How Not to Lie Without Statistics[1]

Gary King[2] and Eleanor Neff Powell[3]

August 22, 2008

[1]Our thanks to Mitchell Duneier, Gary Goertz, Dan Hopkins, Phil Jones for many helpful comments.

[2]Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; http://GKing.harvard.edu, king@harvard.edu, (617) 495-2027.

[3]Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; http://www.people.fas.harvard.edu/∼enpowell, enpowell@fas.harvard.edu

Abstract

We highlight, and suggest ways to avoid, a large number of common misunderstandings in the literature about best practices in qualitative research. We discuss these issues in four areas: theory and data, qualitative and quantitative strategies, causation and explanation, and selection bias. Some of the misunderstandings involve incendiary debates within our discipline that are readily resolved either directly or with results known in research areas that happen to be unknown to political scientists. Many of these misunderstandings can also be found in quantitative research, often with different names, and some of which can be fixed with reference to ideas better understood in the qualitative methods literature. Our goal is to improve the ability of quantitatively and qualitatively oriented scholars to enjoy the advantages of insights from both areas. Thus, throughout, we attempt to construct specific practical guidelines that can be used to improve actual qualitative research designs, not only the qualitative methods literatures that talk about them.

1 Introduction

We highlight errors commonly made in qualitative research, and in various qualitative methods literatures. These literatures span political science, other social sciences, and many related nondisciplinary or professional areas. The signposts we erect at these problems, and the practical guidelines we offer for each, are designed to enable scholars to avoid common pitfalls and to construct more valid research designs. Many of the problems we raise may at first seem specific to qualitative research, but all the same underlying inferential issues also affect quantitative research. Similarly, although many of the errors Darrell Huff (1954) laid out in *How to Lie With Statistics* ("the most widely read statistics book in the history of the world," Steele 2005) are unique to the statistical technology, all the key design issues discussed in his book, and much of the quantitative literature that followed, are relevant to qualitative research too. Indeed, some problems we identify are better understood and resolved via reference to the qualitative methods literature and others to the quantitative literature; but the resolutions usually apply to both.

The errors we discuss include misinterpretations, misunderstandings, and some false claims.[1] They cover a wide range of research design issues. We begin in Section 2 by discussing the role of theory and evidence separately and in how they interact. Section 3 addresses problems related to the distinction between quantitative and qualitative data collection strategies. And finally, we discuss problems in both causation and explanation in Section 4 and selection in Section 5.

2 Theoretical vs. Empirical Research

Some of the loudest disagreements among social scientists can be traced to different tastes for doing and learning about research at particular places on the continuum from theory to evidence. Theoretically minded social scientists complain that they "do not have the stomach for the endless discussions in seminars about getting the methods right when no one seems to care about whether the idea being tested is worthwhile in the first place". Empiricists report "finding it hard to cope with theoretical navel-contemplation that has no hope of being proven right or wrong, or which ignores relevant existing evidence". As political science is among the most diverse of all academic disciplines, and includes scholars along the whole range of the science-humanities continuum, these disagreements manifest themselves most frequently here. Of course, some similar disputes can be found within other social sciences, as well as education, public health, law, and other areas.

The goal of political science research is to describe, explain, and sometimes improve government and politics. To accomplish this task, we must recognize that neither extreme perspective on theory and empricism is right or wrong; these are normative preferences for what type of knowledge any one of us chooses to pursue at any one time. In the end, we need *both* theoretical creativity and empirical validation. Creative theorizing is important even without immediate prospects for validation and for some purposes even without the hope that the theory will eventually predict or explain empirical reality. We ought to recognize the value of an idea in and of itself, separate from the data (i.e., information of any relevant kind). But we also ought to recognize the value of rock-solid empirical validation. Progress requires both. And neither should have priority or even precede the other in the research process: We need "empirical implications of theoretical models"

[1] We avoid pointing fingers with specific citations when discussing false claims and methodological mistakes in prior research since our goal is to build on, not berate, those who have come before.

(Granato and Scioli, 2004) as well as the study of theoretical implications of empirical research. In the four subsections that follow, we elaborate and extend this point.

2.1 Iterating Between Theoretical and Empirical Research

A large component of scholarly research, and a regular topic in the qualitative methods literature, is iterating between theory development and building empirical evidence (e.g., George and Bennett, 2005). The iteration may occur within a single scholarly work or across publications within a field of inquiry. (Although this practice is not often discussed in quantitative methods textbooks, it is of course also as regular a feature of applied quantitative research too.) The issue we address here is that the prospect of getting anywhere productive by iterating between theory development based on flawed data and empirical observation based on flawed theory seems highly dubious. We clarify this fundamental issue in three steps: first, by describing it, then by highlighting and separating out a political argument it has spawned, and finally by offering the first set of formal mathematical conditions under which this iterative procedure will yield the desired result.

First, the idea of iterating between evidence collection assuming some theory, and theoretical development assuming the veracity of some empirical evidence ultimately stems from two key, and seemingly contradictory, points:

1. Social science theories do not come from thin air. All useful theories are ultimately based on *some* empirical observations, no matter how tenuous.

2. Empirical data cannot be collected without at least implicitly making *some* theoretical assumptions about what is to be observed.

The problem is not that these points are contradictory, as they are not: even the basic categories we use to measure apparently raw, unvarnished facts require some type of theory or explanatory typology (Elman, 2005), and no useful theory of social reality can be constructed without assuming some aspects of that reality. Instead, the real problem is that the iterative process may lead us down the wrong path when theories are based on flawed evidence, or evidence is collected when conditioning on the wrong theoretical assumptions. And yet being certain about the veracity of any one observation or theory is impossible, and learning about the world without both is a fantasy.

Second, the uncertain scientific status of iterating between theory and evidence has led to an unproductive, recurring, and largely political, debate in the field. The debate can be summarized by two arguments stridently made by different groups, typically for the purpose of opposing the other group:

1. If you derive your theory from an existing data set, you cannot use that same data to validate the theory empirically. Fitting a theory to data makes you invulnerable to being proven wrong by those data and thus incapable of learning whether the theory is valid.

2. We can greatly improve a theory by firmly basing it in important features of empirical reality and by continually adjusting the theory to fit new observations.

Both points are manifestly true, even though they may seem to contradict each other; each is usually ignored by those emphasizing the other. Although we will never hear the end of either point, we must always recognize both points in all research, and design research with both simultaneously in mind. After all, constructing theories known to violate important features of empirical reality is waste of time (although because theories are meant to be

abstractions, they are designed to miss less important aspects of the empirical world), and so the second point certainly holds. But at the same time, it is too easy to think you have developed an important idea when instead your theory merely has more "but if" statements and "just so" stories tuned to each new datum, and so the first point is essential too. Any time you notice one of these points being made without full consideration of the other, at least implicitly, you're also likely to find some serious inferential mistakes. Sometimes the motivation for ignoring the one of the points is political, but other times we are naturally focused too much on the problem we happened to identify. Clarifying the big picture, which will always involve both points, has the potential to improve much research.

