring, Kingstone, Lange, and Sinha 2011). We found it a useful exercise. But it was certainly not without its methodological difficulties, and it certainly did not meet the test of inferential validity that case study researchers aspire to. So, again, I found that an available method added to the sum total of human knowledge on our subject but lay very far from the methodological standards currently being advocated.

Robustness Tests

Critics of messy observational data point out that results are generally unstable when slight changes are made in the measurement of key variables, the specification of a model, or the chosen estimator (including corrections for autocorrelation and the like). Robustness tests show few robust results, and virtually no stable results (where the coefficient on a key variable of interest remains stable). Under the circumstances, it seems clear that coefficients and standard errors are not to be taken literally. This is especially true for the cursed format of the crossnational regression, for all the reasons we have discussed (Kittel 2006; Rodrik 2005; Seawright 2010; Summers 1991; Treisman 2007). Because we don't know which (if any) operationalization, specification, and estimator correctly models the data generation process (DGP), we are at sea.

Critics are right to be skeptical of studies that show only one empirical test for an argument. Appeals to "theory" are generally not very convincing. (Note that if the theory is strong there is little point in testing; we already know what's out there. If the theory is weak, we are not strengthening our faith in assumptions by appealing to it.) Usually, there are a variety of plausible ways to model the DGP. However, the researcher typically only shows a subset of these possibilities (one suspects that alternative models have been suppressed, by virtue of their non-corroborating results). This is a serious problem. Consequently, consumers are not in a good position to judge the veracity of an argument based on messy observational data, unless they have played with the data themselves.

Yet, this final clause suggests something important. Impartial examinations of the data generally reveal that some relationships are more robust than others. Of course, this could be the product of persistent X/Y endogeneity, unmeasured confounders, a biased sample, and so forth. Robustness tests will never offer the explicit demonstration of causality that is provided by experimental or superb case-based analysis. And they will never provide a precise estimate of X's impact on Y. The purpose, rather, is to test whether a very generally stated hypothesis—conceptualized vaguely as "positive" or "negative"—is likely to be true or not.

With this modest objective in mind robustness tests are more than just window dressing. Note that although few results withstand all possible and plausible robustness tests, some are more robust than others. These results deserve to be taken very seriously—provided they are plausible (knowing what we know about the world). As an example, one might return to the relationship between development and democratic consolidation. Here is a result that seems unlikely to go away, no matter how much researchers torture the data. Likewise, weaker results—those that are fragile in the face of robustness tests, such as the relationship between development and democratization—deserve to be treated with greater skepticism. This does not mean they should be dismissed; it means simply that the estimate must be surrounded by very large confidence intervals.

This is more—much more—than nothing. Note that all causal inference is based on assumptions. (Even experiments rest on assumptions, though they are much fewer and usually less problematic.) In the words of Donald Rubin,

> Causal inference is impossible without making assumptions, and they are the strands that link statistics to science. It is the scientific quality of those assumptions, not their existence, that is critical. There is always a trade-off between assumptions and data—both bring information. With better data, fewer assumptions are needed. But in the causal inference setting, assumptions are always needed, and it is imperative that they be explicated and justified. One reason for providing this detail is so that readers can understand the basis of conclusions. A related reason is that such understanding should lead to scrutiny of the assumptions, investigation of them, and, ideally, improvements. Sadly, this stating of assumptions is typically absent in many analyses purporting to be causal and replaced by a statement of what computer programs were run.[8]

Robustness tests are tests of assumptions, usually understood by reference to a benchmark model (which the researcher considers to represent the most plausible rendering of the DGP). The purpose of each test is to verify the main finding under slightly different assumptions. If the finding holds, it is considered robust.[9]

Conclusions

This short opinion-piece has discussed only one type of large-N observational data inference, where countries serve as units of analysis. Evidently, I have not offered anything like a comprehensive review of this gargantuan subject. Yet, if messy data offers a viable strategy in crossnational analysis, which might be considered the worst-case scenario for causal inference, it ought to be viable in other settings.

Thus, I submit that large-N observational research where the treatment bears no resemblance to randomized assignment and where there is no opportunity for process-tracing or strong causal-process observations has made a fundamental contribution to some areas of political science research.[10] Plausibly, it may continue to do so. But it will do so in a productive fashion only if its achievements are recognized and if reasonable standards for publication are accepted by the discipline. If not, I fear that broad questions like the relationship of development to democracy will go unanswered—or will be answered only by journalists and amateur prognosticators. And if this occurs, the cause of truth will be set back immeasurably. Perhaps social science will be purer, more scientific (from a certain angle). But it will be less consequential. And society will not be well-served.

Methodologists who are depressed about the uncertainty of knowledge in political science would do well to contemplate the field of archaeology. Here, researchers are in a much worse position vis-à-vis the things they want to find out (presumably, exactly the same sort of things that social scientists wish to find out about contemporary society). Their subjects are long-departed, leaving few remains. All is conjecture. Yet, this does not stop archaeologists from drawing conclusions—however tentative—about their subject. And these conclusions are generally regarded as an advance over popular myths about the past (though they may incorporate myths as a form of evidence).[11]

Likewise, the difficulties presented by observational data should not prevent political scientists from drawing conclusions about causality—with the critical caveat that those conclusions be framed with appropriate confidence intervals. As Christopher Achen (1982: 77–78) has observed, *all* evidence is descriptive, for causation is an inferential form of knowledge. Even experiments don't speak for themselves.

It is true, of course, that drawing inferences based on weak data is perilous. High uncertainty means that conclusions will often be wrong—less than 50% of the time, one would hope, but a lot of the time nonetheless. One must ponder carefully the ramifications of giving bad policy advice based on messy data analysis. Bad policies may be pursued, lives may be lost, and the credibility of social science may suffer accordingly. This fear prompts some researchers to identify scientific virtue with reticence. Professors should speak only when they are pretty certain of an answer. Otherwise, they should keep mum.

Sometimes it is important to resist the temptation to prognosticate, i.e., to insist that we do not know the answer to a question, however important that question might be. The flip side of the coin is that by refusing to engage questions of public concern, members of the academy withdraw from the debate. The questions do not go away, nor do the—quite possibly faulty—answers. Likewise, the policies persist, based on those faulty answers.

Suppose that political scientists, as a group, decide to take a principled stand on the question of democratization by resolutely insisting on our ignorance. That is, we do not know why countries democratize, much less how to promote this process. We are prepared to tell you why extant studies are faulty, or at least highly uncertain. But we are not prepared to say which policies the U.S., or any other country, might pursue in order to foster political freedoms in the world because there is no secure causal knowledge on this question. Is this a responsible position to take?

Moral philosophers sometimes distinguish between "negative" and "positive" duties. The first is our duty is to avoid inflicting harm on others. The second is our duty to do good (e.g., to alleviate suffering where we can do so). If one subscribes to this distinction, the harm caused by doing harm is much greater than the corresponding harm caused by not doing good. From this perspective, the idea of a social-scientific Hippocratic oath—i.e., pronounce only on issues where there are high levels of certainty—is appealing.

Yet, in the policy sphere the distinction between negative and positive duties is difficult to sustain. It is not clear, for example, that withdrawing assistance to democratization processes around the world would be a virtuous act. It is not even clear what such a withdrawal would consist of. Countries must have foreign policies, unless they are to withdraw entirely from the world, and any foreign policy will presumably have some effect on the pace and progress of democratization. It is difficult to construct a neutral foreign policy because in not taking a position on democratization a country still has a causal effect on that outcome. This means that for a policy science such as political science, it would be difficult to define and maintain a principled stance of "doing no harm." Likewise, we are not a sect of priests whose moral purity is more important than the well being of society at large, so a deontological (person-centered) morality is in the end difficult to justify.

Of course, there is an argument for reticence if uncertainty is too complicated a notion to communicate to the general public or to policymakers. These consumers of political science look to academics for certainty and are illprepared, either psychologically or professionally, for confidence intervals. No matter how carefully advice is tendered, no matter how many caveats are attached, the message will be transmitted in the popular media as "Professor A says X causes Y, and we should follow policy Z."

And yet, I do not see a way around it. We cannot, in good conscience, avoid communicating the knowledge that we possess about pressing issues of the day if they touch upon our area of expertise, even if our knowledge is based on messy data, and hence highly uncertain, and even if there is a risk that it might be misinterpreted and thereby lead to bad policies.

Let us return to the main argument briefly, so as to recapitulate and to clear up any misconceptions. Experiments, quasi-experiments, and slam-dunk case-based evidence are strongly preferred wherever viable, as they generally provide superior internal validity. It is my hope that political scientists will find ingenious ways to widen the applicability of these methods to questions that animate the field.

But where they are (a) impossible to implement, (b) of questionable internal or external validity, or (c) irrelevant for building general theory or addressing questions of public concern, these Grade A methods must be supplemented or replaced by other methods, crude those they may be. This large class of Grade B methods may be categorized broadly as large-N observational research, aka *messy data* analysis.

Thus, I propose one cheer—perhaps even two cheers—for this much-maligned but hardy breed of causal strategies. Messy data is often the least-bad of all feasible alternatives. And for this, it should be honored.

The problem is that political scientists have generally assumed that there is, or ought to be, one standard of causal inference with a very high level of certainty (say, 90% or 95%) applying to all work in the field. This is an unrealistic standard if we are to continue to pursue the panoply of diverse causal questions that have traditionally motivated the field—and that seem to have great policy and practical significance. Lower standards of certainty are required for some questions that are not amenable to Grade A methods.

Likewise, those who work with messy data need to muster the courage to state honestly and forthrightly the high level of uncertainty that usually accompanies their causal inferences. Do not simply recite the t statistic and p value. Messy data calls for grappling with research design issues that are not summarizable with asterisks.

Notes

[1] Many of the points in this short essay are dealt with in a more detailed (and more nuanced) manner in Gerring (2011b). My thanks to Taylor Boas, Jake Bowers, Michael Coppedge, Adam Glynn, Evan Lieberman, Jay Seawright, and David Waldner for their feedback on earlier versions of this polemic. Needless to say, they are not to be implicated in the argument.

[2] The subordination of substantive arguments to methodological considerations is discussed in Mead (2010), Shapiro (2005), and Smith (2003).

[3] I have written a book on case study methods (Gerring 2007) and am a strong proponent of experimental and quasi-experimental methods (Gerring 2011a).

[4] These and other issues are addressed in Deaton (2010), Harrington (2000), Heckman (2010), Humphreys and Weinstein (2009), Leamer (2010), Lieberson and Horwich (2008), Scriven (2008).

[5] Difficulties of case-based knowledge are discussed in Gerring (2007), King and Powell (2008), Lieberson (1991), Sekhon (2004).

[6] Various studies comparing analyses of the same phenomenon with experimental and non-experimental data show significant disparities in results, offering direct evidence that observational research is flawed (e.g., Benson and Hartz 2000; Friedlander and Robins 1995; Glazerman, Levy, and Myers 2003; LaLonde 1986). Cook, Shaddish, and Wong (2008) offer a more optimistic appraisal.

[7] Achen (2005), Berk (2004), Brady and Collier (2010), Clogg and Haritou (1997), Freedman (1991, 1997, 2010), Gerber, Green, and Kaplan (2004), Gigerenzer (2004), Heckman (2000; 2008: 3), Kittel (2006), Longford (2005), Rodrik (2005), Schrodt (2006), Seawright (2010), Summers (1991).

[8] Rubin (2005: 324). See also Day and Kincaid (1994), Freedman (2008), Garfinkel (1981), Heckman (2008: 3), Robins and Wasserman (1999), Rosenbaum (1999, 2005).

[9] Bartels (1997), Imbens (2003), Leamer (1983), Levine and Renelt (1992), Montgomery and Nyhan (2010), Rosenbaum (2002), Rosen-

baum and Rubin (1983), Sala-i-Martin (1997), Sims (1988), Western (1995), Young (2009). Note that my understanding of a "robustness" test incorporates both alternative specifications (the usual focus of extreme bounds analysis and sensitivity testing) and alternative estimators (as discussed in most econometrics texts), as well as various operationalizations for key variables (*X*, *Y*, *C*). *Any* important assumption underlying an empirical model that can be tested by altering some element of that model ought to be included in a set of robustness tests.

[10] Here, my conclusions echo the traditional wisdom, as articulated by Laitin (2002) and Fearon and Laitin (2008).

[11] Consider if an archaeologist were to discover a cross-cultural dataset of the sort that comparativists wrestle with—a first-century World Development Indicators. Presumably, he or she would be thrilled, and our knowledge of the first-century AD would be dramatically advanced.

References

Achen, Christopher H. 1982. *Interpreting and Using Regression*. Beverly Hills: Sage.

Achen, Christopher H. 2005. "Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong." *Conflict Management and Peace Science* 22, 1–13.

Angrist, Joshua D. and Jorn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24:2 (Spring), 3–30.

Bartels, Larry M. 1997. "Specification Uncertainty and Model Averaging." *American Journal of Political Science* 41 (April), 641–674.

Baum, Matthew A. and David A. Lake. 2003. "The Political Economy of Growth: Democracy and Human Capital." *American Journal of Political Science* 47:2, 333–347.

Bennett, Andrew. 2010. "Process Tracing and Causal Inference." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 2nd ed. Henry E. Brady and David Collier, eds. (Lanham, MD: Rowman & Littlefield), 207–220.

Bennett, Andrew and Colin Elman. 2006. "Qualitative Research: Recent Developments in Case Study Methods." *Annual Review of Political Science* 9, 455–476.

Benson, Kjell and Arthur Jr. Hartz. 2000. "A Comparison of Observational Studies and Randomized Controlled Trials." *New England Journal of Medicine* 342:25, 1878–1886.

Berk, Richard A. 2004. *Regression Analysis: A Constructive Critique*. Thousand Oaks, CA: Sage.

Boix, Carles and Susan C. Stokes. 2003. "Endogenous Democratization." *World Politics* 55:4 (July), 517–549.

Brady, Henry E. and David Collier, eds. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 2nd ed. Lanham: Rowman & Littlefield.

Casper, Gretchen and Claudiu Tufis. 2003. "Correlation Versus Interchangeability: The Limited Robustness of Empirical Findings on Democracy Using Highly Correlated Data Sets." *Political Analysis* 11, 196–203.

Clogg, Clifford C. and Adamantios Haritou. 1997. "The Regression Method of Causal Inference and a Dilemma Confronting this Method." In *Causality in Crisis?: Statistical Methods and the Search for Causal Knowledge in the Social Sciences*. Vaughn R. McKim and Stephen P. Turner, eds. (Notre Dame, IN: Notre Dame Press), 83–112.

Cook, Thomas D., Michael Scriven, Chris L. S. Coryn, and Stephanie D. H. Evergreen. 2010. "Contemporary Thinking about Causation in Evaluation: A Dialogue With Tom Cook and Michael Scriven." *American Journal of Evaluation* 31, 105–117.

Cook, Thomas D., William R. Shaddish, and Vivian C. Wong. 2008. "Three Conditions under which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons." *Journal of Policy Analysis and Management* 27, 724–750.

Coppedge, Michael. Forthcoming. *Approaching Democracy: Research Methods in Comparative Politics*. Cambridge: Cambridge University Press.

Day, Timothy and Harold Kincaid. 1994. "Putting Inference to the Best Explanation in its Place." *Synthese* 98, 271–295.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development. *Journal of Economic Literature* 48:2 (June), 424–455.

Dunning, Thad. 2008a. "Improving Causal Inference: Strengths and Limitations of Quasi-Experiments." *Political Research Quarterly* 61:2, 282–293.

Dunning, Thad. 2008b. "Natural and Field Experiments: The Role of Qualitative Methods." *Qualitative and Multi-Method Research* 6:2 (Fall), 17–23.

Epstein, David L., Robert Bates, Jack Goldstone, Ida Kristensen, and Sharyn O'Halloran. 2006. "Democratic Transitions." *American Journal of Political Science* 50:3 (July), 551–569.

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

Freedman, David A. 1991. "Statistical Models and Shoe Leather." *Sociological Methodology* 21, 291–313.

Freedman, David A. 1997. "From Association to Causation via Regression." In *Causality in Crisis?: Statistical Methods and the Search for Causal Knowledge in the Social Sciences*. Vaughn R. McKim and Stephen P. Turner, eds. (Notre Dame, IN: Notre Dame Press), 113–162.

Freedman, David A. 2008. "On Types of Scientific Inquiry: The Role of Qualitative Reasoning." In *The Oxford Handbook of Political Methodology.* Janet Box-Steffensmeier, Henry Brady, and David Collier, eds. (Oxford: Oxford University Press), 300–318.

Freedman, David A. 2010. *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*, David Collier, Jasjeet Sekhon, and Philip B. Stark, eds. Cambridge: Cambridge University Press.

Friedlander, Daniel and Philip K. Robins. 1995. "Evaluating Program Evaluations: New Evidence on Commonly Used Non-Experimental Methods." *American Economic Review* 85:4 (September), 923–937.

Garfinkel, Alan. 1981. *Forms of Explanation: Rethinking the Questions of Social Theory*. New Haven: Yale University Press.

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

Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2004. "The Illusion of Learning from Observational Research." In *Problems and Methods in the Study of Politics*. Ian Shapiro, Rogers M. Smith, and Tarek E. Masoud, eds. (Cambridge: Cambridge University Press), 251–273.

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

Gerring, John. 2011a. "How Good is Good Enough? A Multidimensional, Best-Possible Standard for Research Design." *Political Research Quarterly* (forthcoming).

Gerring, John. 2011b. *Social Science Methodology: A Unified Framework*. Cambridge: Cambridge University Press.

Gerring, John, Peter Kingstone, Matthew Lange, and Aseema Sinha. 2011. "Democracy, History, and Economic Performance: A Case-Study Approach." *World Development* (forthcoming).

Gerring, John, Philip Bond, William Barndt, and Carola Moreno. 2005. "Democracy and Growth: A Historical Perspective." *World Politics* 57:3 (April), 323–364.

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

Gerring, John and Strom Thacker. 2008. *A Centripetal Theory of Democratic Governance*. Cambridge: Cambridge University Press.

Gigerenzer, Gerd. 2004. "Mindless Statistics." *Journal of Socio-Economics* 33, 587–606.

Glazerman, Steven, Dan M. Levy, and David Myers. 2003. "Non-Experimental versus Experimental Estimates of Earnings Impacts." *The Annals of the American Academy of Political and Social Science* 589:1, 63–93.

Harrington, David P. 2000. "The Randomized Clinical Trial." *Journal of the American Statistical Association* 95:449, 312–315.

Heckman, James J. 2000. "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective." *Quarterly Journal of Economics* 115:45–97.

Heckman, James J. 2008. "Econometric Causality." *International Statistical Review* 76:1, 1–27.

Heckman, James J. 2010. "Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy." *Journal of Economic Literature* 48:2 (June), 356–398.

Humphreys, Macartan and Jeremy Weinstein. 2009. "Field Experiments and the Political Economy of Development." *Annual Review of Political Science* 12 (June), 367–378.

Imbens, Guido W. 2003. "Sensitivity to Exogeneity Assumptions in Program Evaluation." American Economic Review 93:2, 126–132.

Imbens, Guido W. and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47:1, 5–86.

King, Gary; Eleanor Neff Powell. 2008. "How Not to Lie Without Statistics." Unpublished manuscript, Harvard University.

Kittel, Bernhard. 2006. "A Crazy Methodology?: On the Limits of Macroquantitative Social Science Research." *International Sociology* 21, 647–677.

Laitin, David D. 2002. "Comparative Politics: The State of the Subdiscipline." In *Political Science: State of the Discipline*. Ira Katznelson and Helen Milner, eds. (New York: Norton).

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76:4 (September), 604–620.

Leamer, Edward E. 1983. "Let's Take the Con out of Econometrics." *American Economic Review* 73:1, 31–44.

Leamer, Edward E. 2010. "Tantalus on the Road to Asymptopia." *Journal of Economic Perspectives* 24:2 (Spring), 31–46.

Levine, Ross and David Renelt. 1992. "A Sensitivity Analysis of Cross-Country Growth Regressions." *American Economic Review* 82:4 (September), 942–963.

Lieberson, Stanley. 1991. "Small N's and Big Conclusions: An Examination of the Reasoning in Comparative Studies Based on a Small Number of Cases." *Social Forces* 70:2 (December), 307–320.

Lieberson, Stanley and Joel Horwich. 2008. "Implication Analysis: A Pragmatic Proposal for Linking Theory and Data in the Social Sciences." *Sociological Methodology* 38:1 (December), 1–50.

Lipset, Seymour Martin. 1959. "Some Social Requisites of Democracy: Economic Development and Political Development." *American Political Science Review* 53 (March), 69–105.

Longford, Nicholas T. 2005. "Editorial: Model Selection and Efficiency–Is 'Which Model…' the Right Question?" *Journal of the Royal Statistical Society,* Ser. A 168, 469–472.

Mahoney, James. 2000. "Strategies of Causal Inference in Small-N Analysis." *Sociological Methods and Research* 28:4 (May), 387–424.

Mahoney, James. 2010. "Process Tracing: A Method for Evaluating Hypotheses in Particular Cases." In process.

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

Mead, Lawrence M. 2010. "Scholasticism in Political Science." *Perspectives on Politics* 8:2 (June), 453–464.