Finally, we now shore up the mathemtical foundation of this procedure by noting that iterating in this way is a qualitative version of Gibbs sampling in Markov Chain Monte Carlo statistical algorithms (e.g., Gill, 2008, Sec. 9.3). The idea is that in statistics and other areas, we often need to be able to draw random samples of two variables, say x and y (as analogies to data and theory, respectively), from their joint bivariate distribution $p(x, y)$ (which indicates the law governing how particular values of x and y occur together), but we may only know how to draw from the two simpler univariate conditional distributions, $p(x|y)$ (i.e., how x varies when y takes on a specific value) and the reverse, $p(y|x)$. Gibbs sampling helps us resolve the problem by starting with a guess (even a flawed guess) for y, drawing x from the distribution of x given the guessed value of y, drawing a new x from the distribution of x given the drawn value of y, and continuing to iterate.

Under the right conditions, we can prove mathematically that this iterative process will converge to draws from the desired joint distribution — which under the analogy to our case, should give the right theory and the right evidence for the theory. So what are the conditions? First, the joint distribution $p(x, y)$ must actually exist. In qualitative research, this means that there is a common process governing the connection, if any, between the theory and evidence. If there is no common process, then trying to learn about it with irrelevant steps in an iterative process will obviously fail.

Second, the way we draw from each conditional distribution needs to stay fixed or at least remain consistent over time. The point here is that for a given project, every time you are confronted with the same evidence you need to have the same view about what theory is likely to be correct; and every time you consider a specific theory, the process by which you select and evaluate data from a given source must remain the same. If we counted the same observation as supporting a theory in one iteration and opposing it on another, we violate this condition.

And finally, convergence to the joint distribution under Gibbs sampling requires that we iterate enough times, collect enough data implied by the theories, and explore enough theories consistent with the data. Exactly how many times we need to iterate depends on how much closer each iteration takes us towards our goal (i.e., how efficient the methods are) and how complex the theory we are developing (more complexity requires more iterations). In practice, the only real check on whether we have reached convergence is to see whether in a long string of iterations we find the same theory along with consistent observations. But in both the formal mathematical version of Gibbs sampling and the analogous qualitative research design we can never be certain we have iterated long enough, and so some more iteration can always be helpful. In a sense, this merely reflects the fact that inference is always uncertain to some degree, and so continuing to iterate — within your research project, or by other scholars as part of a larger research program or literature — may always improve our knowledge of the world we seek to understand.

2.2 Maximizing Leverage

The tension between fitting theory to the data and testing theory can be resolved in part by well-constructed research designs and some creativity. We need to condition theories on as much information as we can be reasonably sure of. But once we do that, the theory fits all known data and we are not vulnerable to being proven wrong — which of course is another way of saying that we cannot learn whether the theory is accurate or whether instead we have pasted together a post-hoc story that seems to fit the facts but does not explain them.

So the key is to make some elbow room between the theory and the data. How do we accomplish this? We suggest two answers that can be applied when feasible, one theoretical and one empirical: (1) reduce the complexity of the theory so that a simpler theory explains the same empirical facts or (2) find new observable implications of the same theory, collect those data, and see whether they are consistent with the theory. Both of these increase *leverage*, the amount of empirical evidence relative to the degree of theoretical complexity.

Why is maximizing leverage so important? Consider three reasons. One is that when our theorizing reveals a new observable implication, we have the chance to shore up the theory by bringing more information to bear on our problem. This procedure is advantageous whether the existing theory is already conditioned on all available data or only some. It is useful whether the implication can be observed by collecting additional data of the same type, the same data in a new time period, or entirely new data from different areas or units of analysis. Data on new observable implications are most valuable when least related to the existing observed implications because then the new data provide independent and thus more informative tests. Thus, for example, complementing an abstract quantitative analysis of a large collection of countries with a detailed ethnography in one city could be highly useful if both measured observable implications of the same idea. Collecting a few more countries would also be helpful, but probably not as much, and certainly not as much if they are highly similar to the countries already in your data. You should always take data where you can get it, but if the same effort can produce data that comes from a very different source or is for another reason unrelated to the data you have, and it is still an implication of the same theory, it would normally be preferable.

A second advantage of maximizing leverage is that data in the social sciences, and indeed in many sciences, is often in short supply relative to the enormous theoretical creativity of individual scholars and the scholarly community as a whole. Creating theories to fit any empirical observation can be done so fast that it is too easy to fool oneself into thinking that you have found something even when you have not. Whereas human beings are stunningly accurate at recognizing patterns, we are miserable at recognizing nonpatterns. In a split-second, we can detect patterns in ink blots and cloud formations, but we are less good at reliably detecting theories without an empirical basis. If you are at all unsure of this, try the following experiment on your colleague in the next office or your spouse: make up a "fact" on almost any subject (e.g., Russia just invaded Iceland! The president has reserved TV air time to make an important announcement! etc.) and see how long it takes before you hear an explanation. Be prepared to count in milliseconds, since there is typically no detectable delay. We must therefore always remain vigilant in putting our theories at risk and continuing to confront them with new sources of evidence. We learn when we try to prove ourselves wrong. And by judging scholarly work by the extent to which it puts its claims at risk of being proven wrong, we can sometimes avoid this key stumbling block in scholarly research.

A final reason why maximizing leverage is important is fundamentally biological. Many

subjects are enormously complicated, and human beings are constructed so that we can only keep a tiny fraction of these complexities in our heads at any one time. As such, and in a variety of different ways, many *define* the concept of "explanation" as requiring simplification — as summarizing, understanding, or accounting for many facts with few. Without this process of simplification, without theories that can maximize leverage, we cannot understand or convey to others the nature of the data and its underlying patterns.

2.3 Relative vs. Absolute Parsimony

The previous section explains that theories with high leverage are valuable because they explain a lot of otherwise unrelated facts, because they help us test our claims, and because of biological facts about how humans think. Although we prefer theories that are relatively more parsimonious than the empirical facts they explain, and more parsimonious than other theories that explain the same facts, no reason exists to value a theory merely because it is simple in an absolute sense. As a result, many claims about parsimony in the literature are misstated.

For example, we should be happy in some circumstances to *add* to the complexity of a theory if doing so accounted for a disproportionately larger array of empirical observations. Parsimony, then, is important only *relative* the the facts it seeks to explain. Unlike implicit claims in the literature that seem to treat absolute parsimony as a mysterious law of nature, whether a parsimonious theory is more likely to be correct (or useful) than a more complex theory is an entirely empirical proposition. To test this theory requires new data that serve as new observable implications.

2.4 The Goals of Empirical Research

The goals of empirical research involve at least three fundamental distinctions. All research projects confront these distinctions as basic choices in the process of research. They are not always considered as explicitly as we are about to, but they are always present. We give them here to give readers a sense of the goals of the enterprise and to orient their work in the broader context of inquiry and to set the stage for the rest of this paper. We portray these three distinctions in the branch points of Figure 1.

A key point in Figure 1 is that *none* of the boxes or distinctions involve an opposition between quantitative and qualitative research. In fact, every box on the page portrays a goal of empirical research that can be pursued via quantitative or qualitative research.

The first distinction in the figure, at the top branch point, is that between *summarizing data* (what King, Keohane and Verba (1994) call "summarizing historical detail") and *inference*. Inference, is simply using facts we have to learn about facts we do not have (about which more in Section 3.1). In contrast, summarizing data involves only examining and summarizing the observations before us rather than trying to learn about facts not observed. Any project with a goal of some type of inference is well advised to first examine the data we have. This step can reveal measurement problems, suggest new inferential targets, or sometimes be of use in and of itself. Of course, to be fair, although maintaining the distinction and distinguishing between facts we know and facts we wish to know is crucial for reducing biases in research, all observation requires some theory and so any amount of summarizing and observing the raw data will always involve some inference; in fact, the discussion about iterations between theory and evidence in Section 2.1 also applies writ small to iteration between observation given some theory about what we are observing and inference given some observation we think we have made.

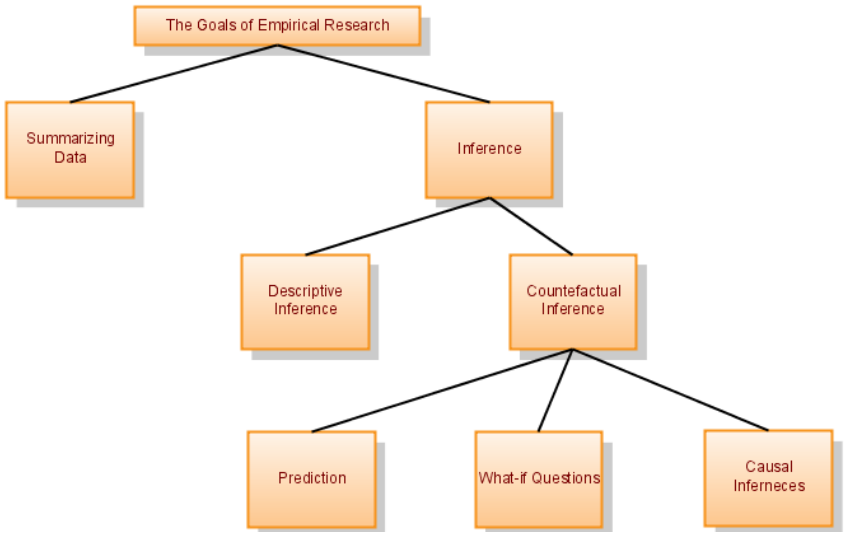

Figure 1: The Goals of Empirical Research. Note that the distinction between quantitative and qualitative styles of data collection does not appear, as the entire figure is the same for both.

The second distinction in Figure 1 (reading from the top down) is the choice between *descriptive inference* and *counterfactual inference*. Descriptive inference is the process of learning about facts which exist but are presently unknown to the researcher whereas counterfactual inference is trying to learn about facts that would or will exist in some other time or hypothetical situation. The term measurement is sometimes used to mean summarizing data is more often reserved as a synonym of descriptive inference. What Americans think of the president is a fact that is not fully known and so can be the subject of a descriptive inference. Sometimes measurement refers to learning what each American thinks of the president, which itself is not fully known or even knowable with certainty by a survey researcher or intensive interviewer, with descriptive inference referring to learning about all Americans. See Adcock and Collier (2001) for more sophisticated definitions.

The final distinction, at the bottom of the figure, are the three types of counterfactual inference (see King and Zeng, 2007). *Prediction* involves facts that will exist in the future (when the time is at a counterfactual, i.e. future, value). *What-if questions* ask about facts that would exist if the world were different in some way; since the world is not different in this way, these facts are unknown. And lastly, a *causal inference* is the difference between some factual detail and the answer to a what-if question, such as the vote a candidate receives minus the vote that that candidate would have received if he or she had a different policy stance. Counterfactual inference (and each of its three parts) also includes the broad goal of *explanation*, a concept often confused with causal inference (and a distinction we take up in Section 4.2).

Finally, although none of the goals listed in Figure 1 are inherently more important than the others, some are more valued in some fields. For example, in public health, descriptive inference, under the banner of measurement, is enormously valued. Finding out exactly where the problems are is crucially important for helping to prioritize funding, research, and amelioration efforts. In contrast, political science and most of the social sciences are primarily concerned with causal inference and less interested in measurement. These are not hard and fast rules, and they are not necessarily right. They are merely the conventional practices and normative preferences of groups of people. But causal inferences

can also be highly valuable in public health, and measurement can add a great deal of knowledge to social scientific inquiries. In fact, when looking across the broad range of scientific fields, it seems clear that many areas of the social sciences have underemphasized basic measurement, which would seem to imply tremendous opportunities for enterprising researchers.

3 Quantitative vs. Qualitative Data Collection Strategies

In this section, we discuss aspects of the divide between quantitative and qualitative styles of research and data collection strategies. Qualitative research is any inquiry other than that requiring numerical measurements. Examples include the parts of ethnographic methods (such as participant observation and interviewing), archival work, historical analysis, and others that do not use quantitative measures. In contrast, quantitative-style research includes explicit measurements of some type. The measurements can be ordinal, interval, or ratio, or they can include labels such as on nonordered categorical data. Typically, there is some fixed unit of analysis, such as the person, country, dyad, social relationship, etc., over which each measurement is taken over, but even this may change across a data set.

In both types of research, the units of analysis may be uniform or varied across the evidence at hand. Since conducting research without any qualitative insights is impossible, even with quantitative measures, we conclude that *all research is qualitative, and a subset is also quantitative.*

3.1 Theories of Inference

To make reliable scientific progress in political science requires at least an understanding of a coherent theory of inference with which specific methods in particular applications can be built, derived, and adapted. Such a theory exists for quantitative analysis. This theory has been developed, elaborated, and extended within the disciplines of statistics, philosophy, and many of the social sciences. It has led to the development of a wide range of methods for the analysis of particular types of data and information, and has made it possible to evaluate and improve numerous approaches.[2]

What theory of inference under-girds qualitative research in political science? At present, two options are on offer. The first is to appeal to the well-developed theory of inference used in the statistical and philosophical literatures (as suggested by King, Keohane and Verba, 1994). The second is to wait for some type of unique and coherent theory of inference to emerge from the qualitative methods literature itself. Although qualitative researchers sometimes object to their work being evaluated by a theory of inference so closely associated with quantitative research, no such "qualitative theory of inference" has emerged, no coherent arguments have developed to support it, and no effort is underway to make it happen. Of course, there is little reason to develop a new theory of inference since the existing one, although developed in large part within and for quantitative research, does not require information to be quantified and can be used in most situations directly in qualitative research, without quantification. Indeed, the

[2]What we call *the* theory of statistical inference is actually a set of fundamentally distinct theories which differ mathematically in important ways. These include Bayesian, likelihood, robust, and nonparametric theories, among others. However, from the broader perspective of this discussion, and the qualitative literature in general, they are all sufficiently closely related that it is reasonable to treat them as a single theory (Imai, King and Lau, 2007). Indeed contributors to each normally recognize and frequently build on the contributions from the other theories of inference.

degree to which the theory of inference has to be modified and adapted to apply to most qualitative research settings is no greater than routinely occurs when applying it to new forms of quantitative data.