Montgomery, Jacob M. and Brendan Nyhan. 2010. "Bayesian Model Averaging: Theoretical Developments and Practical Applications." *Political Analysis* 18:2 (Spring), 245–270.

Morgan, Stephen L. and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

Paluck, Elizabeth Levy. 2010. "The Promising Integration of Qualitative Methods and Field Experiments." *Annals of the American Academy of Political and Social Science* 628:1, 59–71.

Persson, Torsten and Guido Tabellini. 2009. "Democratic Capital: The Nexus of Political and Economic Change." *American Eco nomic Journal: Macroeconomics* 1:2, 88–126.

Przeworski, Adam and Fernando Limongi. 1997. "Modernization: Theories and Facts." *World Politics* 49 (January), 155–183.

Robins, James M. and L. Wasserman. 1999. "On the Impossibility of Inferring Causation from Association without Background Knowledge." In *Computation, Causation, and Discovery*. Clark Glymour and G. Cooper, eds. (Menlo Park, CA, Cambridge, MA: AAAI Press/The MIT Press), 305–321.

Rodrik, Dani. 2005. "Why We Learn Nothing from Regressing Economic Growth on Policies." Unpublished manuscript, Harvard University.

Rosenbaum, Paul R. 1999. "Choice as an Alternative to Control in Observational Studies." *Statistical Science* 14:3 (August), 259–278.

Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer.

Rosenbaum, Paul R. 2005. "Reasons for Effects." *Chance* 18, 5–10.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. New York: Springer.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome." *Journal of the Royal Statistical Society*: Series B 45, 212–218.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100:469 (March), 322–331.

Sala-i-Martin, Xavier X. 1997. "I Just Ran Two Million Regressions." *American Economic Review* 87:2, 178–183.

Schrodt, Philip A. 2006. "Beyond the Linear Frequentist Orthodoxy." *Political Analysis* 14:3 (Summer), 335–339.

Scriven, Michael. 2008. "A Summative Evaluation of RCT Methodology: & An Alternative Approach to Causal Research." *Journal of Multidisciplinary Evaluation* 5, 11–24.

Seawright, Jason. 2010. "Regression-Based Inference: A Case Study in Failed Causal Assessment." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 2nd ed. Henry E. Brady and David Collier, eds. (Lanham, MD: Rowman & Littlefield), 247–271.

Sekhon, Jasjeet S. 2004. "Quality Meets Quantity: Case Studies, Conditional Probability and Counterfactuals." *Perspectives on Politics* 2:2 (June), 281–293.

Shapiro, Ian. 2005. *The Flight from Reality in the Human Sciences*. Princeton: Princeton University Press.

Sims, Christopher A. 1988. "Uncertainty across Models." *American Economic Review* 78:2, 163–167.

Smith, Rogers M. 2003. "Reconnecting Political Theory to Empirical Inquiry, or A Return to the Cave?" In *The Evolution of Political Knowledge: Theory and Inquiry in American Politics*. Edward D. Mansfield and Richard Sisson, eds. (Ohio State University Press), 60–88.
Summers, Lawrence H. 1991. "The Scientific Illusion in Empirical Macroeconomics." *The Scandinavian Journal of Economics* 93:2 (June), 129–148.
Treisman, Daniel. 2007. "What Have We Learned About the Causes of Corruption from Ten Years of Cross-National Empirical Research?" *Annual Review of Political Science* 10, 211–244.
Western, Bruce. 1995. "Concepts and Suggestions for Robust Regression Analysis." *American Journal of Political Science* 39:3, 786–817.
Young, Cristobal. 2009. "Model Uncertainty in Sociological Research: An Application to Religion and Economic Growth." *American Sociological Review* 74:3 (June), 380–397.

Regression's Weaknesses and Strengths: A Reply to Gerring

Jason Seawright
Northwestern University
j-seawright@northwestern.edu

Assumptions are the rule, not the exception, in both descriptive and causal inference in the social sciences. This fact has long been used as a defense of the specific families of assumptions used to make causal inferences on the basis of regression-type models (Freedman 2004: 195). Yet the defense is weak. Inferences differ in terms of the strength, complexity, plausibility, and testability of the assumptions they require. On all of these fronts, regression-type analysis of observational data often performs so poorly that it is difficult to give the results a persuasive causal interpretation. In what follows, I will make this argument by showing how hard it can be to assign causal interpretations to regression models that show either unstable or stable results across the range of models that the discipline considers plausible, as well as the challenges involved with drawing causal conclusions from either the unconditional or the conditional analysis of quantitative observational data. For these reasons, I disagree with Gerring's argument that the regression analysis of messy data is a good default option for social scientists; instead, it is a weak default, and one far more suited to describing initial facts, discovering puzzles, and characterizing patterns than to causal inference. I then argue that the importance of difficult-to-research questions is a weak defense of the status quo, and I conclude by briefly sketching the valuable but carefully delimited role that regression should play in our research.

The Trouble with Unstable Results

Gerring helpfully discusses the range of studies showing that cross-national regressions, in particular, routinely show unstable patterns of results across plausible specifications regarding the most important relationships. While this difficulty is by now well known, two points are nonetheless worth emphasizing.

First, instability invites the tempting but unjustified inference that the true relationship of interest is weak. Gerring's commentary provides an example of this interpretation:

> …if there were a reasonably strong (and therefore practically and theoretically relevant) causal relationship between democracy and growth, one would expect it to appear in cross-national empirical tests and to be at least somewhat stable across various (plausible) robustness tests.

While this expectation makes intuitive sense, it is nonetheless unreliable. With respect to the relationship between democracy and growth, the range of models which scholars have regarded as potentially credible produce results ranging from substantively and statistically significant negative effects to similarly significant positive effects (Seawright 2010). Within this range of results, there is no special reason to believe that the truth lies in the middle. It might instead be the case that the largest negative estimate produced to date in fact reflects the causal truth; or, perhaps, a very positive estimate corresponds with the correct model. If one model captures the structure of the data-generating process, or one estimate is correct, then all the others are incorrect and irrelevant. Instability across accepted specifications thus should not be seen as providing evidence that the true relationship is weak. Such instability only provides evidence that our consensus about how to write down regression models is weak.

Second, the set of models which are currently regarded by the scholarly community as plausible and which can be estimated using existing data comprise a quite unusual sample from the population of possible models for a given relationship. The distinctiveness of this sample is in part healthy: presumably, knowledge of cases and substance rules out a range of specifications that are statistically possible but in some sense foolish. Thus, we rarely estimate models in which the positions of planets, for example, are taken to predict economic performance or political institutions.

However, the extreme winnowing that produces our collection of plausible models also includes less salutary forms of selection. Some of these reflect ossified convention. For historical reasons, additive models which are linear in both the parameters and the independent variables, and which feature an independent, additive, approximately normally distributed error term, are our collective default for the analysis of continuous dependent variables (Stigler 1990).

Our sample of plausible models is further constrained by the set of available indicators. While scholars sometimes create new indicators to capture novel hypotheses that lie at the center of their explanatory agendas, they rarely go to the same amount of work to measure potential confounding variables. Instead, the control variables in our plausible models are generally some subset of the current collective stock of data. Some subset of that stock of variables becomes defined as the core control variables, without which a model is inherently implausible; this process of definition, I think, reflects in part an accumulation of past findings and arguments and in part a process of social consensus. But, regardless of the mix of these two

Table 1: Two Simple Models of Democracy and Development

	Model 1 Estimates (P Values)	Model 2 Estimates (P Values)
Intercept	-46.6 (< 0.01)	-13.3 (< 0.01)
Logged Per Capita GDP	5.6 (< 0.01)	
GDP Rank		0.2 (< 0.01)
GDP Residual		8.5 (0.59)
R^2	0.16	0.16
N	128	128

components, such norms certainly further constrain the range of plausible models.

Last but obviously not least, the set of plausible models is limited by our contemporary repertoire of concepts and indicators. Scholars working before the development of systematic conceptualizations of, and survey measures for, the ideas of retrospective economic evaluations or strategic voting pressures would have an obvious excuse for failing to include those variables in their models of vote choice—but, good excuse or no, the models remain misspecified. The variables that will be discovered or invented over the next century quite evidently cannot be included in today's models, even though they may be necessary for causal inference.

The net result of these and the other constraints listed above is that the range of results found in today's set of published plausible models cannot even be taken as providing logical upper and lower bounds for the true causal effect. Some scholars might be tempted to argue that, while it is possible for the true causal effect to fall outside the range of contemporary statistical estimates, it is unlikely. This argument is not an implication of regression theory and is not even universally supported by tests that compare observational regression estimates with experimental benchmarks.

To sum up, unstable regression results on observational data simply do not teach us about the direction, magnitude, practical relevance, or theoretical importance of the underlying causal relations. We may tend to believe less in causal effects that cannot be consistently demonstrated using messy data, but such disbelief is not well grounded and should probably be resisted. That is to say, "we do not know" does not imply "it is not so."

The Trouble with Stable Results

While researchers are likely to be broadly familiar with the argument that instability in statistical results demonstrates significant uncertainty in our knowledge about causal relations, it is much less widely discussed but nonetheless true that stable statistical results can also be compatible with uncertainty in causal knowledge. To see this point, let us consider one of the most stable findings in comparative politics: that GDP per capita is significantly associated with democracy. Some scholars make much of the distinction between predicting transitions to democracy and predicting democratic breakdown; for the moment, I will disregard this distinction, for reasons to be discussed below.

It is true that democracy and development are strongly related, for a wide variety of measures of democracy, a range of operationalizations of development, and a broad class of statistical models. Yet it nonetheless remains uncertain whether development in fact causes democracy.

While most models reproduce the widely accepted result that development increases the probability of democracy, some do not. In particular, Acemoglu, Johnson, Robinson, and Yared (2008) show that including country fixed effects in an analysis almost completely removes this relationship. A convergent finding can be shown using two simple cross-sectional regression models, shown in Table 1.

Model 1 in the table shows a bivariate regression predicting democracy on the basis of per capita GDP (logged, as is often the case in this literature, to deal with the skewness of the variable). The analysis is carried out using 1985 data, although the year is not important and similar findings can be produced for a wide range of years. Here we find the standard result: wealth strongly and positively predicts democracy.

Model 2 refines this finding, partitioning the democracy variable into two orthogonal components. The first component is a country's rank in the global 1985 distribution of wealth, while the second is that country's residual in a regression predicting logged GDP using GDP rank as an explanation. In other words, the rank variable shows countries' relative order in the global economic hierarchy but not the fine detail of their level of wealth, while the residual shows the component of the level of wealth that cannot be predicted by rank order. The two components can be linearly combined to recover the original GDP variable.

This model allows us to ask which aspect of wealth—relative position in the world hierarchy or absolute resources—is in fact correlated with level of democracy. The question is crucial given that most theorizing about this relationship, from the days of modernization theory to the present, has treated the absolute level of economic resources as the cause of interest. Hence, if relative rather than absolute wealth is key, most theoretical work on this central issue has been misdirected in important ways.

If wealth per se is a cause of democracy, then both components in this partition of GDP should be associated with level of democracy. Moving up the rank order should help because it generally involves a gain in level of wealth, but increases in level of wealth that are not quite large enough to produce a change in rank order should also help. But in fact, as Model 2 shows, virtually all of the predictive power of the GDP variable is captured by the rank component; the coefficient for the residual component is not even close to achieving statistical significance.

The distinction between rank order and level of GDP is crucial because, while levels of GDP change substantially over time, rank orders do not. Between 1960 and 1990, for example, the correlation in GDP rank orders is 0.88. For this reason, the 1990 GDP rank order is almost as good a predictor of a country's 1960 level of democracy as is that country's 1960 level of GDP. In my judgment, these findings are consistent with the hypothesis that both long-term development trajectories and long-term regime trajectories are caused by decisions or institutional patterns at critical junctures well before the 20th century, an idea that is supported by much more robust case-study research (e.g., Mahoney 2010).

To the extent that these findings imply path dependence, most panel analyses of wealth and regime type are statistically problematic because they omit the critical historical events that set countries on one path or another (whatever those might be). Furthermore, findings relating wealth and democratic consolidation become causally ambiguous. Consolidation may be a consequence of a country's wealth, in absolute or relative terms, or alternatively may be a component of an institutional package that helps propel high levels of long-term economic performance.

As this example shows, stable results across specifications may simply mean that all of those specifications omit the same key confounder. These issues do not arise in the same way for experiments and other strong research designs. Because of their reliance on randomization or detailed case information, findings from these kinds of studies are, in comparison with the regression analysis of observational data, much less fragile to alternative model specifications.

The Trouble with Unconditional Inference

If neither stable results nor unstable results, with reference to the regression-type analysis of observational studies, can be logically taken to have clear implications for causal inference, the reader may begin to doubt that we could ever be confident that we have found causal knowledge with such a model. This doubt is, I think, healthy. To further nourish it, let us consider the same dilemma along the lines of another dichotomy, that between unconditional and conditional inference.

Unconditional inference involves a simple bivariate analysis of the relationship between the hypothesized cause and the outcome. For experiments, and many natural experiments, unconditional inference should be seen as the gold standard for causal inference (Freedman 2008, Dunning 2010). However, for observational studies, scholars have long been taught to regard unconditional inferences as entirely suspect.

The reason for this suspicion is the very real possibility of confounders, i.e., variables which belong in the model but are excluded from it and that distort the relationship between the independent and dependent variables. Experiments greatly reduce the problem of confounding by randomly assigning cases to treatment groups; successful natural experiments similarly abate confounding through a randomization, albeit one not controlled by the scholar. In regression-type observational studies, however, there is no randomization. Instead, there is every reason to believe that cases take on their observed scores on the independent variable because of complex social, economic, and political processes that may well also directly affect the outcome. Confounding, we anticipate, is therefore ubiquitous.

This does not necessarily mean that an unconditional inference is incorrect—there may by some miracle be no confounding in this particular analytic instance, or it might by extreme coincidence be the case that the various biases brought about by confounders happen to more or less cancel out. But it is nonetheless clear that confounding will usually be a problem, that we have no tools for identifying the handful of instances in which it might not be a problem, and therefore that unconditional analysis will rarely provide reliable causal inference.

The Trouble with Conditional Inference

The conclusion that unconditional inferences are unreliable for observational studies should not surprise. The following argument may be more surprising: conditional inferences, i.e., inferences that introduce control variables, are typically no more reliable than unconditional inferences in observational studies. I will develop this argument in two stages. First, there are some variables that, when added to an otherwise correct model as controls, distort causal inference. Second, even variables which appear as controls in the correct model may often, in imperfect real-world models, make causal inference worse, not better.

For decades, the literature on causal inference has warned against conditioning on post-treatment variables, i.e., variables that are caused by the independent variable (for useful recent discussions, see Rosenbaum 2002, King and Zeng 2006, and Morgan and Winship 2006). When a scholar conditions on a post-treatment variable, she inadvertently subtracts the effect of any causal pathway from the main independent variable, through that post-treatment variable, and to the outcome. If this subtraction is not taken account of analytically, the result will be a biased estimate of the overall causal effect of the independent variable of interest. It is somewhat less widely known that other categories of impermissible control variables exist; in particular, conditioning on "collider" variables can create new problems of confounding even when none existed before (Pearl 2000: 17–18, Cole et al. 2010). What happens if a variable is a confounder but also meets the criteria for post-treatment or collider status? If we are to follow the standard advice for achieving unbiased causal inference, such variables must be simultaneously included and excluded from our models. In a typical observational study, we lack the ability to identify with confidence which of the potential control variables belong in any of these categories, so it is hard to be sure whether we are making things better or worse by conditioning.

Suppose that, for some potential control variable, we are somehow entirely confident that the variable is a confounder and is neither a collider nor in any part post-treatment. Surely inference conditional on such a control variable is more reliable and closer to the causal truth than unconditional inference?

In fact, there is no certainty about this at all. The problem is that, while we may have identified a confounder, we are almost never certain that we have identified the last confounder. Thus, it remains probable that other omitted variables bias the inference even when conditioning on the known confounder. If the net bias produced by the set of remaining confounders is zero or points in the same direction as the bias connected with our known confounder, then the conditional inference will be superior to unconditional inference. However, the net remaining bias can point in the opposite direction, in which case conditional inference will often be worse than unconditional inference—a circumstance which, in some simulation studies, holds for 50% of potential control variables (Clarke 2005).

So, as every introductory methods text will tell us, in observational studies we cannot trust unconditional inferences. Yet barring unusual sorts of a priori causal knowledge, we also cannot trust that our conditional inferences will be closer to, rather than farther from, the truth than the unconditional inference. The value added by control variables can be obscure.

When the Stakes are High

The above arguments, together with the preference I and other scholars express against regression-type analysis and for in-depth case-based arguments, on the one hand, and experimental or natural-experimental designs, on the other, are sometimes seen, by Gerring and others, as an unhelpful form of "methodological perfectionism." Are there important questions that cannot be studied using these stronger designs? For such questions, does regression not offer a best-available approach?

I am unsure. It is true that there are many important substantive domains in political science that have been dominated by regression-type studies of observational data. Such designs have been the stock-in-trade of our discipline and the centerpiece of our methodological training for decades, so their dominance should not surprise us. Nor should we take the de facto dominance of these techniques as an indication that other approaches cannot work. Until relatively recently, experimental and natural experimental research had peripheral status in most political science subfields, and powerful voices made arguments denigrating the inferential value of case studies vis-à-vis regression.

What is certain is that political scientists have already, over the last decade, found ways of using these techniques to address questions at both macro and micro levels that have long been central to our discipline (e.g., Wantchekon 2003, Brady and McNulty 2004, Bhavnani 2009, Humphreys and Weinstein 2009, Corstange 2010, Dunning and Harrison 2010). It seems at least possible that an ongoing emphasis on the importance of research design and the relative inferential weakness of regression-type studies will motivate the hard work and ingenuity necessary to bring these techniques into full engagement with a broader range of issues.

In the end, however, I expect it to be the case that some important questions remain inaccessible for these methods. Of course, one might remark, there are always important questions that remain beyond the scope of all scientific methods; that a question matters does not guarantee that we can answer it well. And, for the most important questions, is it not true that the quality of our answers is unusually important?

Where Regression Shines

None of this should be taken as an attack on regression analysis, or a call for a ban on the technique. What regression does well, it does very well indeed—in fact, sometimes optimally well, as statistical theory can show. Trouble arises when we push regression too far outside its domain of competence.

What, then, are the strengths of regression? The technique is a powerful tool for the summary of complex cross-tabulations and scatter plots. Regression can sometimes make consistent but small descriptive relationships among variables more visible and can often dramatically aid comprehension of central themes in data by replacing an overwhelming mass of numbers or dots with a few key estimates (Berk 2003).

When scholars move beyond the tasks of summarizing and clarifying which constitute the key area of regression's strength in the social sciences, trouble can arise. It is important to understand that, in terms of inferential logic, regression is no different from the (potentially multidimensional) scatter plot or cross-tabulation that it summarizes. Matrix algebra simply does not convert observational data into causal laws (Humphreys and Freedman 1996, Freedman 1997, Freedman 1999).

When regression is used with careful attention to its real strengths, it can be a powerful tool, along with difference-in-means tests, graphs, cross-tabulations, and other such techniques, in the analyst's arsenal for descriptive and exploratory analysis. Furthermore, there are certainly moments when one or another piece of descriptive knowledge has strong causal implications; in such instances, regression may sometimes play a pivotal role in a causal argument.

However, we must accept that regression analysis of observational data will usually leave a great deal of causal uncertainty in its wake. Indeed, we cannot know in general whether regression analysis of messy data moves us closer to, or farther from, causal understanding. Our theorems cannot help us here; those which show regression-type analysis in a positive causal light do not apply to messy data, and those that do apply for messy data usually lack causal implications. So any defense of regression analysis of messy data must be pragmatic: the technique has to be shown to work for some important goal. That demonstration of efficacy has to be specific to the subject matter at hand and independent of the regression analysis itself. An example is regression work on forecasting election results;[1] here the regression analysis of messy data has been shown to have some practical (predictive, although not causal) value through out-of-sample prediction. However, we rarely produce such demonstrations of practical value for our regression research. As such, we simply cannot say whether we are better off with or without regression-type research in these contexts.

To the extent that our discipline values causal over descriptive knowledge, we must consider the possibility that regression-type studies of observational data have been significantly overvalued and overrepresented in our history over the last several decades. It may be time to shift some portion of resources such as funding, training, institutional support, and pages in our journals away from regression-type studies and toward case studies, experiments, natural experiments, and related approaches.

Note

[1] See, e.g, a symposium of ten articles on U.S. election forecasting in the October, 2008, issue of *PS: Political Science & Politics*.

References

Acemoglu, Daron, Simon Johnson, James A. Robinson, and Pierre Yared. 2008. "Income and Democracy." *American Economic Review* 98:3, 808–842.

Berk, Richard A. 2003. *Regression Analysis: A Constructive Critique*. Thousand Oaks, CA: Sage.

Bhavnani, Rikhil R. 2009. "Do Electoral Quotas Work After They are Withdrawn? Evidence from a Natural Experiment in India." *American Political Science Review* 103:1, 23–35.

Brady, Henry E. and John E. McNulty. 2004. "The Costs of Voting: Evidence from a Natural Experiment." Presented at the Annual Meeting of the Society for Political Methodology, Palo Alto, CA.

Clarke, Kevin A. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Journal of Conflict Resolution* 47:1, 72–93.

Cole, Stephen R., Robert W. Platt, Enrique F. Schisterman, Haitao Chu, Daniel Westreich, David Richardson, and Charles Poole. 2010. "Illustrating Bias Due to Conditioning on a Collider." *International Journal of Epidemiology* 39:2, 417–420.

Corstange, Daniel. 2010. "Vote Buying Under Competition and Monopsony: Evidence from a List Experiment in Lebanon." Presented at the Annual Conference of the American Political Science Association, Washington, D.C.

Dunning, Thad. 2010. "Design-Based Inference: Beyond the Pitfalls of Regression Analysis?" In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 2nd ed. Henry E. Brady and David Collier, eds. (Lanham, MD: Rowman & Littlefield), 273–311.

Dunning, Thad and Lauren Harrison. 2010. "Cross-Cutting Cleavages and Ethnic Voting: An Experimental Study of Cousinage in Mali." *American Political Science Review* 104:1, 21–39.

Freedman, David A. 1997. "From Association to Causation via Regression." *Advances in Applied Mathematics* 18:1, 59–110.

Freedman, David A. 1999. "From Association to Causation: Some Remarks on the History of Statistics." *Statistical Science* 14:3, 243–258.

Freedman, David A. 2004. *Statistical Models: Theory and Practice*. Cambridge: Cambridge University Press.

Freedman, David A. 2008. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics* 40:2, 180–193.

Humphreys, Paul and David A. Freedman. 1996. "The Grand Leap." *British Journal for the Philosophy of Science* 47:1, 113–123.

Humphreys, Macartan and Jeremy M. Weinstein. 2009. "Field Experiments and the Political Economy of Development." *Annual Review of Political Science* 12, 367–378.

King, Gary and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14:2, 131–159.

Mahoney, James. 2010. "After KKV: The New Methodology of Qualitative Research." *World Politics* 62:1 (January), 120–147.

Morgan, Stephen L. and Christopher Winship. 2006. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer.

Seawright, Jason. 2010. "Regression-Based Inference: A Case Study in Failed Causal Assessment." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 2nd ed. Henry E. Brady and David Collier, eds. (Lanham, MD: Rowman & Littlefield), 247–271.

Stigler, Stephen M. 1990. *The History of Statistics: The Measurement of Uncertainty Before 1990*. Cambridge: Belknap Press of Harvard University Press.

Wantchekon, Leonard. 2003. "Clientelism and Voting Behavior: Evidence from a Field Experiment in Benin." *World Politics* 55:3, 399–422.

Messy Data, Messy Conclusions: A Response to Gerring

Adam Glynn
Harvard University
aglynn@fas.harvard.edu

What can we learn from the analysis of a large-N observational data set (aka "messy" data)? Gerring argues that despite warnings from a number of critics, such an analysis may be deemed adequate as long as

> ... it allows us to update our priors, it beats the alternatives, and it presents a plausible uncertainty estimate.

This seems a rigorous benchmark. Depending on our definition of plausible, even randomized trials may fail this standard when issues of treatment compliance, treatment heterogeneity, experimenter effects, or interference between units muddy the interpretation of results. Of course, observational studies may fail this standard even when such issues are not a concern. Regression results from observational studies have two additional sources of uncertainty when compared to randomized studies.

The first is due to a lack of specificity about the manipulation of explanatory variables. In an experiment, the researcher controls the explanatory variable, and hence it is manipulable by definition. In observational studies, explanatory variables might hypothetically be manipulated in a number of ways, and these different manipulations can imply different effects. In the democratic consolidation example cited by Gerring, the "effect" of income (on the likelihood of a relapse into authoritarianism) depends on exactly how one intends to manipulate income. If income is increased by the discovery of oil, this may have different consequences than if income is increased by improved education.

The second source of uncertainty that is special to observational studies is the instability of regression results across different sets of pre-treatment conditioning variables. Randomized experiments do not exhibit this instability because in large samples, different treatment groups are guaranteed to have similar distributions for all possible pre-treatment variables.

However, despite these two complications, it is certainly the case that observational data allow us to "update our priors" regarding effects. To see this, it is helpful to more closely consider a single observation (country) from the democratic consolidation example. Suppose we only know that this country is poor, the "effect" of income is a comparison between the outcome for this country (relapse or consolidation) under their current state of income (poor), and the outcome we would have observed (relapse or consolidation) if we had somehow increased income (rich). There are three possible effects of increasing income from poor to rich: negative (consolidation to relapse), positive (relapse to consolidation), no effect (relapse to relapse or consolidation to consolidation). Our priors in this case might represent probabilities over these three effects.[1]

Now suppose that we observe the outcome variable for this country, and in fact this poor democracy relapsed into authoritarianism. For this country, we now know that increasing income (from poor to rich) would not have had a "negative effect" (from consolidation to relapse), because we observed relapse when poor. Increasing income could only have had either a positive effect (from relapse to consolidation) or no effect (from relapse to relapse). Therefore, if our prior beliefs put any positive probability on a negative effect, we must update because we now know that there is no probability of a negative effect for this country.

Furthermore, note that our method for ruling out the "negative effects" relied only on our knowledge that this country relapsed when poor—we did not impose any constraints on the outcome we would have observed if this country had been rich. Therefore, we can rule out negative effects regardless of the exact method of manipulation. To be specific, for this country we know that the effect of increasing income due to discovering oil, or the effect of increasing income due to improving education, or the effect of increasing income through any other means, could not have had a negative effect—it is logically impossible to move from consolidation to relapse if we know that the starting point is relapse.

While these sorts of logical arguments are straightforward when considering individual cases, they can also be used to consider the limits of instability for regression results. These limits provide conservative estimates of uncertainty. In the next section, I provide a stylized example.

The Limits of Instability for Regression Slopes with Observational Data: A Stylized Example

While the potential instability of regression results with observational data is well known, it is helpful to consider a stylized example in order to explore the source and the limits of this instability. Figure 1 (a) presents an example of a simple regression. In this case, both X and Y are dichotomous and only take on the values 0 and 1, therefore this plot can also be thought of as a 2x2 table (the points in the plot have been jittered so all 12 of them can be seen).[2] Notice that the arrangement of points in the plot implies a positive regression slope of 1/3. Two thirds of the X=1 points have Y=1, while one third of the X=0 points have Y=1, so the difference is 1/3.[3]

Of course, the regression slope of 1/3 that we see in Figure 1 (a) merely represents a simple regression result. We could get different regression results by conditioning on a third variable. For example, consider a hypothetical dichotomous conditioning variable that takes the values "a" and "b." Figure 1 (b) presents the same data that were presented in Figure 1 (a), except now the values of this conditioning variable are represented on the plot. The conditional regression analysis (based on this dichotomous conditioning variable) works by fitting two a separate regression lines—one to the "a" points and one to the "b" points. For this example, we see that the slope is zero for both of these lines, and therefore our conditional regression estimate is zero (quite different from the 1/3 we got in the simple regression).

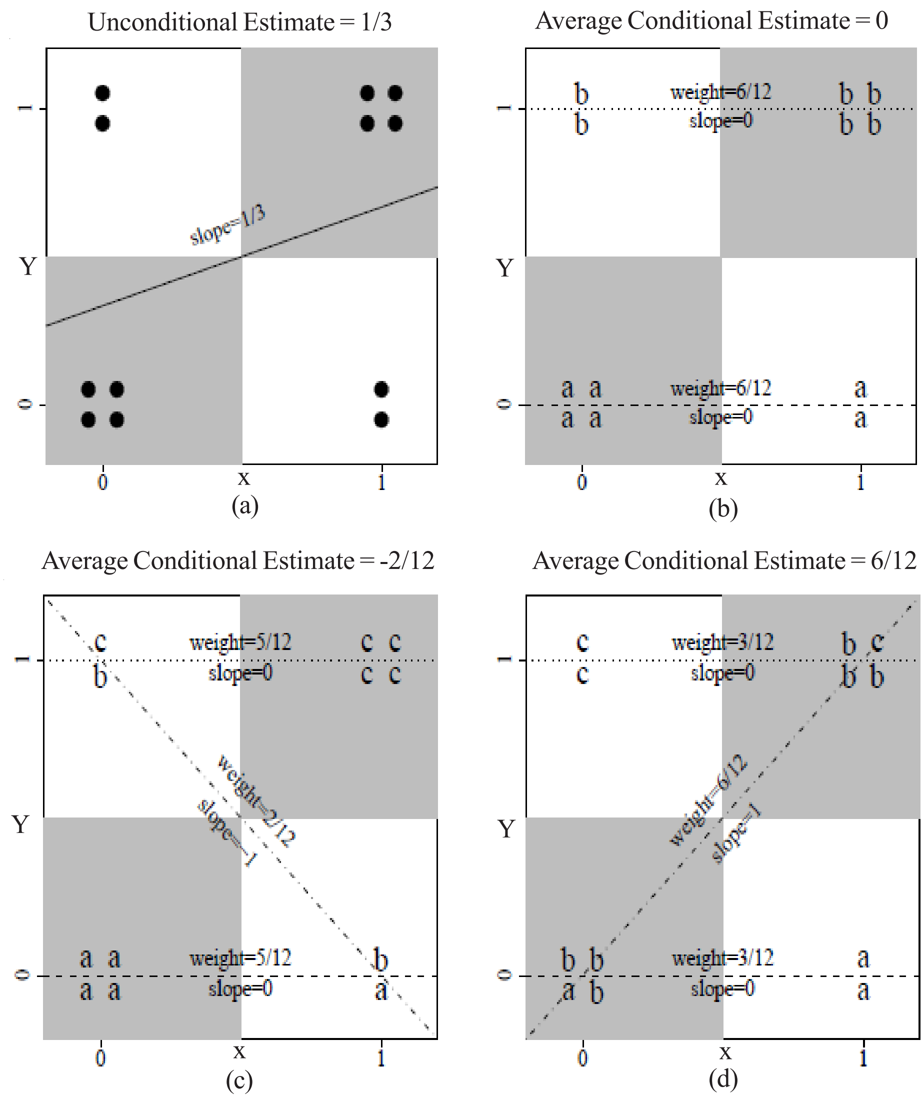
This instability of regression results is well known to most practitioners, and we can get many different answers by considering alternative conditioning variables. For example, Figure 1 (c) demonstrates how we can get a very small regression estimate. This plot presents the same data again with a different conditioning variable that is trichotomous (taking the values "a", "b", and "c"). In Figure 1 (c), the regression slope is zero for both the points that take the conditioning value "a" and for the points that take the conditioning value "c." However, the regression slope is -1 for the points taking the conditioning value "b." In a model with different conditional slopes for different observations (i.e., a model with interactions), we obtain the overall estimate by averaging across the conditional slopes according to the proportion of observations associated with each slope (Cochrane 1968). Using weights according to the proportion of the "a," "b," and "c" observations, the conditional estimate is the average slope, which is -2/12.[4]

Similarly, Figure 1 (d) demonstrates how we can get a very large regression estimate. This plot presents the same data again with a different trichotomous conditioning variable. In this plot, the regression slope is zero for both the points that take the conditioning value "a" and for the points that take the conditioning value "c." However, the regression slope is 1 for the points taking the conditioning value "b." Using weights according to the proportion of observations taking each value of the conditioning variable, the average conditional estimate is the average slope, which is 6/12.

Again, it is not remarkable that Figures 1 (a), (b), (c), and (d) demonstrate the instability of regression results. What is remarkable, or at least is less well known, is that the plots in Figure 1 demonstrate a limit to this instability.

To see this, first imagine all possible conditioning variables that are not collinear with X. That is, if you are allowed to label the 12 points in any manner you wish ("a," "b," "c," "d," …), so long as each label has a representative in the X=0 and

Figure 1



X=1 groups, there is no labeling that will produce an average slope smaller than -2/12 (as in Figure 1 (c)) or greater than 6/12 (as in Figure 1 (d)). In other words, if we restrict ourselves to conditioning variables that are not collinear with X, there is no set of conditioning variables that would produce an average slope estimate that is smaller than -2/12 or greater than 6/12—there is a limit to the instability of the regression results when we assume an absence of collinearity.

In fact, even if we allow conditioning variables that are collinear with X (i.e., we are allowed to label the 12 points in any manner we wish), the average slope must at least be -1/3 and can at most be 2/3—there is a limit to the instability of the regression results. This limit is often known as the Manski bounds (Manski 1990, 2003). In the language of the previous section, we can rule out average slopes smaller than -1/3 or larger than 2/3.

Development, Democracy, and Gerring's Proposed Standards?

If the Manski bounds demonstrate limits to the instability of regression results, then messy observational data may add to our knowledge of a given subject area—even if we cannot provide a single answer, we can rule out some answers. However, while we learn something from messy data, we may be unsatisfied by the wide range of answers we obtain. Without additional information or assumptions, we cannot state that any particular slope estimate (or set of slope estimates) in the bounds is more likely than any other. Perhaps most importantly, without additional information or assumptions, we can never rule out a slope of zero, and therefore we can never establish the presence of an effect.

As an example, consider the previously discussed result on democratic consolidation. Has this finding added to our knowledge on the subject? In light of the stylized example above, the answer seems to be yes. A positive correlation between development and retained democracy does not conclusively indicate a positive causal effect; however, it does rule out highly negative effects.[5]

However, while the Manski bounds demonstrate conclusively that we learn something from this observational data (because they represent all the results we might get from any conceivable set of conditioning variables), they are likely to be conservative because they only utilize data on the explanatory variable and the outcome variable. By utilizing other variables, we may be able to tighten these bounds, but only when the information from these variables is combined with additional assumptions.

As an example, consider the typical practice of running regressions with different sets of conditioning variables, and using the range of results we get from these regressions as bounds. These bounds will be valid only if we assume that one of our sets of conditioning variables is the correct set, or if we assume that the answer we would get from the correct set of variables is contained within these bounds. In other words, the conditioning variables may allow us to tighten the Manski bounds, but only if we are willing to make assumptions on the basis of these variables.[6]

Unfortunately, as demonstrated by many of the regression critics cited by Gerring, these alternative bounds (the range of estimates produced by running different regression specifications) are likely to be anti-conservative. Because regression results can, in practice, only utilize non-collinear sets of measured confounders, we will often find it untenable to believe that any of the regressions we have run use the correct set of conditioning variables.

This leaves the analyst in a quandary. Given our data, we have a conservative answer (the Manski bounds) and an anti-conservative answer (the bounds on regression results), but we may not have a "plausible" answer. In order to move forward, the analyst must be willing to make and defend assumptions. In some cases, the bounds on regression results are expanded by making assumptions about the effects of unobserved conditioning variables (Brumback et al. 2004, Lin et al. 1998, Rosenbaum 1987, 2002, Rosenbaum and Rubin 1983). In others, the Manski bounds are tightened by making a variety of assumptions (Manski 2003), sometimes in a Bayesian framework (Quinn 2011).

Specifying and defending these assumptions is hard work, and we may never arrive at a set of assumptions that will be agreeable to all readers—messy data lead to messy conclusions. Still, this sort of work with messy observational data can lead to useful conclusions—see for example the Cornfield et al. (1959) study on smoking and lung cancer. It is unfortunate that typical practice eschews this hard work and presents only the anti-conservative bounds implicit in regression results.

Notes

[1] More typically, we might propose that these probabilities follow a Dirichlet distribution, and specify our priors on the parameters of this distribution.

[2] Linearity is justified in this case due to the binary treatment.

[3] To avoid concerns about a lack of significance, we can interpret the points in this plot to merely represent the proportions in the cells of a 2x2 table. Therefore, the sample size might actually be quite large.

[4] This process of averaging over the slopes of regression lines in an interactive model is sometimes known as direct adjustment (Cochran 1968), and is the most straightforward way to estimate an average effect using an interactive regression model.

[5] With a continuous explanatory variable, the causal question of interest must be stated more precisely in order to estimate effects or specify the regression bounds. Often, parametric assumptions are made that obviate the need for this precision.

[6] It is worth noting that this is also true for post-treatment variables, although standard regression analysis cannot be used when conditioning on post-treatment variables. Glynn and Quinn (2011) provides an example on the use of post-treatment variables to tighten the Manski bounds.

References

Brumback, Babette A., Miguel A. Hernan, Sebastien J. P. A. Haneuse, and James M. Robins. 2004. "Sensitivity Analyses for Unmeasured Confounding Assuming a Marginal Structural Model for Repeated Measures." *Statistics in Medicine* 23, 749–767.

Cochran, W. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrics* 24, 295–314.

Cornfield, J., W. Haenszel, E. Hammond, A. Lilienfeld, M. Shimkin, and E. Wynder. 1959. "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions." *Journal of the National Cancer Institute* 22, 173–203.

Glynn, Adam N. and Kevin M. Quinn. 2011. "Why Process Matters for Causal Inference." *Political Analysis*, forthcoming.

Lin, D. Y., B. M. Psaty, and R. A. Kronmal. 1998. "Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies." *Biometrics* 54, 948–963.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *The American Economic Review Papers and Proceedings* 80:2, 319–323.

Manski, Charles F. 2003. *Partial Identification of Probability Distributions*. New York: Springer Verlag.

Rosenbaum, Paul R. 1987. "Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies." *Biometrika* 74, 13–26.

Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome." *Journal of the Royal Statistical Society*: Series B 45, 212–218.

Quinn, Kevin M. 2011. "What Can We Learn From a 2x2 Table?" Unpublished Manuscript.

Qualitative Research: Progress Despite Imperfection

Andrew Bennett
Georgetown University
bennetta@georgetown.edu

I have encountered what John Gerring aptly describes as a fear of "the specter of methodological perfectionism" in my students and colleagues. In my view this fear imputes to methodologists more optimism on the perfectibility of research methods and more pessimism on the contributions of imperfect methods than most of us actually hold, but like any phobia, this fear is sufficiently real in the minds of those who hold it that it deserves remediation.

My own view, similar to that in Gerring's *Social Science*

Methodology: A Criterial Framework (2001), is that there are no perfect research methods for observational data or even for the experimental and quasi-experimental approaches that are generating renewed interest, so we must make methodological choices among imperfect alternatives that represent different tradeoffs of desiderata such as external and internal validity. Alexander George and I tried to make this clear in our 2005 book by discussing the limits as well as the comparative advantages of case study methods (2005: 22–34) and explicitly rejecting perfectionism in research (2005: 10–11, fn. 14). Soon after we published our book, however, I noticed that when they presented their work, graduate students, visiting speakers, and even colleagues would cast a nervous glance in my direction whenever the issue of qualitative methods arose, apparently concerned that I would accuse them of grievous methodological mistakes.

Of course I do critique other scholars' research methods, just as I expect them to critically evaluate my own research methods and findings. My critiques, however, assume that all methods are imperfect, that methodological choices hinge in part on prior theoretical and empirical knowledge that other scholars usually possess to a greater degree than I do regarding their particular research topics, and that reasonable people can disagree on the complex question of which methodological choices are better than others in a particular project with particular research objectives (theory generation, theory testing, etc.). The fear of methodological perfectionism should thus be relatively easy to lay to rest, even if we must all continuously concern ourselves with how our methods might be better.

In addressing the specter of methodological perfectionism, however, we must also address the opposite worry of methodological fallibilism: if our methods are inevitably imperfect, and if it is difficult to get inter-subjective agreement on what methods to use in a particular project, how can we make any claims that our research leads to cumulative, progressive, and general knowledge that is useful to policymakers?

The challenge, then, is to prevent the "best" from being the enemy of the "good," or to ensure that our aspiration to optimize our methods does not paralyze us and keep us from pursuing work that is imperfect but that makes genuine contributions worthy of the time and effort invested. In this essay, I address the questions of both perfectionism and fallibilism by exploring an example from my own work that I consider to be methodologically imperfect but theoretically and empirically progressive and policy relevant. I proceed as follows: after a very brief discussion of the general issues that Gerring raises, I re-examine research that colleagues and I did in the 1990s on burden-sharing in security coalitions (Bennett, Lepgold, and Unger, 1994, 1997). I use this example to demonstrate how even imperfect work such as ours, together with subsequent imperfect research by other scholars, can contribute to cumulative, generalizable, and policy-relevant insights. After summarizing our research, which focused on burden-sharing in the 1990–1991 Gulf War coalition, I review some of the subsequent data and research on burden-sharing in later security coalitions, focusing on those involved in the Balkans in the 1990s, Afghanistan since 2001, and Iraq since 2003. This discussion shows the strengths of our research in anticipating burden-sharing outcomes in later security coalitions as well as the weaknesses of our work in overlooking important phenomena. My analysis builds upon and updates the discussion of this example in George and Bennett (2005: 255–61), and I present it in such a way that colleagues can teach this example Socratically to get students to think about issues of case selection, theory-generation, typological theorizing, policy relevance, and theoretical progress.[1]

Methodological Perfectionism

John Gerring's essay argues that "the current vision of perfection prizes causal knowledge over descriptive knowledge, theory appraisal over theory discovery, micro-theory (aka micro-mechanisms) over macro-theory, and internal validity over external validity." Gerring does not cite any specific works in this passage as embodying these four perspectives, but as a prelude to re-analyzing my earlier research I briefly summarize my views on them here.