Fortunately, the applicability and validity of the reigning theory of inference to qualitative data is widely and increasingly recognized and respected by many qualitative researchers. Even those rankled by analogies to statistical inference regularly engage with and contribute to the application of notions from this theory of inference like omitted variable bias, selection bias, and many other issues. However, when researchers fail to recognize this theory of inference, they sometimes act as if they are free to invent methods without constraints or evaluation except introspection or intuition. This of course can be a real mistake, as intuition fails so frequently in making inferences (see Section 3.4). Indeed, recognizing the inferential foundation of existing and future qualitative methods is essential, since without it cumulative progress is extremely unlikely. Indeed, when there is no commonly recognized theory of inference, even understanding different proposals for qualitative methods can be difficult. This issue is exacerbated by the propensity of each qualitative researcher to invent a new language to describe what are often the same issues. Numerous examples, like the equivalence of "most similar systems" designs and Mills' "method of difference," would be easy to recite. Multiple languages are also used across the different methodological subfields of various substantive disciplines, although they have the same underlying mathematical representations to help with translation.

3.2 Quantitative vs. Qualitative Methodology Subfields

The nature of the professional fields of quantitative and qualitative research methodology could hardly be more different. The core of quantitative research methodology is the discipline of statistics. This basic field is supported by the closely related quantitative methods subfields existing within most applied social science disciplines and nondisciplinary professional areas. These subfields include political methodology within political science, psychometrics and other statistical analyses within psychology, biostatistics and epidemiology within public health, econometrics within economics, sociological methodology within sociology, and many others. These areas are united by common or closely related mathematical representations of statistical models and approaches. The methods subfields innovate by developing methods that address new data problems and quantities of interest, and the discipline of statistics innovates with rigorous proofs, and to some extent vice versa, but the underlying theories of statistical inference that give rise to these models are shared across all these fields. Intellectual progress from the collective effort has been remarkably fast and unambiguous.

A summary of these features of quantitative methods is available by looking at how this information is taught. Across fields and universities, training usually includes *sequences* of courses, logically taken in order, covering mathematics, mathematical statistics, statistical modeling, data analysis and graphics, measurement, and numerous methods tuned for diverse data problems and aimed at many different inferential targets. The specific sequence of courses differ across universities and fields depending on the mathematical background expected of incoming students, the types of substantive applications, and the depth of what will be taught, but the underlying mathematical, statistical, and inferential framework is remarkably systematic and uniformly accepted.

In contrast, research in qualitative methods seems closer to a grab bag of ideas than a coherent disciplinary area. As a measure of this claim, in no political science department of which we are aware are qualitative methods courses taught in a sequence, with one

building on, and required by, the next. In our own department, more than a third of the senior faculty have at one time or another taught a class on some aspect of qualitative methods, none with a qualitative course as a required prerequisite.

Perhaps this will change through efforts by the Consortium on Qualitative Research Methods, and its popular summer training program, to promote the teaching of qualitative courses in the social sciences. But it is likely to be most successful, as this group also emphasizes, only if these courses are integrated with regular statistics courses, because of the conceptual importance of the theory of statistical inference and other topics frequently covered in statistics classes.

3.3 Multiple Data Sources, not Multi-Methods

A popular banner among qualitative researchers in political science in recent years is "multi-method research." The phrase is even included in the newly renamed "Qualitative and Multi-method Research Section" of the American Political Science Association. The phrase multi-methods is a call for pluralism in the choice of research styles, data sources, and analytic methods. It is also a call attempting to make some room for approaches that do not involve statistical analyses. But, to be clear, from the perspective of learning about the world, the literal meaning of the phrase "multi-method research" makes little sense and is too easily confused with the desirable goal of having multiple sources of data (except of course when it refers to different methods of collecting data rather than analyzing it; Lieberman 2005).

That is, for any one type of information collected from a single data source, a near optimal method (or range of methods distinguishable only by unverifiable assumptions) is available or can be constructed. Getting agreement on this method or range of methods, even among scholars with divergent backgrounds, is typically not difficult. Moreover, adding analyses based on methods outside this range, for the purpose of analyzing the same data, can then only mean using suboptimal approaches.

In contrast, collecting additional, diverse sources of information that are implications of the same theory is uniformly beneficial. It directly advances the research enterprise by enabling us to maximize leverage. More data, in as diverse forms as possible (i.e., such that each new source of information is minimally related to the existing sources), is in this sense always better. Thus, you may sometimes require multiple methods to deal with multiple data sources, but *multi-data sources and not multi-methods is the value that should be maximized*.

To be more specific, diversity comes in several flavors. Diversity of data sources helps ensure against bias of the source, so long as the sources are unrelated. Diversity of data type (such as a detailed ethnographic studies of one city vs. abstract quantitative cross-country summaries) may ensure against bias, but they are especially useful for increasing efficiency since they will often bring observable implications of a theory least likely to be related, and thus most likely to be additionally informative, compared to the existing data. In contrast, diversity of method, given a particular set of quantitative or qualitative data, only takes us away from whatever the optimal method or methods are in that situation. Applying different methods to the same data is only useful in studying model dependence (about which more in Section 5.4).

A related point concerns the popular notion of emphasizing the collection of both quantitative and qualitative evidence in the same scholarly work, especially in dissertations. As should now be clear, combining both sources is good only if it increases the diversity and amount of observable implications. But there is no *additional* value in emphasizing both:

no magic occurs if they are mixed in the right proportions beyond how the collection of data can increase one's leverage. It may be true that your job prospects may be enhanced if you do both in some subfields, but this is career advancement, and only sometimes; its not necessarily the best way to make scientific advances. From the perspective of learning about the world, the optimal allocation of your scarce resources, including your own time, suggests collecting the most data and the most diverse forms of data. That may suggest adding quantitative data to qualitative, qualitative cases to quantitative data, or it may mean collecting additional forms of the same type of data. The issue is what maximizes leverage, not whether you can collect one leaf from every tree.

3.4 Is Quantitative or Qualitative Data Better for Your Research

Remarkably, this is a question with an extremely well justified answer based on considerable supporting research — research that seems to have been thoroughly ignored by the qualitative methods literature. The fact that the answer may be incendiary for some does not make it any less true. Coming to terms with the answer should greatly improve research of all kinds in our field. We explain the answer via two separate facts:

The first fact is:

> When insufficient information about a problem has been quantified to make statistical analyses useful, and additional qualitative information exists, qualitative judgment and analysis are typically superior to the application of statistical methods. The evidence in support of this point is not seriously debatable. No application of statistical methods, no matter how fancy, can overcome an inadequate information source. Yes, theory can greatly improve a statistical analysis, but the theory must at some point be based in empirical fact or else the same story holds: if the available quantified information is inadequate, statistical analyses will work poorly and often worse than qualitative analyses.