On the first point, I support the view, nicely expressed by King, Keohane, and Verba (1994: Chapter 2), that both causal and descriptive knowledge are important and that we can't make useful causal inferences unless we get our descriptions right. Much of the value of the historical explanation of cases lies in providing detailed descriptions of the historical events that we deem relevant to building and testing theories, and many important contributions to the social sciences are empirical rather than theoretical.

Regarding the appraisal and discovery of theories, George and I (2005:12) critiqued King, Keohane, and Verba (1994: 14) for putting too much emphasis on theory appraisal relative to theory discovery. In contrast, we included the term "theory development" in the title of our book to embrace both the discovery and appraisal of theories.

On the micro-macro issue, George and I have argued that explanation via reference to causal mechanisms requires researchers to commit in principle to make their theories consistent with the finest level of detail that they can observe. We also explicitly stated, however, that this commitment "does not mean that the explanatory weight or meaningful variation occurs at this level" and we added that "macrosocial mechanisms can be posited and tested at the macrolevel" (141–42).

Finally, although some might read George's and my book as emphasizing internal validity over external validity, this misses the crux of our argument. Naturally, we all want both external and internal validity, but the point George and I made is that statistical methods have some advantages at the former and case studies are in some ways better at achieving the latter (hence our encouragement of multi-method research). We did not privilege internal over external validity (see pp. 109–124 on generalizing from case studies), but we warned against doing the reverse, critiquing King, Keohane, and Verba for offering advice on ways of increasing sample sizes without noting the tradeoffs this can involve regarding internal validity and conceptual stretching (George and Bennett 2005: 172–176).

In short, I see the dichotomies Gerring lays out as being

Table 1: A Typological Theory on Burden-Sharing in the 1991 Gulf War (George and Bennett 2005: 257)

	Collective Action	Balance of Threat	Alliance Dilemma	Domestic Politics	Expected Outcome
Saudi Arabia	Y	Y	Y	Y	Contribute
Turkey	Y	Y	Y	N	Contribute*
	Y	Y	N	Y	Contribute
	Y	N	Y	Y	Contribute
	N	Y	Y	Y	Contribute
	Y	Y	N	N	Contribute*
United States	Y	N	N	Y	Contribute
	Y	N	Y	N	Contribute*
Great Britain	N	N	Y	Y	Contribute
Egypt	N	Y	Y	N	Contribute
	N	Y	N	Y	Contribute*
France, Canada, Australia	N	N	N	Y	Contribute
Germany, Japan	N	N	Y	N	Contribute*
Iran, Syria	N	Y	N	N	No Contrib.
	Y	N	N	N	No Contrib.
China, USSR	N	N	N	N	No Contrib.

*In types marked with an asterisk, countries are expected to contribute only if strong state leaders override domestic opposition, as happened in Egypt and Turkey. In Japan and Germany, security dependence on the United States was so high that domestic opposition was muted.

"both-and" desiderata rather than "either-or" propositions, and I welcome research projects that make widely different tradeoffs among these desirable features of social science. If I understand Gerring correctly, this is his view, as well.

Progress Despite Imperfection: Burden Sharing in Security Coalitions

In the early 1990s, with my colleagues Joseph Lepgold and Danny Unger, I set out to study the puzzle of why coalition contributions to the Desert Storm coalition that ousted Iraqi forces from Kuwait in 1991, totaling $70 billion dollars and 240,000 troops, were far more extensive than would have been predicted by collective action theory, which at that time was the dominant theoretical approach to alliance behavior. By itself, collective action theory should have predicted free-riding from all states except the United States in response to Iraq's invasion of Kuwait, as the United States appeared to be capable of and willing to evict Iraq from Kuwait all by itself if necessary (President Bush publicly vowed in August 1990 that the Iraqi invasion of Kuwait would not stand even before he had concrete commitments of support from U.S. allies).

To address the surprising lack of free riding, we developed a typological theory that integrated collective action theory with three other explanations. First, all other things equal, balance of threat theory should have predicted contributions from states like Kuwait and Saudi Arabia whose occupation by or proximity to Iraq made them feel threatened and who feared the United States would not provide the public good of security against Iraq to a sufficient degree without assistance. Second, alliance security dilemma theory should have predicted contributions from states which were at that time heavily dependent on the United States for their security, such as Germany and Japan. Finally, theories of domestic politics suggested that a state would contribute to the Desert Storm coalition if public opinion in that state favored a contribution or if state leaders wanted to make a contribution and were powerful enough to override public opinion.[2]

Our typological theory on how combinations of the variables in each of these theories would interact in producing outcomes is summarized in Table 1. Note that the columns in Table 1 indicate how the variables in one theory, acting in isolation from those of other theories, would predict a contribution from a state. For example, collective action theory would look at several variables, including the relative capabilities of a state, the expected costs of forcing Iraqi troops out of Kuwait, and whether a state was needed as a basing ground (as in Saudi Arabia and Turkey), to determine whether collective action theory should have predicted a contribution from or free riding by that state. The rows of Table 1 indicate how states with the specified combinations of predictions from the four individual theories should be expected to behave.

Our typological theory did quite well in fitting and explaining the contributions of the cases we studied, including the United States, Britain, Egypt, France, Germany, and Japan, and the process tracing evidence in these cases largely validated the theory (Syria's contribution was an anomaly for our theory but the rest of the cases listed largely fit the theory [Bennett, Lepgold, and Unger, 1997]). In particular, the study highlighted the fact that alliance dependence was in itself sufficient to engender large economic contributions from Japan and Germany even though essentially all the other variables in these cases created incentives to try to free ride on American efforts.

Yet the first Gulf War was not an entirely independent test of our theory, as we had constructed our theory with some preliminary knowledge of the outcomes of the cases (the process tracing evidence, of which we were largely ignorant before actually carrying out the case studies, provided more independent corroboration of the theory's hypothesized mechanisms).

A tougher test concerns contributions and non-contributions to subsequent security coalitions. Table 2 presents some of the cases that do and do not fit our initial theory well from security coalitions in Bosnia (first the United Nations Protection force from 1992–1995 [UNRPOFOR] and then the NATO implementation force 1995–1996 [IFOR]), Iraq 2003–2011, and Afghanistan 2001–2011.[3] Cases are identified by the name of the (non)contributor, the year the coalition started (not necessarily the year of a country's contribution), and the number of troops contributed (in bold). The cases that arguably fit our theory are in plain type, while those that do not are in italics. Table 2 includes only countries that contributed more than 500 troops, except for the case of the small U.S. contribution to UNPROFOR to allow a comparison to the large U.S. contribution to IFOR. I also include a few instances of countries that might have been expected to contribute more than 500 troops but did not. Each coalition also had a large number of contributions of fewer than 500 troops.

The UNPROFOR and IFOR cases suggest our theory got a lot right (Lord and Lord, 1997): the more powerful and proximate countries were the biggest contributors (collective action and balance of threat); coalition actions were more effective once the United States, which had stayed out of UNPROFOR, got involved in IFOR (collective action); the United States felt only a modest threat from the Balkan crises and was reluctant to get involved until the credibility of NATO was at stake (domestic politics, balance of threat); and once the United States did become involved through IFOR, many smaller states seeking to join NATO made contributions to win favor from the U.S. (alliance security dilemma).

Yet the high contributions of Jordan, Pakistan, and others to the UNPROFOR coalition do not fit our theory well. Our theory would have predicted modest symbolic contributions from each of these countries. One possible explanation is the Muslim countries' religious affinity for the Bosnian Muslims. A more likely interpretation, however, is evident in the comparison with troop contributions to NATO's IFOR operations: U.N. peacekeeping payments of about $1,000 per soldier per month, well more than what many developing countries paid their soldiers, may have helped motivate some countries' contributions. This is evident in the fact that nearly all the developing countries that were big UNPROFOR contributors dropped out of the coalition when it transitioned to the pay-your-own-way IFOR force (Turkey, aspiring to be a member of the EU, stayed in the coalition). The role of such side payments does not contravene the logic of our original theory—collective action theory allows for private goods and side payments, and our case study of Egypt noted that one of the motivating factors behind its contribution to the 1991 Gulf War was that the United States forgave Egypt's debts in return. Yet our initial research could have put more emphasis on the potential role of side payments (Tago 2008).

A second interesting phenomenon not captured in Table 2 is the large number of countries giving token contributions of a few soldiers. We understood during our initial research that in order to limit coordination costs the U.S. military wanted only a few large contributions from other countries, and niche contributions of capabilities, like mine-sweeping, that it lacked. The State Department, meanwhile, wanted to count as many countries as possible as contributors, regardless of the size of their contributions, to show the legitimacy of U.S. operations. Clearly, the latter concern often won out, but how, when, and at what cost are interesting and underexplored questions. In particular, when the number of coalition contributors with a say over key issues like prioritization of bombing targets grows larger, the transactions costs of achieving agreement on military tactics and strategies grows (Auerswald 2004). Also, the motives of small contributors may have extended beyond alliance dependence to a desire to share the spoils (or contracts) of U.S. coalitions and perhaps even a desire to learn from (or spy upon) U.S. military operations. Here again, more research is warranted.

The data on contributions in Iraq and Afghanistan similarly show the value of our theory while at the same time suggesting several other researchable puzzles that we failed to anticipate or overlooked. The combination of collective action theory, balance of threat, alliance dependence, and domestic politics clearly remained powerful as an explanation of states' contributions. It is not surprising that the 2003 Iraq coalition was much smaller than that in 1991, for example, due to the lower sense of a shared threat from a weakened Iraq in 2003 and the decline in alliance dependence on the U.S. as the Cold War receded. Our domestic politics hypothesis helps explain why the newly elected government of Turkey refused in 2003 to allow the United States to use Turkey to stage a northern prong of its invasion of Iraq (Baltrusitis 2010).[4]

Table 2: Contributions and Non-Contributions That Do and Do Not Fit the 1994 Theory

	Collective Action	Balance of Threat	Alliance Dilemma	Domestic Politics	Expected Outcome
1992 France **4,600** 1992 UK **3,500** 1995 UK **7,500** 1995 France **7,500** 1995 Italy **2,500**	Y	Y	Y	Y	Contribute
	Y	Y	Y	N	Contribute*
1995 US **20,000** 2001 US **23,600** 2003 US **250,000**	Y	Y	N	Y	Contribute
	Y	N	Y	Y	Contribute
2003 UK **46,000**	N	Y	Y	Y	Contribute
	Y	Y	N	N	Contribute*
	Y	N	N	Y	Contribute
	Y	N	Y	N	Contribute*
1992 Canada **2,000** 1992 Neth. **1,700** 1992 Turkey **1,500** 1992 Spain **1,400** 1992 Denmark **1,100** 1992 Sweden **1,100** 1992 Poland **1,100** 1992 Ukraine **1,100** 1992 Czech. **1,000** 1992 Belgium **900** 1992 Norway **800** 1992 Slovak Rep. **600** 1995 Neth. **2,100** 1995 Turkey **1,600** 1995 Norway **1,000** 1995 Sweden **900** 1995 Czech. **900** 1995 Denmark **800** 1995 Poland **700** 2001 UK **8,500** 2001 France **1,700** 2001 Canada **2,500** 2001 Italy **2,400** 2001 Neth. **1,800** 2001 Poland **1,100** 2001 Spain **800** 2001 Turkey **800** 2001 Denmark **700** 2001 Norway **600** 2001 Romania **600** 2003 Italy **3,200** 2003 Poland **2,500** 2003 Georgia **2,000**	N	N	Y	Y	Contribute

Table 2 continued.

	Collective Action	Balance of Threat	Alliance Dilemma	Domestic Politics	Expected Outcome
2003 Ukraine **1,700** 2003 Spain **1,300** 2003 Neth. **1,200** 2003 Romania **700** 2003 Japan **600** 2003 Denmark **500**					
1995 Germany **4,000** 2003 Turkey **NC**	N	Y	Y	N	Contribute*
1992 Russia **1,300** 1995 Russia **1,600** 2001 India **NC**	N	Y	N	Y	Contribute
1992 Jordan **3,400** *1992 Pakistan* **3,000** *1992 Malaysia* **1,500** *1992 Bangl.* **1,200** *1992 Kenya* **1,000** *1992 Nepal* **900** 1992 Argentina **900** *2001 Australia* **1,100**	N	N	N	Y	Contribute*
2001 Germany **3,400** 2003 Korea **3,600**	N	N	Y	N	Contribute*
	N	Y	N	N	No Contrib.
1992 U.S. **300**	Y	N	N	N	No Contrib.
2003 Australia **2,000**	N	N	N	N	No Contrib.

Notes on Table 2: 1992 is UNPROFOR deployments as of 2005; 1995 is IFOR deployments as of September 1996; 2001 is Afghanistan deployments as of June 2008; 2003 is peak deployments in Iraq between 2003 and 2007. Anomalous cases are in italics; troop numbers are in bold (troop numbers often changed over time and these numbers are indicative rather than definitive). **NC** is for surprising non-contributors. Pakistan is not listed as a contributor in the Afghanistan coalition because it does not have troops in Afghanistan, but Pakistani soldiers have contributed to the coalition effort by fighting Islamic militants in northern Pakistan.

One interesting and researchable phenomenon that our initial research only partly anticipated is that when two coalition operations (Afghanistan and Iraq) were ongoing simultaneously, the operation in which the alliance leader (in this case, the United States) was having more difficulty rounding up contributions was the one in which the leader called in favors from its most dependent allies. This is evident in the fact that Korea, Poland, Georgia, and Ukraine, all feeling dependent on the United States for their security, were disproportionately large contributors in Iraq (see Baltrusitis 2010 on the Korean case). At the same time, large U.S. NATO allies were bigger contributors in the U.S. led coalition in Afghanistan, which had more legitimacy in the views of these contributors. This is consistent with our original theory's emphasis on the alliance dependence variable, but it is a rather novel example in that we did not anticipate how such dependence might play out when two operations of unequal global legitimacy were ongoing.

In addition, the British contribution in Iraq and the Australian contributions in Iraq and Afghanistan may attest to the power of individual leaders (see Dyson 2007 on the British case). This could be encompassed by our domestic politics variables but it accords a bigger role to specific individuals than we gave in our original articulation of these variables.

Probably the biggest missed opportunity in our original research, however, concerns our decision to set aside an anomalous case that we thought was a one-off phenomenon unlikely to be repeated. In our research on the 1991 Gulf War, we understood the case of Israel to be anomalous. Israel deeply shared the U.S. goal of evicting Iraq from Kuwait and reducing Iraq's military capabilities, but as Israel understood any contribution it made would disrupt the participation of Saudi Arabia, Syria, Egypt, and others in the Desert Storm coalition, Israel contributed by *not* contributing. Even more interesting, Israel endured SCUD missile strikes from Iraq during the war but did not retaliate directly in order to help keep the Desert Storm coalition intact. Israeli leaders made it clear at the time that in exchange they expected the U.S. to support Israel's position strongly in any post-war negotiations with the Palestinians.

I always tell my students to look for anomalies as sites for potential theory development, and here we had a whopper: a state that contributes by not contributing, and then uses its *non-contribution* to try to extract concessions from the coalition leader. Yet we did not do a case study of Israel's behavior or theorize about it more deeply because we thought the case was so unlikely, idiosyncratic, and well-explained by quirky facts of which we were already aware that it did not merit further study.

On this we were woefully wrong. Fast forward to the U.S.-led coalition in Afghanistan and the next question I ask my students when I teach this example: Who is the "Israel" in the Afghan coalition, the country whose contribution would have most disrupted the contributions of another key member or members of the coalition? The answer, clearly, is India, as any contribution of troops from India would have vastly complicated the already problematic Pakistani assistance that is vital to U.S. efforts in Afghanistan. Like Israel, India used its non-contribution to try to extract concessions from the United States. India rather unsubtly let it be known that if the U.S. was not able to successfully pressure Pakistan to sharply reduce its supports of militants in the disputed Kashmir region, India would feel compelled to move more troops into Kashmir as part of the regional push against Islamic militants. This would drive Pakistan to move its troops from the tribal areas in Pakistan's north, where they were fighting against Al Qaeda and Taliban fighters in areas bordering Afghanistan, to Pakistan's southern border along the Kashmir region (Cody 2001; Coll 2008). Such a redeployment of Pakistan's forces, of course, would have made it harder for the U.S. to establish security in Afghanistan. For this reason, the U.S. responded to India's threatened "contribution" by pressuring Pakistan to cut off its support for militants in Kashmir.

India's behavior shows that the case of Israel in our initial study was exactly the kind of potentially fruitful anomaly that I urge my students to prize. On first glance, the case of Israel looks explicable but so unusual that it is unlikely to generalize to many other cases. On closer inspection, however, the case of Israel reveals a counterintuitive outcome that is explicable in terms of a perfectly logical and highly generalizable mechanism: In many coalitions, there are one or more potential coalition partners who would cause a net loss to the strength, size, unity, or decisiveness of the coalition if they joined it because their disagreements or adversarial relations with existing coalition members would cause these members to drop out of the coalition or create debilitating arguments within it. This applies not only to international coalitions, but to domestic political coalitions, where it helps explain the awkward relations between, for example, the Tea Party and the Republican Party. The Tea Party wants to leverage its power by winning policy concessions from the Republican Party, and it has demonstrated its ability to win Republican primary nominations for Tea Party candidates who have then gone on to lose general elections in districts or states that Republicans hoped to win. Republicans, on the other hand, want to co-opt Tea Party supporters but worry that reaching out to or formally endorsing the Tea Party will alienate voters and elites who currently support the Republican Party. By failing to identify and theorize about the mechanisms related to such problematic potential coalition members, we missed a theory-building opportunity worthy of a dissertation or book.

The subsequent development of the security coalition literature reveals a number of additional improvements on our work and raises interesting questions that we had not thought to ask. These include: detailed and nuanced studies of the domestic politics of coalition contributions (Baltrusitis 2010; Spiezo 2008; Tago 2005); studies of how an increase in coalition members with a voice in decision-making complicates military operations (Auersawld, 2004); analysis of how security threats and national prestige affected allied contributions to every major U.S. led coalition from Vietnam in the 1960s to Iraq in 2003 (Davidson 2011); exploration of why states leave security coalitions (Tago 2009); examination of why states channel coercion through international organizations (Thompson 2009); and a study of why Britain joined the U.S. coalition in Iraq but not Vietnam (Dyson 2007). Many, but not all, of the authors of these works cited our study and were perhaps partly inspired by it to pose the questions they raised, but of course I only wish we had the perfect foresight to identify all these questions as the interesting next steps in the research program on security coalitions.

Still, despite these and other imperfections, I am satisfied that our earlier research improved upon the literature extant at that time, provided policy-relevant generalizations that got most of the big contributions (and non-contributions) to subsequent coalitions mostly right, and helped inspire subsequent research into related questions. These important and worthwhile contributions are not perfect, but they are good enough for me.

Notes

[1] When I teach this example, I assign students to read Bennett, Lepgold, and Unger (1994), and then in class I present the tabular data herein from the Balkan, Iraqi, and Afghan security coalitions and ask them what the 1994 article got right, what it missed, and how it could be improved upon in light of subsequent evidence. Of course, assigning the present article could discourage students from thinking hard about their own conclusions to these questions, so the present article should be seen as the beginnings of an "answer key" for teachers to use or to distribute after a class discussion of this example.

[2] Bennett, Lepgold, and Unger (1994). Our later study (Bennett, Lepgold, and Unger, 1997) added a fifth theory, which argued that lessons learned from previous burden sharing episodes would affect behavior in subsequent coalitions. I leave this argument aside here for the sake of simplicity.

[3] For teaching purposes, a web link at: http://www.maxwell.syr.edu/moynihan/cqrm/Newsletters/ provides more complete data, including small coalition contributions, on each coalition. These data were compiled from news accounts and web sites over a number of years for teaching purposes but they are not authoritative and I do not list the sources as would be necessary to use the data for actual research. Also, the data are snapshots of frequently changing contributions and do not represent the totality of states' contributions over time. Nonetheless, the data work well for the pedagogical purpose of asking students to identify which cases they think do and do not fit our original theory, and why. When I teach this example, I present each

table in turn and ask my students what our typological theory got right and what it missed. Here, I present my own answers, to which teachers and their students can add their own insights when discussing this example in class.

[4] Bennett, Lepgold, and Unger (1997) added an additional hypothesis suggesting that countries' "lessons learned" from previous security coalitions or uses of force would shape their subsequent contributions, and this helps explain why Turkey, which felt it did not get as much compensation as it was promised in 1990–1991 for enforcing trade sanctions on Iraq, was less eager to help in 2003. Learning may also explain why Russia did not join the 2001 Afghanistan coalition, and why Germany and Japan started to move toward contributing troops to U.S. led coalitions as a less expensive alternative to the more than $10 billion that each country contributed to the first Gulf War coalition.

References

Auerswald, David. 2004. "Explaining Wars of Choice: An Integrated Decision Model of NATO Policy in Kosovo." *International Studies Quarterly* 48:3, 631–662.

Baltrusitis, Daniel. 2010. *Coalition Politics and the Iraq War*. Boulder: Lynne Rienner.

Bennett, Andrew, Joseph Lepgold, and Danny Unger. 1994. "Burden Sharing in the Persian Gulf War." *International Organization* 48:1, 39–75.

Bennett, Andrew, Joseph Lepgold and Danny Unger. 1997. *Friends in Need: Burden Sharing in the Persian Gulf War*. New York: St. Martin's Press.

Cody, Edward. 2001. "U.S. Anti-Terror Effort May Be Diluted If Pakistan Calls on Army to Fight India." *The Washington Post* (28 December).

Coll, Steve. 2008. "Between India and Pakistan, A Changing Role for the U.S." *The Washington Post* (26 May), B1.

Davidson, Jason W. 2011. *America's Allies and War: Kosovo, Afghanistan, and Iraq*. New York: Palgrave Macmillan.

Dyson, Stephen. 2007. "Alliances, Domestic Politics, and Personality: Why Did Britain Stay out of Vietnam and Go Into Iraq?" *Political Psychology* 28:1, 647–666.

Gerring, John. 2001. *Social Science Methodology: A Criterial Framework*. Cambridge: Cambridge University Press.

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

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

Lord, Jeffrey R. and Kristin M. Lord. 1997. "Additional Cases from the Gulf Crisis and Burden Sharing in UNPROFOR." In *Friends in Need: Burden Sharing in the Persian Gulf War.* Bennett, Lepgold, and Unger, eds. (New York: St. Martin's Press).

Spiezo, Edward. 2008. "Domestic Structure and Collective Security." *Politics and Policy* 25:1, 111–132.

Tago, Atsushi. 2005. "Determinants of Multilateralism in U.S. Use of Force: State of Economy, Election Cycle, and Divided Government." *Journal of Peace Research* 42:5, 585–604.

Tago, Atsushi. 2008. "Is There an Aid-for-Participation Deal? US Economic and Military Aid Policy to Coalition (Non)Participants." *International Relations of the Asia-Pacific* 8:3, 379–398.

Tago, Atsushi. 2009. "When are Democratic Friends Unreliable? The Unilateral Withdrawal of Troops from the Coalition of the Willing." *Journal of Peace Research* 46:2, 219–234.

Thompson, Alexander. 2009. *Channels of Power: The UN Security Council and US Statecraft in Iraq*. Ithaca: Cornell University Press.

Reply

John Gerring
Boston University
jgerring@bu.edu

I want to take this opportunity to thank the contributors to this symposium for their astute comments, which have informed my views of these topics and—I suspect—will also inform the views of many readers. I shall not attempt to address every issue that has arisen, but rather will address only those issues that seem to bear centrally upon my thesis. These comments will be divided across three areas: case study research, large-N observational research, and the larger topic of uncertainty (encompassing both).

Case Study Research (aka Causal Process Observations, Process Tracing)

Seawright seems to be arguing that case-based research is a strong form of causal inference, on par with experimental and quasi-experimental research. Yet, he does not provide any clarifying discussion of this point or any examples to back up this claim—which, to my mind, remains dubious.

Bennett seems to agree with my general argument—that causal inference is problematic in non-experimental settings and that case-based knowledge is no more immune to confounders than messy large-N observational data.

Bennett does not share my impression of a creeping methodological perfectionism in the field of qualitative methods. One might conclude that the intellectual movement has made less headway among quals than among quants. Perhaps so. But whether enforced by members of one camp or the other, the continued marginalization of qualitative methods in the discipline (e.g., their under-representation in top general-interest journals[1]) would seem to provide evidence of the phenomenon of which I speak. Someone is judging that case studies do not surpass the bar of causal inference.

Large-N Observational Research

Seawright, Glynn, Bennett, and I agree that lots of problems attend any attempt to infer causation from large-N observational data, and that regression is not the *deus ex machina* that some members of the discipline once thought it was (or seemed to think it was). We agree that problems may be due to under-specified models as well as over-specified models. We agree that observational studies often suffer from ambiguity due to the nonmanipulable nature of the treatment (as Glynn discusses). We agree that observational studies depend on a great many—often quite heroic—assumptions about the data generating process, assumptions that cannot always be directly tested but that nonetheless must be justified. We agree that observational studies often over-state the certainty of their findings. We agree—to quote from Seawright's concluding sentence—that "regression-type studies of observational data have been significantly overvalued and overrepresented in our history over the last several decades," or at least given an overly optimistic interpretation. We also agree that experimental and quasi-experimental evidence should be pursued, wherever possible. And we agree, finally, that the growth in stature of these methods within political science is a very healthy development. In short, there is a high level of agreement among the contributors to this symposium.

Seawright and I part company when it comes to the following question. Is large-N observational data dispensable for causal inference? At times, Seawright seems to argue that it is. He writes: "Unstable regression results on observational data simply do not teach us about the direction, magnitude, practical relevance, or theoretical importance of the underlying causal relations." In other words, he appears to take the position that we learn nothing whatsoever from such results. We are no better off with them than without them. We might as well toss them out entirely.

Later in the essay, he softens his conclusions. Here, the issue seems to be *what percent* of worthwhile social science questions can be solved by first-class research designs—and how large, therefore, is the residual that must be approached with second- and third-class research designs (e.g., regression)? I believe this residual category is quite large. Seawright believes it is quite small, or at least that it will become so over time, as political scientists apply their ingenuity to recalcitrant questions.

Now we come to another ambiguity. Seawright finds that regression's strengths are *descriptive* and *exploratory*. He believes it to be…

> a powerful tool for the summary of complex cross-tabulations and scatter plots. Regression can sometimes make consistent but small descriptive relationships among variables more visible and can often dramatically aid comprehension of central themes in data by replacing an overwhelming mass of numbers or dots with a few key estimates… Furthermore, there are certainly moments when one or another piece of descriptive knowledge has strong causal implications; in such instances, regression may sometimes play a pivotal role in a causal argument.

In one way of thinking, *all* evidence is descriptive, from which the researcher must infer causality (Achen 1982: 77–78). However, some writers prefer to distinguish between evidence that is causal—in the sense that one can infer causality from it without making too many assumptions (e.g, from a well-designed experiment)—and evidence that is descriptive or exploratory, in the sense that many assumptions are required to infer causality. If this is the way one likes to dice up the methodological world then I will happily place large-N observational data analysis into the latter category.

Of course, this way of defining things creates other ambiguities, since some research really is descriptive and has no ambition to infer causality. What do we call that species of animal? My personal preference would be to classify data according to the role that it plays—either as evidence of descriptive relationships or relationships that are presumed to be causal—rather than the number of assumptions required to obtain a strong causal inference. But I realize that this is a matter of taste.

The point of importance is that *large-N observational data serves as the primary empirical basis for many extant causal inferences*. Calling this data descriptive or exploratory cannot take that away; it merely calls attention to the questionable status of the evidence, a message I fully endorse. In any case, Seawright seems to move closer to my own position in the concluding section of his essay.

With respect to the specific question of development and democracy, I don't claim that this relationship is robust to every possible gyration of the data, i.e., every plausible robustness test. Seawright, explicating Acemoglu, Johnson, Robinson, and Yared (2008), offers the intriguing speculation that the time-honored relationship reflects not economic development per se but rather each country's long-term historical position in the global economic hierarchy. It is not clear what causal mechanisms are operative here, or how they would differ from those associated with economic development. In any case, it's an interesting idea and certainly proves the point that observational data about macro-historical phenomena are open to a wide range of interpretations. (Indeed, even those who agree that development causes democracy often disagree on their explanation for the relationship.) I still find the standard explanations more plausible.[2] But I must quickly add that this is not a question I have worked with myself, so I don't have an intimate feel of the data (as Seawright does). Nor do I have a dog in the fight, so to speak.

However, the methodological point at stake is the following: if we want to know about the relationship between development and democracy, isn't regression (or some other large-N observational tool) a useful tool—and quite possibly the *most* useful of all available tools? It is telling that Seawright (following Acemoglu et al 2008) resorts to regression in his attempt to deconstruct the standard wisdom and to suggest a new hypothesis. This suggests that he is learning something from the regression analysis in Table 1 (and he admits as much in his later discussion, quoted above).

Uncertainty

There are many possible answers to what Popper called the demarcation problem—distinguishing scientific from non-scientific arguments. Among these possible answers, I would propose that *validity* and *certainty* are not nearly as important as *reasonable estimates of uncertainty*.

To be sure, there is a common impression that scientific studies ought to be valid more often than journalistic reports and that science ought to adhere to a higher (more certain) standard of truth. However, I cannot say if this is actually true or not (in the sense of describing an empirical reality in which the work of scientists [so-called] is compared with the work of non-scientists). Lots of what scientists say turns out to be wrong and a good deal of it is highly uncertain at the time of publication. Indeed, this is what defines a frontier of knowledge.

By contrast, I would propose that what distinguishes science from non-science is the serious attention paid to uncertainty estimates. We expect scientists to tell us not only what their best guess is but also how good a guess it is likely to be. Journalists and other popular prognosticators are unlikely to dwell on this aspect of a question, at least not in a self-conscious and explicit fashion.

Let me press further on this issue, for I suspect that there is also an important distinction with respect to uncertainty between natural science and social science.

Within most natural-science fields, my sense is that there is usually general agreement about the relative (un)certainty of a finding. This stems from the fact that the methods of analysis in use within a field are fairly limited. Whether the field is experimental or non-experimental, whether it is dominated by data or by mathematical models, there are usually a few common methods that all practitioners employ, or are at least intimately familiar with. Consequently, it is usually fairly clear what standards ought to apply and what levels of uncertainty each finding implies.

Within the social sciences, our predicament is that the toolkit of available methods is, well, virtually limitless—including laboratory experiments and messy observational data, large-N and small-N samples, and so forth. Consequently, in assessing uncertainty one must span a great swath of diverse approaches to social knowledge. This constrains our ability to arrive at—and agree upon—uncertainty estimates. We have no commonly recognized metric to appeal to.

At the same time, arriving at reasonable estimates of overall uncertainty—ones that all practitioners can agree upon — may offer the only hope of attaining greater methodological unity across the diverse methods and disciplines of the social sciences.

Consider the following experience, which I expect many readers have shared. You are asked to review a paper or book that addresses some problem of importance to the discipline. You find the argument and evidence ingenious—the work clearly qualifies as a contribution to knowledge—but also highly dubious. Worse, in his/her effort to convince, the researcher has assumed the guise of a lawyer arguing a case, offering all the reasons why his/her argument might be true and none of the reasons why it might be false. In this situation, my willingness to endorse publication of the manuscript rests on an explicit and reasonable estimate of uncertainty. If the author is willing to oblige, I am willing to open the door. If not, then not.

This brings me to a central point of critique. The discipline is too focused on getting estimates right and not focused enough on getting estimates of uncertainty right. Note that we may never agree on whether democracy causes growth. But we should be able to agree that any causal inferences on this matter are highly uncertain. This, itself, is vital for the progress of the field. And it should allow for new studies of democracy and growth to appear (in top journals), without giving a misleading impression of certainty to unwary policymakers.

Of course, arriving at reasonable uncertainty estimates is not an easy task. Quantitative work with messy observational data is rightly criticized for equating t statistics and confidence intervals with the real (overall) uncertainty of an argument. Qualitative work is not prone to this error. But it is prone to an equally grave problem: uncertainties are often unaddressed, or not explicitly addressed, and there is no standard format for assessing and comparing uncertainty across studies.

Extreme bounds offers an intriguing possibility. However, as Glynn points out, the bounds identified as extreme may be too broad to be of much use. Further assumptions may be required in order to narrow the bounds of a causal proposition to levels that are informative, as discussed at the end of Glynn's essay. And yet doing so brings us back to the central problematic: causal inference often requires assumptions that are not directly testable.

Notes

[1] Bennett et al (2003) find qualitative work well represented in the pages of some leading journals, such as *Comparative Political Studies*, *International Organization*, and *World Politics*, but their data indicate that qualitative work is infrequent in the top general-interest journals, including the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics*, and in articles in the American Government sub-field.

[2] With respect to the models shown in Table 1, I wonder about what is being tested here. Recall that in order to be meaningful, a regression model must replicate some actual data-generating process. What, then, is the process by which a change in (a) GDP rank and (b) GDP residual leads to regime-change (or stasis)? If this question cannot be answered, then the regression model is nonsensical. And if these independent variables do not resemble actual (real-life) interventions then there is no way of interpreting them in a causal fashion. By contrast, the attempt to distinguish between various elements of GDPpc, or elements that the GDPpc term might be proxying for, makes a lot of sense—so long as each element is interpretable in a causal fashion. Thus, scholars have examined the relationship between GDP (without the per capita denominator), population density, urbanization, education, infrastructure, and other modernization variables on the one hand, and regime-type on the other. But these efforts have not, to my knowledge, dethroned the empirical status of GDPpc; they have merely clarified some of the elements within the modernization rubric that might be carrying the causal burden.

References

Acemoglu, Daron, Simon Johnson, James A. Robinson, and Pierre Yared. 2008. "Income and Democracy." *American Economic Review* 98:3, 808–842.

Achen, Christopher H. 1982. *Interpreting and Using Regression*. Beverly Hills: Sage.

Bennett, Andrew, Aharon Barth, and Kenneth R. Rutherford. 2003. "Do We Preach What We Practice? A Survey of Methods in Political Science Journals and Curricula." *PS: Political Science and Politics* 36:3 (July), 373–379.

Causal Homogeneity in Mechanismic Research

Martin Austvoll Nome
University of Oslo & Centre for the Study of Civil War (CSCW), Peace Research Institute Oslo (PRIO)
m.a.nome@stv.uio.no

The long-tailed weasel would seem to be a well-adapted creature, changing as it does the color of its fur to blend with the seasonal landscape.[1] According to the caption below, fur color co-varies with day length. Yet the effect of day length is conditional on latitude. Weasels in the north change color; weasels in the south do not. As the caption so well puts it, "it all depends."[2]

Had the whole population of weasels been compared for the impact of day length on color without regard to whether they lived in the north or in the south, a methodologist might complain that the population suffered from causal heterogeneity. Another methodologist might express regret that there is no mechanism explaining how day length affects fur color.

Had the methodologists shifted their attention from the biology of northern mammals to social outcomes in the world of humans, they would discover a community of methodologists of their own ilk. This paper deals with questions that arise when their two methodological concerns intersect—the concern for causal homogeneity in units of analysis, and the commitment to develop mechanismic explanations of outcomes of interest. What is an appropriate definition of causal homogeneity? How does causal homogeneity matter, and what are the implications for research?

As I discover, understandings of causal homogeneity differ. In response, I outline the differences, derive a working definition of causal homogeneity for mechanismic research, and draw implications for the practice of empirical enquiry. In order to arrive at this, I develop an argument in three parts. First, I distinguish between two different but complementary approaches to causal inference, that of a co-variational logic and that of a mechanismic logic. I then go on to discuss the understanding of causal homogeneity and the advice pertaining to it as it appears in some standard works on methodology. Finally, I show a variety of understandings of causal homogeneity in mechanismic work, move towards a unified definition, and draw implications for research. Heterogeneity in populations of cases matter, whether one is of the co-variational or the mechanismic bent, but as I argue, it matters in different ways.

I-centered Versus M-centered Research

The argument in this paper depends on distinguishing between two different but complimentary approaches to causal inference. One seeks to uncover specific robust co-variates of outcomes of interest. The approach typically seeks leverage for probabilistic statements about the relationship between two or more variables. If one thinks of particular values on independent variable as inputs (I), and particular values on dependent variables as outcomes (O), then the sort of statement sought by this approach is "if I, *ceteris paribus*, then we should expect to see O" (I → O).[3] Such statements are expected to be valid across the population to which the I → O relationship pertains. I call this I-centered research.

The other approach seeks explanation in terms of the mechanisms that generate outcomes of interest. Often the focus is on explaining empirical regularities in terms of their mechanisms (M) and the scope conditions (SC) that enable the operation of these mechanisms and determine their impact. The approach typically seeks leverage for conditional statements about the operation of one or more mechanisms that link inputs to outcomes, "if I and SC are present, then we should expect M to generate O" (I → M → O). Such statements are expected to be valid for all cases where the necessary scope conditions are present, in the population to which the I → O relationship pertains. I call this M-centered research.

As I argue, the appropriate understandings of causal homogeneity in I-centered and M-centered research are at best poor analogies and at worst incommensurable. Accordingly, the assumption of causal homogeneity has different meanings and consequences for the practice of I-centered and M-centered research. In order to establish the premise for this argument it is necessary to define mechanisms and how they relate to variables.

In spite of an ostensibly insurmountable range of definitions of mechanisms (Gerring 2007; Mahoney 2001), it is possible to arrive at a definition that transcends these differences. Gerring (2007: 18) does so when he defines mechanisms as "the pathway or process by which an effect is produced or a purpose is accomplished." Gerring's broad church has its costs in terms of precision, so I introduce here two more specific definitions—both compatible with Gerring's, and both compatible with each other—that highlight aspects of mechanisms that are important to my discussion of the causal homogeneity assumption. First, Mayntz (2004: 241) states that "ontologically speaking, the term 'mechanism' refers to recurrent processes linking specified initial conditions and a specific outcome." Second, Falleti and Lynch (2009: abstract, 1143) define mechanisms as "portable concepts that explain how and why a hypothesized cause, in a given context, contributes to a particular outcome." "Context" is very central to Falleti and Lynch's argument about criteria for mechanisms-based explanations. They define context as "the relevant aspects of a setting (analytical, temporal, spatial, or institutional) in which a set of initial conditions leads (probabilistically) to an outcome of a defined scope and meaning via a specified causal mechanism or set of causal mechanisms" (Falleti and Lynch 2009: 1152).

Consider some important aspects of these definitions: Mayntz includes the *recurrent* nature of mechanisms in her definition. Mechanisms are not unique explanations of particular, one-off social outcomes. Instead, the same mechanism can operate across cases, for example in several cases of revolution, democratization, or the adoption of better human rights practices.

Furthermore, to Falleti and Lynch mechanisms are *por-*

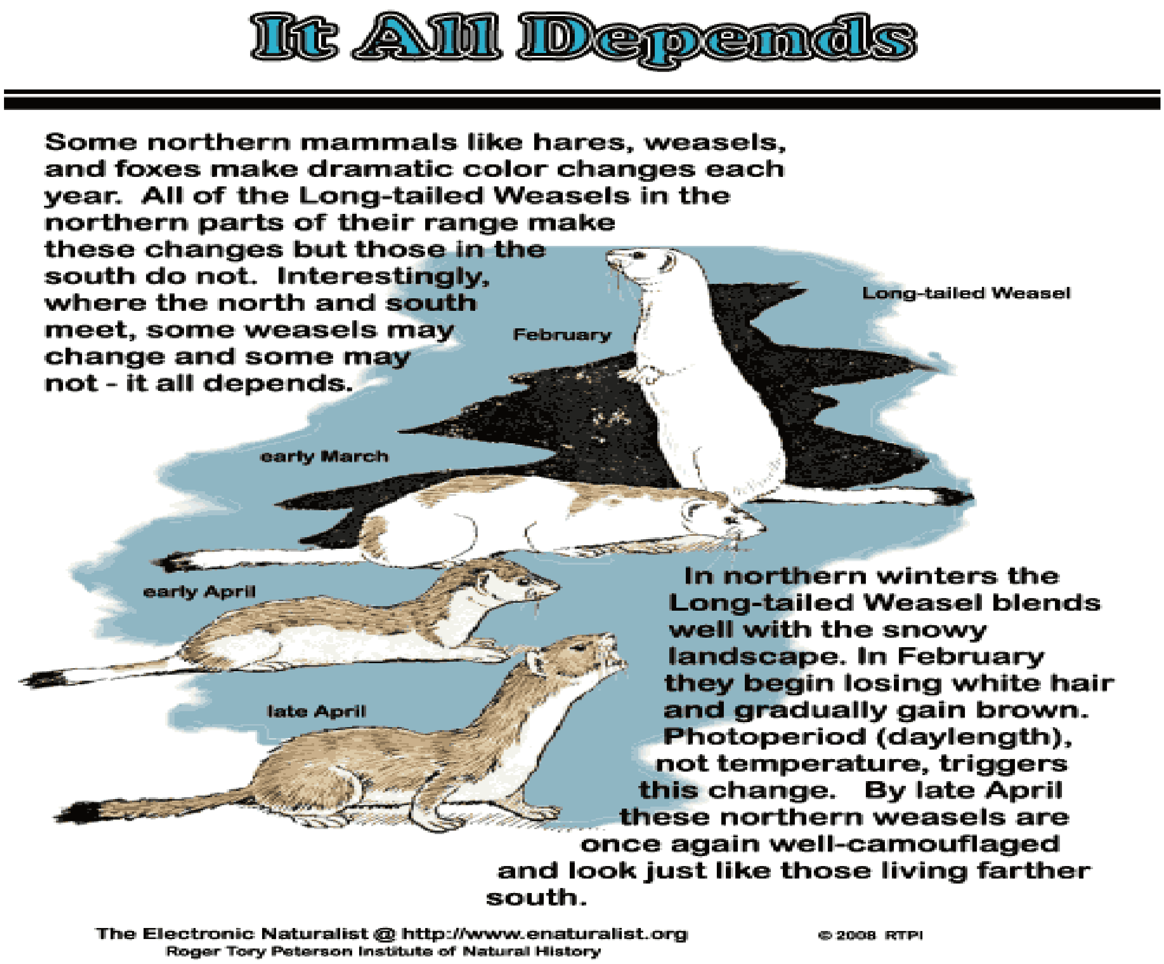


table. The same mechanism can generate diverse social and political phenomena—that is, different I → O relationships. For example, Hedström and Swedberg (1998: 21) discuss a particular belief-formation mechanism by which "the number of individuals who perform a certain act signal to others the likely value or necessity of the act," thereby influencing other individuals' choice of action. This belief-formation mechanism, they argue, explains phenomena as different as a run on the bank (Merton [1948] 1968), the diffusion of a new drug (Coleman, Katz, and Menzel 1957), and the emergence of collective behavior such as rioting (Granovetter 1978).[4]

To Falleti and Lynch, most mechanisms are also *indeterminate*. Mechanisms cannot in themselves predict outcomes, they can only explain them after the fact. As Falleti and Lynch (2009: 1151) put it, "given an initial set of conditions, the same mechanism operating in different contexts may lead to different outcomes." It is only when mechanisms interact with given contexts that one can predict particular outcomes. In other words, it is the context that does the predicting, not the mechanism.