Researchers must understand that it is and will probably always be impossible to quantify the vast majority of information in the world. The last time you wandered into a classroom, you instantly decided your students were not going to eat you for dinner. When you woke up this morning, you quickly decided that there was no emergency and you probably even figured out what city you were in without much delay. When you are served a meal, you can detect with one sniff whether it has spoiled with imperfect but high reliability. If we were not able to make snap decisions like these with some degree of accuracy, humanity's main accomplishment would have been as a food source for saber tooth tigers. In other words, no matter how many statistics recourses you've taken, deciding to collect data and run a regression at moments like these, and in many areas of scientific research, would not be helpful.

The second fact, much less widely known in our field than the first, is:

> When sufficient information about a problem can be quantified (a crucial qualification!), a high quality statistical analysis is far superior to qualitative judgment. Mathematics and statistics enable human beings to reason properly even when informal human reasoning fails. Human reasoning, in turn, fails in highly predictable ways that qualitative experts have not been able to overcome even when the field of statistics has. Qualitative judgments by subject matter experts are routinely out-distanced, out-performed, out-reasoned, and out-predicted by brute force statistical approaches. This is true even when the

the data analysts know little about the substantive problem at hand and the quantified information seems shockingly incomplete to subject matter experts.

This fact will come as a surprise only to those unfamiliar with the quantitative literature, but so many examples of this point now exist in so many fields that it is no longer seriously debatable. For example, in a head-to-head contest two political scientists with a crude six-variable statistical model predicted the outcome of U.S. Supreme Court cases (without reading them) more accurately than a set of 83 law professors and other legal experts reasoning qualitatively and with access to enormously more information and decades of jurisprudential experience (Martin et al., 2004). For another example, political scientists have long been more successful at forecasting presidential elections than pundits, pollsters, and others (Campbell, 2005; Gelman and King, 1993). Tetlock (2005, p.64) has shown that most of his 284 articulate, highly educated, and experienced experts forecast many aspects of the political future with "less skill than simple extrapolation algorithms." Similarly, two political scientists with no medical training built a statistical model that out-performs physicians (assessing individual causes of death) in determining cause-specific mortality rates (King and Lu, 2008). These are but four of hundreds of such examples in many scholarly areas. Indeed, at least since Meehl (1954), numerous similar contests and comparisons have taken place across various fields of study and practice. The result is not always the same, but the same very strong tendency favoring the quantitative estimates is ubiquitous (Grove, 2005). There even exists a wide ranging popular book devoted to the subject (Ayres, 2007). The qualitative methods literature in political science needs to come to terms with these facts. Accurate prediction or estimation is awfully difficult in most fields without measurement precision.

The claim does not mean that incompetent or inadequate statistical analyses, of which there are many, are in any way superior or even necessarily useful. Conducting quantitative analyses is difficult and time consuming, requires preparation and training, and can easily be done wrong, and often is done wrong. The claim is not that statistical analyses are more often done correctly than qualitative analyses, only that high quality statistical analyses based on adequate data are usually superior to qualitative expert analysis when sufficient information has been quantified. The mere presence of numbers in an article or book conveys no necessary assurance of anything else.

Should the fact that many statistical analyses are done badly cause us to conclude that quantitative approaches have no practical value? This would be the case without a fairly unified approach to the theory of inference. With it, different analysts, reaching different conclusions from different approaches to data analysis or sources of quantitative information, can converge to similar or identical answers. The theory of inference provides a framework, a common standard that can be applied to apparently contradictory approaches. This is what good quantitative approaches have going for them now; it is what qualitative scholars only sometimes benefit from now. But even when more qualitative scholars develop their approaches with reference to common theories of inference, quantitative methods will still be superior to qualitative methods when (and only when) sufficient information has been quantified.

Whether quantitative or qualitative data are better for your research, then, depends on how much information is available, the extent to which it can be systematized and quantified, and how much time you can devote to the problem. Quantification for its own sake is a waste of time (as qualitative scholars forced to include baby statistical analyses in their qualitative works by reviewers and dissertation advisors know!). But in the small number of situations where you are able to quantify the essential information, gear up to do a proper statistical analysis, and commit the time and resources necessary, it is worth

going ahead because the increased precision will likely yield far more accurate results. If all these conditions do not hold then its best to proceed qualitatively.

3.5 What Research Can't Statistical Inference Represent?

Adding to knowledge about a research problem by conducting statistical analyses is not always easy, efficient, advisable, or useful, but it it is usually possible, at least in principle. The regular critiques to the contrary in the qualitative methods literature about unavoidable problems unique to quantitative research seem to be based on particular applications of quantitative research circa 1970 or when the critic was in graduate school. Qualitative scholars have regularly argued that quantitative research is incapable of dealing with dichotomous dependent variables, collinear explanatory variables, measurement error, interactions, multiple causal paths, path dependence, nonlinear functional forms, selection on the dependent variable, models without specified functional forms, "overdetermined" problems, analyses without models, and numerous other patterns and issues.

These claims, and many others like them, are false. Indeed, we would go further and make the following alternative claim:

> Every inferential statement, empirical pattern, and notion of uncertainty can be represented sufficiently well, for the purposes of social science analysis, by the statistical theory of inference.

Coming up with a formal statistical approach for any arbitrary idea will not always be easy, and indeed thousands of methodologists in dozens of academic fields are doing the same with their own problems, but it should always be possible. At the least, no impossibility theorems have been stated or proven. For some recent examples of new approaches in our field related to discussions in this paper, see Braumoeller (2003) and Glynn and Quinn (2008).

Of course, just because an inferential statement can be given in formal statistical terms does not make other qualitative versions useless. They may be technically unnecessary, since there exists other terms for the same concepts, but different emphases can be very useful in guiding scholarship to new or underappreciated questions and approaches. Consider the following example.

Example: What is *path dependence*? The broadest view of path dependence in qualitative research is the simple and important idea that history matters, or sometimes that institutions matter. In this broad incarnation, everything discussed in the qualitative literature on path dependence has been or could easily be formalized within the extensive existing quantitative literature on *time series analysis* (e.g., Hamilton, 1994): or in other words, literally nothing is new. In other formulations, path dependence refers to the more specific idea of historical processes that have "increasing returns," which is when apparently small events turn out to have larger (or at least permanent) consequences as time passes (Pierson, 2000; Hall, 2009, forthcoming). This more specific notion of path dependence is also extremely well studied in the time series literature under the names "nonstationary" processes or "unit roots", and so technically nothing is new here either. However, the emphases of the two literatures are almost exact opposites, with almost no cross-citations and thus little possibility of building on each other's work. The lack of contact between these complementary fields poses a substantial opportunity; the diametrically opposed emphases serves an important purpose in encouraging scholars to focus on ideas that might be lost in the other field.