Finally, mechanisms have an ontological status that is different from variables. This is an important point to make in order to appreciate how mechanisms and variables can complement each other in causal inference. Regretably, King, Keohane, and Verba (1994: 86) conflate mechanisms and independent variables; so does Sambanis (2004: 263), and in passing, so does Gerring (2007: 163). Mechanisms should never be mentioned in the same breath as independent variables.[5] Whereas variables are dimensions along which units of analysis can be placed, mechanisms are *relational* and *processual* (Falleti and Lynch 2009: 1149). Units of analysis can have variable values along particular dimensions, be they nominal, ordinal, interval, or ratio. In contrast, mechanisms "describe the relationships or the actions among the units of analysis or in the cases of study" (Falleti and Lynch 2009: 1147). The difference in the ontological status of variables and mechanisms suggests a division of labor between I-centered and M-centered research.

Given the affinity of mechanisms-oriented research with qualitative methods, one would think that standard methodological work such as *Designing Social Inquiry* by King, Keohane, and Verba (1994) and their critics in *Rethinking Social Inquiry* (Brady and Collier 2010) would be the place to begin in order to learn about causal homogeneity. As I show in the following section, however, their understanding of causal homogeneity is only compatible with I-centered research and a co-variational logic of causal inference.

Causal Homogeneity in I-centered Research

Causal Homogeneity in Designing Social Inquiry (DSI)

DSI's oft-cited definition of causal homogeneity is as follows:[6] "Two units are homogenous when the expected values of the dependent variables from each unit are the same when our explanatory variable takes on a particular value" (King, Keohane, and Verba 1994: 91). A less cited and somewhat weaker version of causal homogeneity is the assumption of constant effect. According to this version, units are homogenous when a change in an independent variable leads to the same expected change in the dependent variable, regardless of which value on the dependent variable the units had to begin with (King, Keohane, and Verba 1994: 92–93). Either way, *DSI* offers an understanding of causal homogeneity that rather technically links model and data in a large-N cross-case template.

As it goes on to discuss causal homogeneity, *DSI* makes some points that are worth bearing in mind. If one is interested in estimating causal effects, it says, then causal homogeneity is an important assumption (91). Yet, causal homogeneity is only an assumption; it can be wrong (91). The ostensibly troubling elaboration of this is that the "unit homogeneity assumption is generally untestable" (93). Attaining it is often impossible (93). Still, it would seem to be desirable to somehow investigate the diversity in one's population. For the poor researcher who realizes that causal homogeneity is a difficult criterion to meet, *DSI* offers encouragement. Learn from it, *DSI* suggests: "understanding the degree of heterogeneity in our units of analysis will help us to estimate the degree of uncertainty or likely biases to be attributed to our inferences" (93–94).

Yet applying this to qualitative comparative case studies is not straightforward. The literal interpretation of *DSI*'s advice is the basic point that the standard errors of regression coefficients are greater the more one's units of analysis are scattered around the regression line. To those using regression analysis, this relationship between model and data is an obvious manifestation of causal heterogeneity, and a likely interpretation of *DSI*'s advice given its reliance on a quantitative template. Yet, its stated concern is qualitative research, not quantitative analysis, and so an analogous interpretation of the advice is more appropriate. Here one runs into trouble. An interpretation of the advice as it pertains to qualitative analysis leads one to question what *DSI* means by "degree of heterogeneity" and "estimating the degree of uncertainty" (93–94). Such concepts are impossible to quantify in case study research. It is also difficult to see how *DSI*'s advice can be applied to M-centered research given that its definition of causal homogeneity is tied to the effects of independent variables. The question remains how one is to assess causal heterogeneity in case studies in general and in mechanismic case studies in particular.

Causal Homogeneity in Rethinking Social Inquiry (RSI)

In *Rethinking Social Inquiry* (Brady and Collier 2010), Collier, Seawright, and Munck (2010) take it upon themselves to codify and comment on *Designing Social Inquiry*. According to Collier et al. (2010: 41), *DSI*'s main concern with the causal homogeneity assumption is that any causal model is only appropriate to a particular domain of cases; additional cases may have distinctive causal features that necessitate a more complex causal model. Such general formulations would seem to have relevance for M-centered research. The term "causal model" could refer to a model in the I→M→O form. Yet as *RSI* goes on, it is evident that it argues on *DSI*'s premise of a Humean understanding of causality. For example, it (rightly) advises that "if the causal homogeneity assumption is not met, and a researcher analyzes the data as if it were, the inference will be a misleading average that lumps together differences among subgroups of cases" (Collier et al. 2010: 42). By referring to such faulty inferences as a "misleading average," it suggests that the task Collier et al. have in mind is the correct estimation of causal effects, not the identification of the appropriate causal mechanisms. Like *DSI*, Collier et al. offer advice based on the premise of I-centered research. Then, how would they deal with samples where different causal models apply to different sub-samples? They suggest that for regression analysis, the way is to include an interaction term that effectively separates the different sub-samples, and for qualitative comparison, the way is to conduct separate comparisons for each sub-sample (Collier et al. 2010: 43).

So far, *DSI*'s and *RSI*'s understandings of causal heterogeneity and how to deal with it are coherent. By considering their arguments on a more general level, it is not clear how definitive their solution is.

The Limits of Disaggregation

The methods of introducing an interaction term in regression analyses and separating comparisons for each sub-sample in case study analysis are both special cases of the more general procedure of disaggregation. Yet how far should one disaggregate? Nature, or the social world, provides no natural limits on the process of sub-dividing samples, units of analysis, concepts, mechanisms, or variables. Collier et al. (2010: 21) warn against lumping together sub-groups of cases. Yet, the modeling that social scientists do, involving abstraction and simplification no matter how close to micro-mechanisms one gets, is by nature an exercise in lumping together sub-groups of cases. There is no obvious standard for how far one is to disaggregate, except the current state of theory on the particular question with which one is dealing.

This argument is not unique to I-centered research. Also, mechanisms-based explanations have to deal with questions of aggregation, "how micro?," and the appropriate level of abstraction. Heterogeneity is a feature of any level of aggregation. Disaggregating one step further only transposes the phenomenon of heterogeneity to another level. This is not to deny the value of disaggregation or to invalidate accusations of bad modeling. I only seek to argue that disaggregation provides no final solution to causal heterogeneity.

Causal Homogeneity in M-centered Research

The definition and significance of causal homogeneity may be relatively straightforward for I-centered research. The meaning of causal homogeneity in M-centered research is much less so. Considering that mechanisms are characterized by *recurrence*, *portability*, and *indeterminacy*, one might imagine at least three different meanings of homogeneity. Since I wish to arrive at one working definition of causal homogeneity, let me in the meantime use the loose concept of "sameness."

1: Same Mechanism, Different Phenomena

First, because mechanisms are recurrent and portable, the same mechanism can generate a variety of social phenomena. Formally, the same mechanism can explain different I$\rightarrow$O relationships (see Figure 1).

Figure 1: The Same Mechanism Explaining Different Social Phenomena

$$I_1 \longrightarrow M \longrightarrow O_1$$
$$I_2 \longrightarrow M \longrightarrow O_2$$
$$\ldots$$
$$I_{n-1} \longrightarrow M \longrightarrow O_{n-1}$$
$$I_n \longrightarrow M \longrightarrow O_n$$

Take the mechanism of framing as an example. Framing is the strategic adaptation of information for a target audience in order to make desired behavior more likely. As scholars of social movements put it, "by rendering events or occurrences meaningful, frames function to organize experience and guide action, whether individual or collective" (Snow, Rochford, Worden, and Benford 1986: 464). As a causal mechanism, framing is portable because it can explain such different social phenomena as participation in social movement organizations (Snow et al. 1986), the non-intervention of Great Britain in the American Civil War (Steele 2005), and the constitution of a Kurdish diaspora in Germany (Lyon and Ucarer 2001). The sameness among these different instances is presumably in the aspects of the phenomena that allow framing to operate and generate an outcome. Would this be a sensible meaning of causal homogeneity? The answer is not obvious, considering that there are at least two more ways in which sameness is manifested in mechanismic explanations.

2: Same Phenomenon, Same Mechanism, Different Outcomes

Second, because mechanisms are indeterminate, sameness can manifest itself when the same mechanism operates between a given independent and dependent variable, yet generating different outcomes across cases. This is possible when the impact of a mechanism depends on unobserved conditions that vary across cases. Here, the mechanism is constant, but outcomes are not (see Figure 2).

Figure 2: The Same Mechanism Generating Different Outcomes on the Same Dependent Variable

$$I \longrightarrow M \longrightarrow (O_A \text{ or } O_B)$$

For military interventions in civil conflicts, the mechanism of framing can explain both intervention and non-intervention. Consider Gagnon's (2004: 93–96) account of the Serbian intervention in Croatia, 1991–1995. To justify the intervention, the Serbian leadership used its control of the mass media to frame the violence it imposed on the Croatian police and moderate Serbs as extreme injustices suffered by Yugoslav Serbs living outside Serbia. As Gagnon (2004: 96) describes the way violence was framed, "the discourse itself was structured around injustices that Serbs outside of Serbia were suffering only because they were Serbs, at the hands of crazed, bloodthirsty extremists defined in ethnic and/or religious terms." In an illustration of how framing can cut both ways, Steele (2005) argues that British non-intervention in the American Civil War can be explained by a re-framing of the war that occurred in the period 1862–1863. The decisive factor was President Abraham Lincoln's Emancipation Proclamation. Steele suggests that Lincoln used the Emancipation Proclamation in part to thwart foreign intervention by re-framing the conflict in terms of liberation. In the minds of the British, "the Union went from an army of preservation to an army of liberation. The Confederacy went from an oppressed society to one constituted by the enslavement of four million people" (Steele 2005: 532). As Steele traces a shift in the way Foreign Secretary Russell and Chancellor of the Exchequer William Gladstone argued about a prospective intervention, he finds evidence suggesting that "a British intervention prior to the Emancipation Proclamation could have still been consistent with British self-identity, but an intervention following the Emancipation Declaration would most certainly have not. Thus, the changed meaning of the Civil War from 'Northern Aggression' to 'liberation' meant that any intervention would threaten British identity" (Steele 2005: 521). The sameness among these cases is in the social phenomenon to be explained, and in the mechanism mustered to do the explaining. Is it a candidate definition of causal homogeneity? Consider first the last manifestation of sameness in mechanismic explanations.

3: Same Phenomenon, Same Outcome, Different Mechanisms

The third way in which sameness manifests itself in M-centered explanations is in focusing both on the same independent and dependent variable *and* on the same I$\rightarrow$O relationship. In other words, the interest is in explaining a particular outcome on the dependent variable. Here, the same I$\rightarrow$O relationship can be explained by different mechanisms or combinations of mechanisms (see Figure 3).

Figure 3: Different Mechanisms Explaining the Same Social Phenomenon

$$I \longrightarrow (M_1 \text{ or } M_2 \text{ or } \ldots \text{ or } M_{n-1} \text{ or } M_n) \longrightarrow O$$

Take for example the international diffusion of liberalism. Here, the input in the I → O relationship is the adoption of liberal policies by one government, and the outcome in the I → O relationship is the adoption of liberal policies by another government where its adoption was systematically conditioned by policy choices in the former country (Simmons, Dobbin, and Garrett 2006: 787). Simmons et al. (2006) and contributors to their symposium (2006) explain the international diffusion of liberalism with four different mechanisms—coercion, competition, learning, and emulation. Another example of the use of several different mechanisms to explain a particular outcome is Johnston's (2008) work on the socialization of Chinese diplomats in international security institutions. In order to explain how the attitudes of Chinese participants changed through interactions with other diplomats in small group settings, Johnston draws on the mechanisms of mimicking, social influence, and persuasion.

Towards a Working Definition of Causal Homogeneity for Mechanismic Explanations

The three different versions of sameness in mechanismic explanations all derive from the recurrent, portable, and indeterminate nature of mechanisms. What say the methodologists? Falleti and Lynch (2009) write on causal homogeneity with respect to mechanismic explanations. They state that

> …unit homogeneity in mechanismic explanations requires that mechanisms, and not just variables, be portable and comparable across contexts. To be analytically equivalent (i.e., homogenous) for comparative purposes, these contexts must possess similar values of the attributes that are likely to affect the functioning or meaning of the mechanisms that are involved in the causal process (1145).

This understanding of homogeneity is closely analogous to the first notion of sameness above. There, the same mechanism operates on different arenas because the contexts are analytically equivalent, that is, homogenous. It is perhaps unnecessarily indirect to describe these contexts as having "similar values of the attributes that are likely to affect the… mechanisms" (see quotation above). One might simply say that the contexts are characterized by the same scope conditions.

Scope conditions are of great importance in mechanismic explanations. Scope conditions are "the relevant aspects of a setting (analytical, temporal, spatial, or institutional) in which a set of initial conditions leads (probabilistically) to an outcome of a defined scope and meaning via a specified causal mechanism or set of causal mechanisms" (Falleti and Lynch 2009: 1152). In other words, scope conditions are the characteristics of the cases of a particular I → O relationship that enable mechanisms to operate and determine what outcome they generate. Whereas mechanisms in themselves can only explain outcomes after the fact, scope conditions do the heavy lifting in terms of prediction.

An example of M-centred research that explicitly invokes scope conditions is Checkel's (2005) work on the socialization of bureaucrats in European international institutions. One mechanism of socialization is normative suasion, a mechanism by which agents "present arguments and try to persuade and convince each other," and by which agents "actively and reflectively internalize new understandings of appropriateness" (Checkel 2005: 812). Checkel specifies five scope conditions under which normative suasion is more likely to change the interests of agents. One example of a scope condition is that "the target of the socialization attempt is in a novel and uncertain environment and thus cognitively motivated to analyze new information"; another that "the target has few prior, ingrained beliefs that are inconsistent with the socializing agency's message" (Checkel 2005: 813). These examples illustrate that scope conditions can be operationalised and measured as variables. Scope conditions need not be dichotomous; they can in principle take any appropriate level of measurement. Of the two scope conditions for normative suasion, the "novel environment" condition could well be measured dichotomously, whereas the condition of "few prior beliefs inconsistent with socializer's message" should perhaps be measured at a higher level.

Scope conditions are variables, but the explanatory work they do does not necessarily overlap easily with independent variables. Scope conditions enable mechanisms to operate and determine their impact. Mechanismic explanations are incomplete without them. With respect to my argument on causal homogeneity in M-centered research, scope conditions have pride of place.

In moving towards a definition of causal homogeneity for mechanismic explanations, I assume that empirical research is normally focused on one social phenomenon, or one independent and dependent variable. This is a sensible assumption given how research questions are normally delimited, and given how policy questions are organized. I also assume that research is normally focused on one particular outcome on the dependent variable, that is one I → O relationship. Then, causal homogeneity should pertain to all units of analysis that are cases of the particular I → O relationship.

Given the assumption that one deals with a particular I → O relationship, and given the assumption of the importance of scope conditions, I propose defining causal homogeneity in M-centered research thus: Cases are homogenous to the extent that they display the same scope conditions.

Implications for Research

The definition of causal homogeneity in M-centered research as the extent to which cases display the same scope conditions is simple, but it has important implications for research. The assumption of causal homogeneity in I-centered research does not have a meaningful analogy in M-centered research. In I-centered research, where the objective is to uncover generally valid co-variation between independent and dependent variables, the assumption of causal homogeneity is important when estimating causal effects. If one suspects that a cross-case population is heterogeneous in ways that are relevant for the impact of the independent variable, the solution is to disaggregate. That is, the solution is to muster current theory to identify the subgroups for which effects are different, and organize the modeling around those subgroups to the extent it improves on current theory and meets whatever criteria one has of precision and parsimony. In short, for I-centered research, heterogeneity between units of analysis is a threat to inference.

The assumption of causal homogeneity does not have a meaningful analogy in M-centered research because mechanismic statements that specify scope conditions are valid no matter how few or how many cases they apply to. Instead, the degree of causal homogeneity in a population is an interesting empirical question, and involves mapping the size of the domain for which a particular mechanismic explanation holds. For M-centered research, heterogeneity between units of analysis is not a threat to inference. Mechanismic explanations sit quite comfortably with highly contingent generalization. That being said, mechanismic statements are contingent in one sense, but universal in another.

Mechanismic statements are contingent in that they expect to explain only a limited domain of cases that are characterized by specific scope conditions. This evokes the oft-noted association between mechanismic explanations and middle-range theory.

Mechanismic statements are universal in that they claim that particular mechanisms can always operate when particular scope conditions are present—enabling their operation and determining the outcome—in all contexts that in Falleti and Lynch's (2009) words are "analytically equivalent."

The dual nature of mechanismic statements as both contingent and universal is an appropriate note on which to draw some more specific implications for empirical research. By making the relative heterogeneity among units of analysis in terms of scope conditions into the object of empirical research, the outcome is a procedure that systematizes different middle-range generalizations.

I propose a three-step procedure:

1. Muster theories of the mechanisms and scope conditions that might explain a particular I$\rightarrow$O relationship.

2. Test these mechanisms and their scope conditions by their observable implications on cases from the I$\rightarrow$O population.

3. When one is sufficiently confident that particular mechanisms usually operate to produce an outcome of interest under specific scope conditions, then measure these scope conditions in the remaining cases. At best, scope conditions are so closely associated with the operation and impact of particular mechanisms that knowing which mechanisms generate outcomes becomes a matter of mapping the scope conditions that characterize the remaining cases. As opposed to mechanisms, scope conditions have an ontological status as variables, and are therefore easier to measure.

The ideal here is to find empirical leverage for linking particular scope conditions to the operation of particular mechanisms. It is not necessary to think of heterogeneity as a problem for inference as long as that heterogeneity is known in terms of the different scope conditions that characterize cases. When mechanisms operate under known scope conditions, then heterogeneity among cases simply suggests that their outcomes are generated by different mechanisms. In M-centered research that specifies scope conditions, heterogeneity is not a problem to be solved by introducing interaction terms or separate comparisons as *RSI* (Collier et al. 2010: 43) suggests in its I-centered framework. Heterogeneity is instead a source of knowledge about how mechanisms or combinations of mechanisms relate to their scope conditions, and how these are distributed in the population of cases. When *DSI* (King et al. 1994: 93–94) rightly argues that one ought to understand the degree of heterogeneity in one's cases, its objective is to estimate uncertainty and bias in causal effects. Here, rather, the purpose of understanding heterogeneity is to assess the extent to which outcomes across cases of a particular I$\rightarrow$O relationship are generated by different mechanisms.

The outcome of this three-step procedure is the cumulation of knowledge in terms of both theory and empirics. In terms of theory, one learns more about mechanisms and how they interact with their context. In terms of empirics, one becomes acquainted with the population of cases in a way that suggests which mechanisms operate to produce their outcomes, be it different mechanisms or combinations of mechanisms across different cases of the same I$\rightarrow$O relationship.

Conclusion

The burgeoning literature on social mechanisms testifies to a growing interest in making hypothesized mechanisms into the objects of empirical investigation (see, for example, Elster 2007; Falleti and Lynch 2009; George and Bennett 2004; Gerring 2007; Hedström and Swedberg 1998; Hovi 2004; Johnson 2002, 2006; Mahoney 2001; Mayntz 2004; McAdam, Tarrow, and Tilly 2008; and Tilly 2001). For those of us whose research practices are still being formed, and who might wish to focus on mechanisms, there is a need to acquire principles on the specifics of mechanismic research. Standard work such as *Designing Social Inquiry* (King, Keohane, and Verba 1994) and *Rethinking Social Inquiry* (Brady and Collier 2010) cannot sufficiently offer such principles. As I argue, the way such work treats the central methodological concept of causal homogeneity is difficult to apply to mechanismic research. Either there is no consistent understanding of the meaning of causal homogeneity as it applies to mechanismic explanations, or the concept is incommensurable with any candidate definition because it is based on the assumption that causal theories favor variables over mechanisms. The frequency with which *DSI* and *RSI* appear in syllabi of graduate courses in qualitative methods suggests the need for a more specialized literature on mechanismic research.[7]

Notes

[1] I gratefully acknowledge Øivind Bratberg, Jeff Checkel, and Anders Ravik Jupskås for helpful comments on earlier drafts of the paper. Special thanks go to Robert Adcock for his editorial advice when preparing the paper for *QMMR*.

[2] The diagram is reproduced with permission from the Roger Tory Peterson Institute of Natural History, Jamestown NY. (http://www.

rtpi.org/it-all-depends.html).

[3] The immediate source of the I → O model is Falleti and Lynch (2009), but one suspects that their inspiration is Hedström and Swedberg (1998: 7–10), who in turn refer to work by Bunge (1967).

[4] See Hedström & Swedberg (1998: 17–21) for an elaboration of this.

[5]Waldner (2010) puts this at least as strongly. To him, "mechanisms embody *invariance*" (Waldner 2010: 31, emphasis in original).

[6] *DSI* uses the term "unit homogeneity." I have instead adopted *RSI*'s preference for the term "causal homogeneity." It is less laden with the former term's roots in statistics, and more useful when applying it to mechanismic work. See how Collier, Seawright, and Munck (2010: 41–42) make the case for "causal homogeneity."

[7] Of the 51 Political Science graduate course syllabi in qualitative methods listed in the database of the Consortium on Qualitative Research Methods, 39 syllabi feature *DSI* and/or *RSI* among the readings (http://www.maxwell.syr.edu/moynihan/cqrm/Syllabi_Database/). Granted, some syllabi clearly position *DSI* for major criticism. See, for example, Ido Oren's "Interpretive Methods to Political Science" (http://www.maxwell.syr.edu/uploadedFiles/moynihan/cqrm/Syllabus_%20Spring_%202009.pdf) or Lisa Wedeen's "Interpretive Methods in the Social Sciences" (http://www.maxwell.syr.edu/uploaded Files/moynihan/cqrm/wedeen.pdf).

References

Brady, Henry E. and David Collier, eds. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield.

Bunge, Mario. 1967. "Scientific Research." *Studies of the Foundations, Methodology, and Philosophy of Science*, Volume 3. Berlin: Springer-Verlag.

Checkel, Jeffrey T. 2005. "International Institutions and Socialization in Europe: Introduction and Framework." *International Organization* 59, 801–826.

Coleman, James, Elihu Katz, and Herbert Menzel. 1957. "The Diffusion of an Innovation Among Physicians." *Sociometry* 20:4, 253–270.

Collier, David, Jason Seawright, and Gerardo L. Munck. 2010. *The Quest for Standards: King, Keohane, and Verba's Designing Social Inquiry*. In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Henry E. Brady and David Collier, eds. (Lanham, MD: Rowman & Littlefield), 33–63.

Elster, Jon. 2007. *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. New York: Cambridge University Press.

Falleti, Tulia G. and Julia F. Lynch. 2009. "Context and Causal Mechanisms in Political Analysis." *Comparative Political Studies* 42:9, 1143–1166.

Gagnon, V. P. 2004. *The Myth of Ethnic War: Serbia and Croatia in the 1990s*. Ithaca: Cornell University Press.

George, Alexander L. and Andrew Bennett. 2004. *Case Studies and Theory Development in the Social Sciences*. Cambridge: MIT Press.

Gerring, John. 2007. "Review Article: The Mechanismic Worldview: Thinking Inside the Box." *British Journal of Political Science* 38, 161–179.

Granovetter, Mark. 1978. "Threshold Models of Collective Behavior." *American Journal of Sociology* 83:6, 1420–1443.

Hedström, Peter and Richard Swedberg. 1998. "Social Mechanisms: An Introductory Essay." In *Social Mechanisms: An Analytical Approach to Social Theory.* Peter Hedström and Richard Swedberg, eds. (New York, NY: Cambridge University Press), 1–31.

Hovi, Jon. 2004. "Causal Mechanisms and the Study of International Environmental Regimes." In *Regime Consequences: Methodological Challenges and Research Strategies*. Arild Underdal and Oran R. Young, eds. Dordrecht: Kluwer Academic Publishers.

Johnson, James. 2002. "How Conceptual Problems Migrate: Rational Choice, Interpretation, and the Hazards of Pluralism." *Annual Review of Political Science* 5, 223–248.

Johnson, James. 2006. "Consequences of Positivism: A Pragmatist Assessment." *Comparative Political Studies* 39:2, 224–252.

Johnston, Alastair Iain. 2008. *Social States: China in International Institutions, 1980–2000*. Princeton: Princeton University Press.

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

Lyon, Alynna J. and Emek M. Ucarer. 2001. "Mobilizing Ethnic Conflict: Kurdish Separatism in Germany and the PKK." *Ethnic and Racial Studies* 24:6, 925–948.

Mahoney, James. 2001. "Beyond Correlational Analysis: Recent Innovations in Theory and Method." *Sociological Forum* 16:3, 575–593.

Mayntz, Renate. 2004. "Mechanisms in the Analysis of Social Macro-Phenomena." *Philosophy of the Social Sciences* 34:2, 237–259.

McAdam, Doug, Sidney Tarrow, and Charles Tilly. 2008. "Methods for Measuring Mechanisms of Contention." *Qualitative Sociology* 31, 307–331.

Merton, Robert K. ([1948] 1968). "The Self-Fulfilling Prophecy." In *Social Theory and Social Structure*. (New York, NY: The Free Press), 475–490.

Sambanis, Nicholas. 2004. "Using Case Studies to Expand Economic Models of Civil War." *Perspectives on Politics* 2:2 (June), 259–279.

Simmons, Beth A., Frank Dobbin, and Geoffrey Garrett. 2006. "Introduction: The International Diffusion of Liberalism." *International Organization* 60, 781–810.

Snow, David A., E. Burke Rochford, Jr., Steven K. Worden, and Robert D. Benford. 1986. "Frame Alignment Processes, Micromobilization, and Movement Participation." *American Sociological Review* 51, 464–481.

Steele, Brent J. 2005. "Ontological Security and the Power of Self-Identity: British Neutrality and the American Civil War." *Review of International Studies* 31, 519–540.

Symposium. 2006. "Diffusion of Liberalism." *International Organization* 60, 781–934.

Tilly, Charles. 2001. "Mechanisms in Political Processes." *Annual Review of Political Science* 4, 21–41.

Waldner, David. 2010. "What are Mechanisms and What are They Good For?" *Qualitative & Multi-Method Research* 8:2, 30–34.

Book Notes

Book descriptions are excerpted from publishers' websites. If you would like to recommend a book to be included in this section, email Joshua C. Yesnowitz, the production editor of QMMR*, at* jcyesnow@bu.edu.

Calise, Mauro and Theodore J. Lowi. 2010. *Hyperpolitics: An Interactive Dictionary of Political Science Concepts*. Chicago: University of Chicago Press.

Fifteen years in the making, *Hyperpolitics* is an interactive dictionary offering a wholly original approach for understanding and working with the most central concepts in political science. Designed and authored by two of the discipline's most distinguished scholars, its purpose is to provide its readers with fresh critical insights about what informs these political concepts, as well as a method by which readers—and especially students—can unpack and reconstruct them on their own.

International in scope, *Hyperpolitics* draws upon a global vocabulary in order to turn complex ideas into an innovative teaching aid. Its companion open-access website (www.hyperpolitics.net) has already been widely acknowledged in the fields of education and political science and will continue to serve as a formidable hub for the book's audience. Much more than a dictionary and enhanced by dynamic graphics, *Hyperpolitics* introduces an ingenious means of understanding complicated concepts that will be an invaluable tool for scholars and students alike.

Levitsky, Steven and Lucan Way. 2010. *Competitive Authoritarianism: Hybrid Regimes after the Cold War*. Cambridge: Cambridge University Press.

Based on a detailed study of 35 cases in Africa, Asia, Latin America, and post-communist Eurasia, this book explores the fate of competitive authoritarian regimes between 1990 and 2008. It finds that where social, economic, and technocratic ties to the West were extensive, as in Eastern Europe and the Americas, the external cost of abuse led incumbents to cede power rather than crack down, which led to democratization. Where ties to the West were limited, external democratizing pressure was weaker and countries rarely democratized. In these cases, regime outcomes hinged on the character of state and ruling party organizations. Where incumbents possessed developed and cohesive coercive party structures, they could thwart opposition challenges, and competitive authoritarian regimes survived; where incumbents lacked such organizational tools, regimes were unstable but rarely democratized.

Yanow, Dvora and Peregrine Schwartz-Shea. 2011. *Interpretive Approaches to Research Design: Concepts and Processes*. London: Routledge.

Research design is fundamentally central to all scientific endeavors, at all levels and in all institutional settings. This book is a practical, short, simple, and authoritative examination of the concepts and issues in interpretive research design, looking across this approach's methods of generating and analyzing data. It is meant to set the stage for the more "how-to" volumes that will come later in the *Routledge Series on Interpretive Methods*, which will look at specific methods and the designs that they require. It will, however, engage some very practical issues, such as ethical considerations and the structure of research proposals.

Interpretive research design requires a high degree of flexibility, where the researcher is more likely to think of "hunches" to follow than formal hypotheses to test. Yanow and Schwartz-Shea address what research design is and why it is important, what interpretive research is and how it differs from quantitative and qualitative research in the positivist traditions, how to design interpretive research, and the sections of a research proposal and report. It demonstrates how interpretive researchers engage with "world-making," context, systematicity and flexibility, reflexivity, investigation, bottom-up concepts, and explanatory description.

Article Notes

Ahram, Ariel L. 2011. "The Theory and Method of Comparative Area Studies." *Qualitative Research* 11:1 (February), 69–90.

Though many now downplay the tension between area studies and disciplinary political science, there has been little substantive guidance on how to accomplish complementarity between their respective approaches. This article seeks to develop the idea of comparative area studies (CAS) as a rubric that maintains the importance of regional knowledge while contributing to general theory building using inductive intra-regional, cross-regional, and inter-regional comparison. Treating regions as theoretically-grounded analytical categories, rather than inert or innate geographical entities, can help inform both quantitative and qualitative attempts to build general theory.

Baumgartner, Michael. 2009. "Uncovering Deterministic Causal Structures: A Boolean Approach." *Synthese* 170:1, 71–96.

While standard procedures of causal reasoning as procedures analyzing causal Bayesian networks are custom-built for (non-deterministic) probabilistic structures, this paper introduces a Boolean procedure that uncovers deterministic causal structures. Contrary to existing Boolean methodologies, the procedure advanced here successfully analyzes structures of arbitrary complexity. It roughly involves three parts: first, deterministic dependencies are identified in the data; second, these dependencies are suitably minimalized in order to eliminate redundancies; and third, one or—in case of ambiguities—more than one causal structure is assigned to the minimalized deterministic dependencies.

Capoccia, Giovanni and Daniel Ziblatt. 2010. "The Historical Turn in Democratization Studies: A New Research Agenda for Europe and Beyond." *Comparative Political Studies* 43:8–9 (August/September), 931–968.

This article lays the theoretical and methodological foundations of a new historically minded approach to the comparative study of democratization, centered on the analysis of the creation, development, and interaction of democratic institutions. Historically, democracy did not emerge as a singular coherent whole but rather as a set of different institutions, which resulted from conflicts across multiple lines of social and political cleavage that took place at different moments in time. The theoretical advantage of this approach is illustrated by highlighting the range of new variables that come into focus in explaining democracy's emergence. Rather than class being the single variable that explains how and why democracy came about, scholars can see how religious conflict, ethnic cleavages, and the diffusion of ideas played a much greater role in Europe's democratization than has typically been appreciated. Above all, the authors argue that political parties were decisive players in how and why democracy emerged in Europe and should be at the center of future analyses.

Hedström, Peter and Petri Ylikoski. 2010. "Causal Mechanisms in the Social Sciences." *Annual Review of Sociology* 36, 49–67.

During the past decade, social mechanisms and mechanism-based explanations have received considerable attention in the social sciences as well as in the philosophy of science. This article critically reviews the most important philosophical and social science contributions to the mechanism approach. The first part discusses the idea of mechanism-based explanation from the point of view of philosophy of science and relates it to causation and to the covering-law account of explanation. The second part focuses on how the idea of mechanisms has been used in the social sciences. The final part discusses recent developments in analytical sociology, covering the nature of sociological *explananda*, the role of theory of action in mechanism-based explanations, Merton's idea of middle-range theory, and the role of agent-based simulations in the development of mechanism-based explanations.

Kreuzer, Marcus. 2010. "Historical Knowledge and Quantitative Analysis: The Case of the Origins of Proportional Representation." *American Political Science Review* 104:2, 369–392.

Political scientists commonly draw on history but often do not read actual historians carefully. This limited engagement with historians, and with contextual information more generally, contributes to a loss of historical knowledge that can undermine the validity of quantitative analysis. This article makes this argument by means of an examination of the qualitative evidence underlying the important quantitative arguments about the origins of electoral systems advanced by Carles Boix and by Thomas Cusack, Torben Iversen, and David Soskice. The article explores how their respective attention to historical knowledge affects the quality of their data, the plausibility of their hypotheses, and, ultimately, the robustness of their statistical findings. It also analyzes how such knowledge sheds new light on the causal direction between institutions and their economic effects.

Mahoney, James. 2010. "After KKV: The New Methodology of Qualitative Research." *World Politics* 62:1 (January), 120–147.

This article discusses developments in the field of qualitative methodology since the publication of King, Keohane, and Verba's (KKV's) *Designing Social Inquiry*. Three areas of the new methodology are examined: (1) process tracing and causal-process observations; (2) methods using set theory and logic; and (3) strategies for combining qualitative and quantitative research. In each of these areas, the article argues, the new literature encompasses KKV's helpful insights while avoiding their most obvious missteps. Discussion focuses especially on contrasts between the kind of observations that are used in qualitative versus quantitative research, differences between regression-oriented approaches and those based on set theory and logic, and new approaches for bringing out complementarities between qualitative and quantitative research. The article concludes by discussing research frontiers in the field of qualitative methodology.

Montgomery, Alexander H. and Scott D. Sagan. 2009. "The Perils of Predicting Proliferation." *Journal of Conflict Resolution* 53:2 (April), 302–328.

Studies of nuclear proliferation share five serious problems. First, nuclear programs' initiation and completion dates are ambiguous and difficult to code, but findings are rarely subjected to sufficient robustness tests using alternative codings. Second, independent variables overlook important factors such as prestige and bureaucratic power and often use poor proxies for concepts such as the nonproliferation regime. Third, methodologies and data sets should be tightly coupled to empirical questions but are instead often chosen for convenience. Fourth, some findings provide insights already known or believed to be true. Fifth, findings can ignore or gloss over data crucial for policy making and wider debates. This article reviews new quantitative research on nuclear proliferation, noting improved analysis and lingering problems. It highlights the 1999 Kargil war to explore dangers of relying on stock data sets and the need for research on statistical outliers. It concludes with a future research agenda aimed at correcting problems and a cautionary note regarding hasty application of quantitative results to policy making.

Paluck, Elizabeth Levy. 2010. "The Promising Integration of Qualitative Methods and Field Experiments." *Annals of the American Academy of Political and Social Science* 628:1 (March), 59–71.

Randomized field experiments should take a more central place in qualitative research. Although field experimentation is often considered a quantitative enterprise, this paper illustrates the compatibility of field experimentation with various types of qualitative measurement tools and research questions. Integrating qualitative and quantitative data within field experiments allows investigators to move past simple average treatment effects and explore mechanisms of the identified causal effect. A more novel proposal is to use field experimentation as the organizing methodological framework for archival, ethnographic, or interpretive work, and to use ethnographic methods as the primary source of measurement in "experimental ethnography." Sustained research and theoretical specificity can address some of the seemingly incompatible features of qualitative and field experimental methods. For example, small sample sizes are acceptable as part of a research program, and some theories of historical patterns or rare events could be disaggregated into smaller cause-and-effect linkages to test with field experiments in theoretically relevant contexts.

Rohwer, Götz. 2011. "Qualitative Comparative Analysis: A Discussion of Interpretations." *European Sociological Review* 27, forthcoming.

The article discusses interpretations of "Qualitative Comparative Analysis" (QCA) proposed by Charles Ragin. The first section argues that QCA can be understood alternatively as a method of data description or as a method for the construction of deterministic functional models. It is shown that thinking in terms of models is required for generalizations. The second section discusses causal interpretations of such models. It is argued that one can use deterministic models without supposing a deterministic metaphysics. The third section briefly introduces stochastic functional models and shows how they can be used for QCA applications. In addition to showing that QCA can be well understood as a specific method of model construction, the article argues that deterministic and stochastic functional models are quite similar and, depending on the application context and the available data, both kinds of models could be useful.

Announcements

Call for Nominations: The David Collier Mid-Career Achievement Award

The Award honors David Collier's contributions—through his research, graduate teaching, and institution-building—as a founder of the qualitative and multi-method research movement in contemporary political science. The award will be presented annually to a mid-career political scientist to recognize distinction in methodological publications, innovative application of qualitative and multi-method approaches in substantive research, and/or institutional contributions to this area of methodology.

To be eligible for the mid-career award, nominees must have defended their dissertation within fifteen years of the beginning of the year in which the award is presented. For the 2011 award, nominees must have defended their dissertation in or later than 1996.

Each nomination must include a cover letter summarizing specific merits of the candidate (not merely generic praise), as well as an up-to-date curriculum vitae of the nominee, including the date of the doctoral degree. Self-nominations are welcome.

Nominations for the 2011 award should be submitted by June 15, 2011, via email to the President of the APSA Organized Section for Qualitative and Multi-Method Research, currently Colin Elman (celman@maxwell.syr.edu).

The David Collier Mid-Career Achievement Award has been established by the Consortium for Qualitative Research Methods (the co-host of the Institute for Qualitative and Multi-Method Research). By agreement with CQRM, the award is managed by the APSA Organized Section for Qualitative and Multi-Method Research.

Call for Nominations: The *APSR* Qualitative Submissions Award

The APSA Organized Section for Qualitative and Multi-Method Research is pleased to announce the establishment of the Qualitative Submissions to *APSR* Award, for the best qualitative manuscript submitted to the *American Political Science Review* in the calendar year.

The award will be offered in 2011 through 2014, and the winner in each year will receive $2,000. To be eligible:

(1) the manuscript need only be submitted to (not necessarily published in) the journal;

(2) the manuscript needs to have been submitted during the calendar year, with the date of submission determined by the acknowledgement email from the *APSR*;

(3) both new and subsequent submissions (e.g., resulting from an invitation to submit *de novo* or to revise and resubmit) are eligible for the award, but only one version of the manuscript is eligible for the award in any one calendar year;

(4) the manuscript submitted to the *APSR* must be (a) new research on qualitative methodology *per se*, i.e., a study that introduces specific methodological innovations or that synthesizes and integrates methodological ideas in a way that is in itself a methodological contribution; and/or (b) substantive work that is an exemplar for the application of qualitative methods, or of multi-methods with a substantial qualitative component.

Nominations for the 2011 award should be submitted by January 31, 2012, via email to the President of the APSA Organized Section for Qualitative and Multi-Method Research, who will by then be Gary Goertz (ggoertz@email.arizona.edu). Nominations need to include a copy of the submitted version of the manuscript version, and a copy of the *APSR*'s email acknowledging receipt. Cover letters are optional. Self-nominations are welcome.

2011 ICPSR Summer Program Offers Mixed Methods Workshop

In 2011, the Inter-University Consortium for Political and Social Research (ICPSR) Summer Program in Quantitative Methods of Social Research will offer its first workshop on mixed methods. This workshop will highlight the main elements of the mixed methods approach to empirical research, as well as the advantages that this approach provides in terms of greater insights on interesting phenomena in the social, behavioral, and health sciences. The title of the course is:

Mixed Methods: Approaches for Combining Qualitative and Quantitative Research Strategies

The instructors for this workshop are John W. Creswell (Educational Psychology, University of Nebraska, Lincoln) and Paul Mihas (Odum Institute for Research in Social Science, University of North Carolina, Chapel Hill). The three-day workshop will meet June 6–8, 2011, at the Odum Institute for Research in Social Science, on the campus of the University of North Carolina, Chapel Hill.

The workshop will cover ways to develop a rigorous and persuasive mixed methods project, with particular attention to: the project title; statement of the problem and research purpose; research questions; collecting and analyzing both qualitative and quantitative data; the use of computer software; and interpretation of empirical results. The class will also discuss how philosophical assumptions, worldviews, and theoretical "lenses" can affect the research process. Moving beyond project design considerations, the workshop will also cover how to craft a mixed methods proposal using guidelines from the National Institutes of Health, and how to develop a journal article based upon mixed methods research.

Participants in this workshop should have some prior exposure to the principles and practice of qualitative and quantitative research, at least at the introductory level. Participants are encouraged to bring to the workshop their own mixed method project; this could be a basic idea, a dissertation prospectus, a proposal for funding, or a study that has already been completed. Regardless of its exact nature, the workshop will help the participant develop the project further, using good design principles and techniques.

Registration instructions for this workshop and further information about the ICPSR Summer Program are available on the program website, http://www.icpsr.umich.edu/sumprog. Also, you should feel free to contact the ICPSR Summer Program by e-mail (sumprog@icpsr.umich.edu) or by telephone (734-763-7400) if you have any questions.