To see this point, note that any historical or time series process can be decomposed into stationary and nonstationary components. The stationary components, no matter how complicated, are those which follow the same probabilistic patterns whenever they occur, so that for example the effects of events or shocks to the time series do not grow without limit over time. The nonstationary components are those parts with increasing or decreasing returns, or any feature dependent on a particular historical time. The key point here is that *while the qualitative literature on path dependence puts primary emphasis on the nonstationary component, the statistical literature often views the nonstationary component as a problem to be corrected (or "differenced away" in their language) and instead focuses on the stationary component.* Their emphases do sometimes make it seem like they treat important aspects of the data as a "nuisance" rather than a valid subject of study (Beck and Katz, 1996).

The idea of path dependence is not wrong; it may be superfluous, but saying the same thing in different language is obviously important in this context. We encourage qualitative scholars to integrate their work more effectively with the existing statistical time series literature, as it can make qualitative claims more effective, powerful, far reaching, and precise. At the same time, quantitative scholars would be able to make their work more relevant, far reaching, and influential by further developing their tools to analyze the nonstochastic components of time series, and not to treat them as mere nuisances to be corrected. And of course, no reason whatsoever exists for ignoring the nonstochastic component in quantitative time series or the stochastic component in qualitative research.

One of the best things about statistics is that the field has made stunning progress, with developments appearing quicker every year. The various fields of statistics are multiplying, with new methods being developed and applied to wider and wider arrays of applied problems. The march of quantification through the fields of science has proceeded fast and steady. Judging from articles published in our leading journal, about half of political science research since the late 1960s has been quantitative. Economics is more; Sociology is somewhat less. Medicine is now "evidence based." The hottest trend in law school research is quantitative (which they call "empirical research"). Biology was once highly qualitative but biostatistics and bioinformatics and several other subfields have blossomed to make possible many analyses never before thought possible. Governments and businesses now conduct numerous large scale randomized experiments.

There are fields of study that have not yet been revolutionized by increasing quantification and modern statistics, but its an easy prediction that this will eventually happen given enough enterprising scholars, wherever it would be useful (and unfortunately, at other times too!). Certainly the opportunities for intellectual arbitrage are enormous. To take one example, for clarity outside our field, consider architecture. By far, the most expensive decisions universities make are about buildings and their physical plant. Yet architecture as a field is composed primarily of engineers who keep buildings up and qualitative creative types who invent new designs: quantitative social scientists do not frequently get jobs in schools of design. Imagine instead how much progress could be made by even simple data collection and straightforward statistical analysis. Some relevant questions, with associated explanatory variables might be: Do corridors or suites make the faculty and students produce and learn more? Does vertical circulation work as well as horizontal? Should we put faculty in close proximity to others working on the same projects or should we maximize interdisciplinary adjacencies? (Do graduate students learn more when they are prevented for lack of windows from seeing the outside world during the day?) And if the purpose of a university is roughly to maximize the number of units of knowledge created,

disseminated, and preserved, then collecting measures would not be difficult, such as citation counts, the number of new faculty hired or degrees conferred, the quality of student placements upon graduation, etc. A little quantitative social analysis in architecture could go a long way in putting these most expensive decisions on a sound scientific footing.

It would be easy to continue this story, as statistics has affected many other areas of human affairs. In political science, quantitative analysis is joining qualitative analysis in field after field. At one time, only American politics included any quantitative analysis. Now comparative politics and international relations are almost as quantitative as American.

The point is not that everything will eventually be quantified. It won't, since that end would discard the vast majority of available information. The point is that statistical inference is not a fixed set of methodological tools. It is a set of procedures to develop tools adapted to problems as they arise. Claiming that some inferential need is not encompassed by statistical inference may be appropriate at some time point, but only until someone sees the need and develops a new tool for that purpose. But at the same time no matter how far quantification and systematic measurement proceeds, qualitative analysis will always be part of every analysis. To make the fastest progress, we need better connections between these well developed fields.

4 Causation and Explanation

Fundamental goals of most quantitative and qualitative social science include causality and explanation, as distinct from description and descriptive inference. We regularly ask *why?* Most of our theories and the vast majority of our empirical analyses seek to go beyond measurement to understand the causal structure of social phenomena. Political scientists often do not even accept predictive evidence without some plausible causal story about how the chosen explanatory variables could lead to the dependent variables. Causality seems to define the essential nature of many social science disciplines and distinguish them from related professional fields.

Unfortunately, despite the central role of causality and explanation in most substantive areas of inquiry, there exists a major misunderstanding within the qualitative methods literature on these issues and a separate and just as major misunderstanding within the quantitative literature. The mistake in one area is not made in the other, and so there is hope that some additional communication between quantitative and qualitative researchers can fix the problem. We introduce the qualitative confusion over the definition of causal effects in Section 4.1 and the quantitative literature's confusion over the meaning of explanation in Section 4.2. Section 4.3 then discusses confusions in both areas over how to estimate causal effects.

4.1 Defining Causal Effects: Confusion in the Qualitative Literature

Scholars in the qualitative methods literature sometimes write as if a controversy exists somewhere over the fundamental definition of what a causal effect is, and they routinely introduce new formulations — sometimes reasonable, sometimes logically inconsistent, but usually unnecessary — to try to characterize it. In fact, the definition of causal effects given in King, Keohane and Verba (1994, ch.3) — now widely known as the *potential outcomes* framework (or the "Rubin causal model") — has since become the near consensus position in almost all academic fields where such matters have been discussed. Statisticians attribute this definition to Neyman, Rubin, and Holland; computer scientists to

Pearl; economists to Granger and others; epidemiologists to Robins; and philosophers to Aristotle, Locke, Hume, Mill, or Suppes. Some scholars favor requiring that the potential outcomes framework be supplemented with other features, others slightly adjust the primary definition in some ways, and still others are more interested in concepts other than causal effects. But political scientists should recognize that whichever version we focus on and whomever we attribute it to, the basic potential outcomes definition is now well established and generally agreed upon. In fact, it is difficult to think of many sophisticated concepts that were once the subject of such widespread disagreement that are now as widely agreed upon as the potential outcomes definition of causality.

4.1.1 The Basic Definition

The best way to understand the core of the potential outcomes definition of causality is in the context of a single unit (person, country, or other observation). Suppose for simplicity that the causal factor of interest, which we label T_i for the "treatment" variable applied to unit i, is either applied ($T_i = 1$) or not applied ($T_i = 0$) to person i at any specific time. In some versions the treatment must be manipulable, such as implementing a particular public policy in some states and not others or a political party choosing whether to endorse an incumbent for reelection rather than a nonincumbent; most social scientists now also allow the treatment to include an attribute of the units, such as gender or region, which is for practical purposes not manipulable (Goldthorpe, 2001).