Berlin Summer School in Social Sciences: Linking Theory and Empirical Research Berlin, July 17–29, 2011

The Berlin Summer School in Social Sciences: Linking Theory and Empirical Research aims at promoting young researchers by strengthening their methodological understanding in linking theory and empirical research. In a first step, we tackle the key methodological challenges of causation, micro-macro linkage, and concept-building that occur in all research efforts. In a second step, we look at how certain empirical fields deal with these challenges and—by referring to selected empirical studies—what solutions have been found to overcome them. Furthermore, participants are provided with hands-on research advice and have the opportunity to present their own work and approaches to these issues.

The Berlin Summer School is a joint endeavor of two of Germany's leading social science institutions, the Berlin Graduate School of Social Sciences (BGSS) at Humboldt-Universität zu Berlin and the Social Science Research Center Berlin (WZB).

The two-week summer school both attracts internationally renowned scholars and draws on Berlin-based faculty. Among the confirmed international lecturers are Craig Calhoun (New York University), John Gerring (Boston University), Gary Goertz (University of Arizona), Michèle Lamont (Harvard University), Dietrich Rueschemeyer (Brown University), and Jonathan H. Turner (University of California, Riverside).

The international summer school is open to thirty PhD candidates. Due to generous funding by the BGSS and the WZB, there will be no tuition fee. We have applied for additional funding and thus might be able to cover travel and accommodation costs for all participants; a decision is expected soon. Five ECTS Credit Points can be granted upon successful completion.

The call for applications has begun. Application can be submitted online until May 15, 2011.

For additional information please visit our webpage at http://www.berlinsummerschool.de/ or contact Johannes Gerschewski at gerschewski@wzb.eu.

APSA Short Courses Created (or Co-Organized) by Division 46: Qualitative and Multi-Method Research Wednesday, August 31, 2011, Seattle, Washington

Short Course 1: Multi-Method Research

Time: 9:00am–1:00pm
Instructors: David Collier, University of California, Berkeley; Thad Dunning, Yale University; Jason Seawright, Northwestern University

Attention has increasingly focused on how qualitative methods can be linked to other analytic tools, including large-N quantitative analysis and formal modeling. To this end, methodologists have urged scholars to "nest" their case studies within small- to medium-N comparisons, and/or within large-N quantitative analysis.

Given that many political scientists are now convinced that good research necessarily employs multiple methodologies, how can different approaches be combined to maximize analytic leverage? How useful are the multi-method techniques under discussion here? Is it sometimes better to stick with one method, and to focus on using it with great skill? This short course explores alternative examples and strategies of multi-method research, with the goal of addressing these questions.

Short Course 2: Designing and Conducting Field Research

Time: 2:00pm–7:00pm
Instructors: Diana Kapiszewski, University of California, Irvine; Naomi Levy, Santa Clara University

This short course addresses a variety of field methods and data collection techniques, and aims to help scholars hone their empirical research skills. Two foundational premises of the course are that fieldwork "begins" long before one enters the field, and that the way fieldwork is carried out can have important ramifications for data-analysis and theory-generation. Although field methods are usually associated with "studying politics abroad," we discuss techniques that may be applied inside and outside the U.S.

We will begin by examining different "varieties" of fieldwork and exploring how research design and fieldwork intersect—since planning for the effective use of field methods and the efficient collection of data are crucial aspects of overall research design. The bulk of the course will focus on preparing for field research, including both logistical and intellectual planning; and conducting field research, including both conversational (interviewing, oral histories, and focus groups) and non-conversational (using archival sources, collecting documents and ephemera, and ethnographic study) forms of data collection. Finally, we will discuss challenges involved in managing, analyzing, and evaluating data both in and out of the field. Although field methods are usually associated with "studying politics abroad," we discuss techniques that may be applied inside and outside the U.S. The course will include several hands-on activities.

Scholars typically initiate their projects by mapping out their analytic questions, thinking about how they will measure their variables and what evidence they will need to support their claims, and beginning to identify potential sources for the data they hope to collect. Yet even if the research is well planned and adequately funded, obstacles can arise. Key respondents may be unhelpful or unavailable. Valuable archives and other collections of primary materials may be accessible only on a limited basis or may be poorly organized. Data necessary for constructing sampling frames for formal or informal interviewing may simply not exist. Time or money may run out before essential data have been collected.

This short course will help analysts to anticipate and address many of the challenges involved in designing and conducting field research. We discuss strategies that will allow scholars to: (1) convert their research design into a "to get" list; (2) identify and begin to investigate data sources before leaving their home institution; (3) make optimal use of relevant technologies (e-mail, web, cell phones, portable photocopying equipment, scanners, digital cameras, and voice and video recorders); (4) respond to the availability of data not anticipated in the original research design, and to the inaccessibility of data that was originally to be collected; (5) organize and manage the potentially vast quantities of information gathered; (6) establish key contacts and interact constructively with actors of all types in the host community; (7) cope with professionally, politically, and personally uncomfortable situations; (8) make the transition from data collection to data analysis and writing in a timely manner.

Following the end of the formal class at 6pm, the instructors will hold a "workshop" in which short-course participants will have the opportunity to discuss their own research and the design and conduct of their own fieldwork in a smaller-group setting. We encourage students to stay for this more informal conversation, and to bring along questions about their work.

Participants will be provided with document templates that may be useful when carrying out field research, including sample correspondence. The course is valuable for students planning dissertation projects, for scholars who would like to develop or improve their data collection skills, and for those who teach classes on research methods.

APSA Panels/Roundtables Created (or Co-Organized) by Division 46: Qualitative and Multi-Method Research September 1–September 4, 2011, Seattle, Washington

Innovative Methods in Qualitative and Multi-Method Research

Chair: Robert Kaufman Adcock, George Washington University
Participants:
Derek Beach and Rasmus Pedersen, University of Aarhus: "What is Process-Tracing Actually Tracing? Empirical Narratives Versus Investigating Theorized Causal Mechanisms."
Marcus Kreuzer, Villanova University: "Taking Plausibility Seriously: Theory Projection, Historical Knowledge Selection and Theory Construction in Comparative Politics."
Tanja Pritzlaff, University of Bremen: "The Practice of Politics: A Video-Based Ethnography of Policy-Making."
Sean L. Yom, Temple University and Jan Erk, Leiden University: "Comparative-Historical Analysis after Fieldwork: Why Good Theory-Building Requires Methodological Sin."
Philippe Blanchard, University of Lausanne: "Sequence Analysis for Political Science."
Discussant: Robert K. Adcock, George Washington University

Qualitative Investigations of Identity

Chair: TBA
Participants:
Nadia E. Brown, Saint Louis University: "Negotiating Insider/Outsider Status: How Identity Influences What Political Elites Say."
Jeremy Menchik, University of Wisconsin, Madison: "Conceptualizing and Measuring Group Tolerance."
Victor M. Olivieri, University of Florida: "The Changing of a Flag and the Changing of a Nation: Meaning-Making in Spain and How a Soccer Team Changed State Identity."
Alexandra C. Budabin, New School University: "Diasporas for Peace? The Building of Transnational Coalitions in the Host County."
Dessislava Kirilova, Yale University: "Changes of Identity, Changes of Course: The Non-Crisis Kind of Foreign Policy Transformation."
Discussant: Philip G. Roeder, University of California, San Diego

Investigating Conflict Processes

Chair: Derek Beach, University of Aarhus
Participants:
Michael Weintraub, Georgetown University: "Fighting Together: Rebel Group Alliances in Civil War."
Ariel Ahram, University of Oklahoma: "The Origins and Persistence of State-Sponsored Militias: Path Dependent Processes in Third World Military Development."
Marisa L. Porges, University of London, King's College: "The Future of Counterterrorism? A New Approach to Assessing Terrorist Deradicalisation."
Jennifer Lamm, University of Texas, Austin: "Non-Citizen Soldiers, Veterans, and Their Families: Political Rights and Wrongs."
Nahomi Ichino, Adam Glynn, and Kevin Quinn, Harvard University: "Using Mixed Methods for Counterfactual Causal Inference."
Discussant: Derek Beach, University of Aarhus

Religious Motivations for Political Actions

Chair: Bradley A. Thayer, Baylor University
Participants:
Peter Shane Henne, Georgetown University: "The Domestic Politics of International Religious Defamation."
Ramazan Kilinc, Michigan State University: "Opportunity Junctures as Catalysts: Islam, Secularism, and Democratic Consolidation in Turkey."
Saba Ozyurt, University of California, Irvine: "A Mixed-Method Inquiry of Religion and Immigrant Integration: How and Why Triangulation Matters."
Shawn Richard Schulenberg, Marshall University: "Essentially Contested Subjects: Some Epistemological Considerations when Studying Terrorists and Homosexuals."
Discussant: Bradley A. Thayer, Baylor University

Qualitative Comparative Analysis

Chair: Axel Marx, Katholieke Universiteit Leuven
Participants:
Charles L. Mitchell, Grambling State University: "Qualitative Techniques in Political Science: Accepting Unstructured Methods for Knowledge Building."
Johannes Gerschewski, Social Science Research Center Berlin (WZB): "Causation and Qualitative Comparative Analysis."
Daniel Stockemer, University of Connecticut: "Explaining and Predicting Women's Representation in Parliament: The Limitations of Qualitative Comparative Analysis (QCA)."
Aries A. Arugay, Georgia State University: "From Probable to Possible: Explaining Military Coups through Qualitative Comparative Analysis."
Giovanni Capoccia, Oxford University, and Laura Stoker, University of California, Berkeley: "Choosing and Combining Units in Qualitative and Quantitative Comparative Research."
Discussant: Axel Marx, Katholieke Universiteit Leuven

Cross-Cultural and Multi-Method Investigation of Processes of Democratization

Chair: Dr. Candice D. Ortbals, Pepperdine University
Participants:
Beatrice Reaud, American University: "The Political Economy of Local Democracy: Experimental Approaches to Examining Municipal Performance in Mozambique."
Vsevolod Gunitskiy, Columbia University: "Complex Adaptive Systems and the Diffusion of Democratization."
Jennifer Gauck, University of Kent: "Measuring the Impact of the Immeasurable? Methodological Challenges in Evaluating Democracy and Governance Aid."
Scott G Feinstein, University of Florida: "Why Violent Secession? A Multi-Method Analysis of Evidence from the Former Soviet Union."
Yonatan Morse, Georgetown University: "In the Shadow of the Party: Electoral Hegemony in Modern–Day Tanzania."
Discussant: Candice D. Ortbals, Pepperdine University

Qualitative International Relations Research

Chair: Michael Strausz, Texas Christian University
Participants:
James F. Robinson, Instituto Tecnologico Autonomo de Mexico: "What Causes IR? Differentiating Causal Analysis in International Relations."
Peter Shane Henne, Georgetown University: "Assessing Parallel Explanations: A Method for Disentangling Competing Explanations in Qualitative International Relations Research."
Kyle M. Lascurettes, University of Virginia: "The Disorderly State of 'International Order': In Search of an Elusive Concept."
Jelena Subotic, Georgia State University and Ayse Zarakol, Washington and Lee University: "Cultural Intimacy in International Relations."
Clayton J. Cleveland, University of Oregon: "Critical Junctures in U.S. Foreign Policy Toward the UN Security Council."
Discussant: Michael Strausz, Texas Christian University

Process Tracing

Chair: Andrew Bennett, Georgetown University
Participants:
David Waldner, University of Virginia: "Process Tracing and Causal Mechanisms."
Jeffrey T. Checkel, Simon Fraser University and Andrew Bennett, Georgetown University: "Process Tracing: From Philosophical Roots to Best Practices."
Alan M. Jacobs, University of British Columbia: "Process Tracing on Ideational Arguments."
Discussants: Matthew Evangelista, Cornell University; James A. Caporaso, University of Washington

The Contribution of 'The Politics of State Feminism: Innovation in Comparative Research' to a Comparative Theory of Rights

Chair: Mieke Verloo, Radboud University Nijmegen
Participants: Dorothy E. McBride, Florida Atlantic University; Yvonne Galligan, Queen's University Belfast; Valerie Sperling, Clark University; Fiona S. Mackay, University of Edinburgh; Birgit Sauer, University of Vienna; Joyce V. Outshoorn, Leiden University.

Qualitative Approaches to Public Policy Research

Chair: Laura Sjoberg, University of Florida
Participants:
Noam Schimmel, London School of Economics: "Obama's Delicate Balance: Moral and Practical Emphases in Obama's Healthcare Reform Rhetoric."
Laura S. Jensen and Fatima Sharif, Virginia Tech: "Health Disparities, Local Knowledge, and Public Policy."
Hahrie C. Han, RWJ Fellow, Harvard/Wellesley College: "The Role of Civic Organizations in Mobilizing Citizen and Physician Engagement around Health Policy Reform."
Season Hoard, Washington State University: "Rethinking the Politics of Expertise: Examining the Role of Gender Expertise in Public Policy."
Neil J. Kraus, University of Wisconsin, River Falls: "Triggering Events and the Limits of Local Law Enforcement Reform."
Discussant: Laura J. Hatcher, Southern Illinois University

Strategies of Case Selection

Chair: Vera Eva Troeger, University of Essex
Participants:
Gary Goertz, University of Arizona: "Qualitative versus Quantitative Strategies of Case Selection."
Jason Seawright, Northwestern University: "Matching Quantitative Case Selection Procedures with Case-Study Analytic Goals."
Ingo Rohlfing, University of Cologne and Carsten Schneider, Central European University: "Case Selection for Process Tracing in the Presence of Necessary Conditions."
Richard Nielsen and John Sheffield, Harvard University: "Case Selection via Matching."
James Mahoney, Northwestern University and Gary Goertz, University of Arizona: "Process Tracing: A Qualitative Perspective."
Discussants: Vera Eva Troeger, University of Essex; John Gerring, Boston University

Multimethod Research: From Slogan to Toolkit

Chair: David Collier, University of California, Berkeley
Participants: Taylor C. Boas, Boston University; Prerna Singh, Harvard University; Thad Dunning, Yale University; Alison E. Post, University of California, Berkeley; Jason Seawright, Northwestern University.

Global Datasets: New and Forthcoming

(Co-Sponsored with Comparative Politics Division)

Chair: Gerardo L. Munck, University of Southern California
Participants:
John Gerring, Boston University, James Mahoney, Northwestern University, and Natalie Lam, Boston University: "Charting the Past: A Global Historical Database."
Zachary Elkins, University of Texas, Austin, Tom Ginsburg, University of Chicago, and James Melton, University of Illinois at Urbana-Champaign: "Lessons from the Decoding of National Constitutions: The Comparative Constitutions Project."
Susan Dayton Hyde and Nikolay Marinov, Yale University: "National Elections Across Democracy and Autocracy: Which Elections Can be Lost?"
Peter F. Nardulli, University of Illinois, Urbana-Champaign: "Measuring Cross-National Differences in Law-Based Orders."
Discussant: Simon D. Jackman, Stanford University

Author Meets Critics: Man is by Nature a Political Animal: Evolution, Biology, and Politics
(Co-Sponsored with Political Psychology Division)

Chairs: Rose McDermott, Brown University; Pete Hatemi, University of Sydney
Participants: Kevin B. Smith, University of Nebraska, Lincoln; James H. Fowler, University of California, San Diego; John R. Hibbing, University of Nebraska, Lincoln; Michael Bang Petersen, University of Aarhus; James N. Druckman, Northwestern University; Chris Dawes, University of California, San Diego.

New Frontiers: Neuropolicy, Childhood Development, Evolution, and Emotion
(Co-Sponsored with Political Psychology Division)

Chair: Charles L. Mitchell, Grambling State University
Participants:
Simone Abendschön, University of Frankfurt: "Young Children's Democratic Orientations: the Beginning of Citizenship?"
Francis X. Shen, Vanderbilt University Law School: "How Will Neuroscience Affect the Law? The Influence of Issue Framing on Support for Neurolaw and Neuropolicy."
Peter Foley, R. Alvarez, and Ralph Adolphs, California Institute of Technology: "How Emotional Reactions Alter Survey Responses."
Anthony C. Lopez, Brown University: "The Risk Contract of War: Offense and Defense in the Adapted Mind."
Discussant: Elizabeth Suhay, Lafayette College

Crafting Interpretive Research: Theoretical Possibilities and Practical Challenges
(Co-sponsored with Interpretive Methodologies and Methods Related Group)

Chair: Jan Kubik, Rutgers University, New Brunswick
Participants:
Devorah Manekin, University of California, Los Angeles: "Collecting Sensitive Data: On the Challenges of Studying Violence in Conflict."
Morris D. Bidjerano, Walden University: "The 'Political Effect' of Political Ethnography Revisited: Contributing to a Local Government Election Campaign While Researching It."
Parakh Hoon, Virginia Tech: "Embedded Ecologies: A Political Ethnography of Wildlife Conservation."
Jessica Peet, University of Florida: "Object or Subject? The Role of Interpretive Methods for Studying Gender."
Discussant: Jan Kubik, Rutgers University, New Brunswick

Field Research: Boundaries, Biases, and Evolving Practices in On-The-Ground Data Gathering
(Co-Sponsored with Comparative Politics Division)

Chair: Diana Kapiszewski, University of California, Irvine
Participants:
Diana Kapiszewski and Robert Nyenhuis, University of California, Irvine: "History, Borders, and Varieties of Field Research in Political Science."
Benjamin L. Read, University of California, Santa Cruz: "Surveys and Experiments as Forms of Field Research."
Kamal Sadiq, University of California, Irvine: "Eyes Wide Shut: The Challenge of Producing Data in Developing Countries."
Lee Ann Fujii, University of Toronto: "Ethical Dilemmas of Fieldwork Among Ordinary People."
Lauren M. MacLean, Indiana University: "The Future of Field Research in Political Science."
Discussants: Lisa Wedeen, University of Chicago; Elisabeth Jean Wood, Yale University

History, Conflict, and Democracy: Historical Approaches to Contentious Politics
(Co-Sponsored with Comparative Politics of Developing Countries Division)

Chair: Benjamin Smith, University of Florida
Participants: Guillermo Trejo, Duke University; Benjamin Smith, University of Florida; Jason Wittenberg, University of California, Berkeley; Mark Beissinger, Princeton University; Matthew Lange, McGill University; Deborah J. Yashar, Princeton University

The Methods Café
(Co-Sponsored with Interpretive Methodologies and Methods Related Group)

Chair: Peregrine Schwartz-Shea, University of Utah
Participants: Emily Hauptmann, Western Michigan University: Archival Research; Lisa Wedeen, University of Chicago: Critical Constructivist and Discourse Analysis; Mary Hawkesworth, Rutgers University: Feminist Methods; Dvora Yanow, University of Amsterdam, Faculty of Behavioral and Social Sciences: Interpretive Policy Analysis: Value-Critical, Policy Discourse, Policy Spaces; Ido Oren, University of Florida: Reflexive Historical Analysis; Peregrine Schwartz-Shea, University of Utah: Research Design for Interpretive Projects; Kevin M. Bruyneel, Babson College: Post-Colonial Analysis.

Additional Related Group Panels

Conceptualization: Theoretical Approaches and Empirical Applications
(Co-sponsored by Interpretive Methodologies, Methods Related Group, and IPSA Research Committee 1 [Concepts and Methods])

Chair: Cas Mudde, DePauw University
Presenters:
Ariel Ahram, University of Oklahoma: "Concept Misformation in Multi-Method Research: Mechanism Muddling and Conceptual Slippage."
Frederic C. Schaffer, University of Massachusetts, Amherst: "Thin Descriptions: The Limits of Survey Research on the Meaning of Democracy."
Jürgen Petersen, Johann Wolfgang Goethe-University Frankfurt: "Political Concepts Between Individual Meaning-Making and Socio-Political Traditions: Bridging the Gap."
Wen-Hsuan Tsai, National Cheng-chi University, Taiwan: "The Concept Construction of 'Chinese Democracy.'"
Discussant: Robert K. Adcock, George Washington University

Theme Roundtable: Hanna Pitkin's "Concept of Representation" Revisited: A New Agenda for Studying Representation Rights?
(Sponsored by IPSA Research Committee 1 [Concepts and Methods])

Chair: Amy Mazur, Washington State University
Participants: Karen Bird, McMaster University; Karen Celis, University College Ghent; Michael D. Minta, Washington University in St. Louis; Anne Phillips, London School of Economics; Michael Saward, Open University; Carole Jean Uhlaner, University of California, Irvine; Laurel Weldon, Purdue University.

Qualitative and Multi-Method Research
School of Government and Public Policy
University of Arizona
315 Social Sciences Building
P.O. Box 210027
Tucson, AZ 85721-0027
USA

Qualitative and Multi-Method Research (ISSN 2153-6767) is edited by Gary Goertz (tel: 520-621-1346, fax: 520-621-5051, email: ggoertz@u.arizona.edu). The production editor is Joshua C. Yesnowitz (email: jcyesnow@bu.edu). Published with financial assistance from the Consortium for Qualitative Research Methods (CQRM). Opinions do not represent the official position of CQRM. After a one-year lag, past issues will be available to the general public online, free of charge, at http://www.maxwell.syr.edu/moynihan/programs/cqrm/section.html. Annual section dues are $8.00. You may join the section online (http://www.apsanet.org) or by phone (202-483-2512). Changes of address take place automatically when members change their addresses with APSA. Please do not send change-of-address information to the newsletter.