Then suppose we observe the value of the outcome variable for person i, Y_i, when exposed to the treatment, which we label $Y_i(1)$, and so we obviously do not and cannot observe the value of the outcome variable for person i at the same time when not exposed to treatment, $Y_i(0)$. The causal effect of T on Y for person i is then the difference between the two potential outcomes, such as $Y_i(1) - Y_i(0)$. Because $Y_i(0)$ is unobserved, the causal effect is never known for certain and always must be estimated. We may try to estimate $Y_i(0)$ at a different time for person i or for a different person similar to i who was not exposed to the treatment at the same time, but these are estimation strategies requiring assumptions that are only sometimes valid; however, either way, they have nothing to do with the definition of the causal effect which is for person i at a single point in time. Sometimes scholars are interested in this causal effect averaged over all units in some chosen sample or population, but the core definition is best understood for one unit at a time. (This core definition is also supplemented by some, to require additional information or formal models about aspects of how the cause has its effect; Heckman 2008.)

For an example of the basic definition, consider the causal effect of (say) the election of George W. Bush rather than Al Gore as president in 2000 on the country's gross domestic product (GDP) in the year 2004. The causal effect is the difference between the actual GDP in 2004 (with Bush as president) minus the GDP that the U.S. would have had in 2004 if Gore had won in 2000. This second, unobserved value of GDP is known as a potential outcome since it is counterfactual that could have been observed under different circumstances (i.e., if Gore had been elected) but was not.

This is a fairly specific definition of a causal effect, but it is not yet fully defined until we also detail precisely how the counterfactual world of Gore being elected could have taken place. One way would be to imagine that the Supreme Court decided *Bush v Gore* the other way. The causal effect based on this definition of the treatment effect is well defined, since it is easy to imagine this counterfactual having actually taken place. We might hypothesize that the effect on GDP of the Supreme Court deciding differently would be similar to the effect on GDP of Gore being elected because of the absence of the problems with the butterfly ballot, or Ralph Nadar decided not to run, or for a variety of other

"small" possible changes, but these are different causal effects. For example, the effect of the Court decision would plausibly be very different from some other counterfactuals, such as Gore being elected because the Army rolled tanks through Washington and forced his installation as president.

However the counterfactual, and thus causal effect, is defined, it is must be delineated before we can begin to discuss estimation. In fact, a key point is that the causal effect must be defined without reference to any method of estimation — such as the unsatisfactory approaches of saying that the effect is what happens after controlling for potential confounding control variables, or appealing to "all things being equal" claims, or confusing the causal effect with a regression coefficient, or defining the causal effect as the difference between Y before and after an event, etc. This is essential for logical coherence if nothing else, since a true target quantity of interest must exist prior to and separate from our efforts to estimate it.

This example of the effect on GDP gives the most basic version of a causal effect within the potential outcomes framework. But although the framework is very simple, it has proven to be a remarkably durable and fertile approach with extremely wide applicability. Many more sophisticated versions of causal effects have been built on this simple framework in many literatures, including studies in many types of quantitative and qualitative data; for observations over time, across space, and varying over both; and for studying intermediate effects and causal effects that occur at different levels of analysis. A large number of applications have also appeared, all beginning with this basic distinction (Morgan and Winship, 2007).

The scholars applying qualitative methods need to come to terms with this definition far more than they have. The precision that comes from laying out the counterfactual involved in a causal claim can be tremendously clarifying from a theoretical perspective, can help identify data relevant for estimation and other observable implications, and can assist in weeding out data that have no relevance to the claim at hand. As we see in classes, apparently understanding the potential outcomes framework is not the same as being able to apply it to one's problem; like anything else, it must be practiced to get it right. Doing so has enormous benefits since a huge range of implications for the understanding, definition, and estimation of causal effects has been worked out and could be harvested by qualitative scholars for their work.

4.1.2 Not the Defintion: Necessary and Sufficient Conditions

Many scholars, especially in the qualitative methods literature, are interested in ideas such as necessary and sufficient conditions. These are important concepts, but much confusion would be eliminated if scholars recognized that they are not causal effects. That fact makes them neither unimportant nor unworthy of study.

We begin with a definition of necessary conditions. For expository simplicity, suppose (like T) the outcome variable Y only takes on the value of 1 or 0 for the presence or absence, respectively, of some characteristic. Then a necessary condition is that Y cannot occur when T is absent: $p(Y = 1|T = 0) = 0$. Thus, a necessary condition a well-defined inferential target, as it is always either true or false and is unambiguously separate from any method that might be used to infer whether it is holds. However, the definition of a necessary condition involves no counterfactuals and so is not a causal statement (or at least is different from, and unrelated to, the definition of causal effects given in Section 4.1.1). The probability $p(Y = 1|T = 0)$ is in fact a *descriptive* quantity, requiring descriptive and not counterfactual inference (see Figure 1).

To be more specific, in the context of the dichotomous outcome in this example, the causal effect can be written as $p(Y = 1|T = 1) - p(Y = 1|T = 0)$ (for a single unit). However, if T is necessary for Y, so that the second term is zero, the causal effect becomes the first term, $p(Y = 1|T = 1)$, which is of course an entirely different quantity from, and in no way constrained by, $p(Y = 1|T = 0)$. (In fact, if $p(Y = 1|T = 0)$ is a descriptive quantity then, when defined for the same unit for which $T = 0$, $p(Y = 1|T = 1)$ is a counterfactual quantity.) This proves that causal effects and necessary conditions are separate inferential targets, both logically consistent, both of some substantive interest, and neither of which should be confused with the other.

The same logic can easily be applied to sufficient conditions since a necesary condition always implies at least one sufficient condition. For example, if the necessary condition $p(Y = 1|T = 0) = 0$ holds, then the equivalent sufficient condition $p(Y = 0|T = 0) = 1$ also holds (since the two must sum to one). Many creative and sophisticated combinations and extensions of necessary and sufficient conditions have also been developed (Mahoney, 2008), and of course the same point applies to all: they are worthy inferential targets, but not causal effects. A researcher may wish to choose between the two to apply to a particular applied problem, but those interested in qualitative methods never need to choose one or the other since they are relevant and different objects of inquiry.

Quantitative researchers sometimes belittle necessary and sufficient conditions, because we do not live in a deterministic world. As such, with a little measurement error or stochastic variability, the conditions never seem to hold in practice. However, at least in the context of measurement error the concepts are frequently used (Goertz, 2003) and well-defined, and they can easily be used as subjects for descriptive inference (Braumoeller and Goertz, 2000).

4.1.3 Also Not the Definition

Qualitative and quantitative scholars seem regularly confused by terms related to causal effects, such as *causal mechanisms*, *overdetermined* effects, *equifinality*, and *multicausality*. The definition of a causal effect is given in Section 4.1. These terms are neither required nor necessarily helpful in understanding the definition, applying it, or estimating causal effects. Of course, they are sometimes useful in encouraging scholars to emphasize different data collection approaches, to look for different types of patterns, and in some cases to identify specific relationships in data, but to give a coherent definition of each is best done by building on, rather than ignoring, the prior potential outcomes definition of causal effects.

For example, the recommendation to search for causal mechanisms after establishing a particular causal effect can be very useful in motivating a productive research program, producing a better understanding of the world, or generating or confirming a more powerful theory of the phenemonon under study (Gerring, 2007, p.185). Nevertheless, *defining* causality solely in terms of causal mechanisms makes no logical sense. In the running example from the previous section, one hypothesis about a causal mechanism is that Gore would not have gone to war in Iraq and so the U.S. government would have spent less on war supplies. There is some evidence that spending on war has less positive effects on the economy than other types of expenditures, and so then the election of Bush rather than Gore may have depressed U.S. GDP. This is an interesting and perhaps important subject of study, and one which the potential outcomes framework entirely encompasses (Glynn and Quinn, 2008), but it is not a self-sufficient definition of the causal effect of the election outcome on GDP. The reason is infinite regress: If you try to define the causal effect in terms of the mechanism, with the mechanism merely being another causal hypothesis,

then the question immediately becomes what the mechanism of the mechanism is. In our example, we would then need the mechanism by which war expenditures do not have as positive effect on the economy as social welfare expenditures — such as because many of the products are shipped overseas or blown up. But then we need the mechanism by which this happens, etc. This infinite regress may well define a productive research agenda, but it does not offer a coherent definition of a causal effect.

For another similar confusion, phenomena are often described in the literature as being overdetermined, by which it is meant that more than one cause could have produced the same outcome. As it turns out, this phrase is largely vacuous because by definition almost every event is overdetermined. He died because his heart stopped, or he had too little blood, or his aorta was breached, or the bullet entered his body, or the trigger was pulled, or he was shot, or the murderer carried out his intent, or the murderer was made to do by his friends, his background, or society, or because of the way the biological process of evolution proceeded, or because of the physics of the big bang. The concept of being overdetermined is incoherent.

For the same reasons, every event and outcome variable can be described as an example of multiple causation or equifinality. These terms — like overdetermination and causal mechanisms — do not distinguish any cause from any other. They may help in establishing certain emphases, but in conducting actual empirical research for the purpose of defining and estimating causal effects (whether quantitative or qualitative), the focus should be on the definition of causal effects given in Section 4.1, precisely how your casual hypothesis and the implied counterfactual can be defined, and what assumptions may be necessary to estimate it.

4.2 Defining Explanation: Confusion in the Quantitative Literature

Causal effects are well defined, but they are not the same as explanations. This fact is misunderstood throughout the quantitative literature but is reasonably well understood among qualitative scholars. We clarify these points here.

The problem appears to stem from a statement in the influential article by Holland (1986, p.945) on the potential outcomes framework:

> The emphasis here will be on *measuring the effects of causes* because this seems to be a place where statistics, which is concerned with measurement, has a contribution to make. It is my opinion that an emphasis on the effects of causes rather than the causes of effects is, in itself, an important consequence of bringing statistical reasoning to bear on the analysis of causation and directly opposes more traditional analyses of causation.

Measuring the effects of causes is indeed the notion behind the potential outcomes definition of causality. The treatment variable is defined and then the causal effect is the difference between two outcomes (or "effects"). What Holland means by "the causes of effects" is reasoning backwards from the effect or the event (or process, people, perspective, role, community, etc.) to one or more causes of that event. Defining this process is logically impossible in a well defined way, and has the same infinite regress problems as the concept of "overdetermined" does, discussed above: more causes always exist for any event. All you need to do is to look at different proximate or distal levels.

Holland's decision to focus on the effects of causes is a perfectly reasonable focus for an article, and indeed for whole literatures, but it does not mean that looking for the causes of effects make no sense. In fact, what is widely called the causes of effects in the

quantitative literature is called more generally in the qualitative literature, *explanation*. Statistics types know that they are less likely to succeed in coming up with a precise and well-posed definition of an explanation because it involves finding the causes of effects, which does not fit well into the potential outcomes framework because the cause itself is not well defined, and for which no other precise definition has yet been developed. However, just because statisticians have not come up with a precise and coherent way to do it does not mean that we should expect social scientists to stop trying to explain social phenomena! (Related work touching on issues of explanation in somewhat different ways exist in philosophy Little (1991) and economics (Heckman, 2008).)

If this does not yet sound convincing, how do you explain the advice to consider rival theories when designing research? Keeping an open mind to the possibility of being massively wrong while pursing the observable implications of their currently favored theory is hard to argue with, and so the advice seems highly useful. But a specific version of rival theories is rival causes, and rival causes is precisely looking for the causes of effects. Again, the fact that this is not readily amenable to analysis under the potential outcomes framework does not mean that we should ignore rival explanations.

So what exactly is explanation? Well, its not exactly anything, since its definition has never been made precise. But roughly, it is a set of causes such that the effects of these causes on the phenomenon of interest are "large" and at the "right" level of proximity to the event. Such causes must also have some other attractive, but difficult-to-state properties, such as concerning the subjects or acceptable areas under study. So start with some outcome — U.S. GDP was $11.4 trillion in 2004 — and then go backwards to find a treatment variable that has a large effect on that outcome. One possible answer would be the U.S. Supreme Court's decision in *Bush v Gore*. It should be clear both that this indeed is one type of answer we seek when looking for the explanation of U.S. GDP in 2004 and why it is different from what we might have expected. However, its status as "the explanation" is highly imprecise and ambiguous. For one, another perfectly reasonable explanation might be the price of oil, and another would be the state of the world economy, and another might be the terrorist incidents on 9/11/2001. Another at a different level of proximity might be the size and openness of the U.S. economy. There are thousands of other explanatory candidates too, and no sense in which one set of these is more appropriate as "the explanation" than the others. Indeed, even all the causes we can think of together do not necessarily constitute the "complete" explanation, since such a thing either does not exist or is extremely ill defined. After all, one "good" explanation by these rules for why U.S. GDP was $11.4T on 12/31/04 is that it was just a tad under on 12/30/04. Perhaps that "explains" the outcome, but for other unstated reasons is not very satisfying (see also Hall, 2006).

You can see where this conversation is going: Qualitative scholars understand that the real question social scientists have in many cases is what explains an event: Why did the Cuban missile crisis happen? Why, in free and fair elections in Weimar Germany, did voters choose to elect Hitler? What was the cause of the French Revolution? Why did the terrorists blow up those buildings on 9/11? These are all calls for explanations, for the causes of effects. When quantitative scholars hear questions like these they sometimes scoff at their imprecision, which is a fair complaint but also misses the point. That is, none of these questions are defined as clearly as the potential outcomes framework when looking for the effects of causes.

Quantitative scholars can proceed to define and evaluate causal effects via potential outcomes for each candidate explanatory factor — which is their most helpful suggestion in cases like these — but they have not properly defined or formalized the definition of

an explanation or how to estimate it. They need to understand that explanations are a legitimate target of social scientific inquiry, however vague they may be. The statistics community may insufficiently appreciate this point or in any event has not yet done anything about it. An important research agenda within that field ought to be to formalize "explanation" and show quantitative and qualitative scholars alike how to produce explanations in more accurate and efficient ways. Until then, the qualitative literature will lead, even if imprecisely.

4.3 Causal Inference Without Regression Assumptions

A strange notion exists within the qualitative method literature, and in some areas of quantitative research, that quantitative estimates of causal inferences require some form of regression analysis. This notion is false. Not only does the quantitative literature include numerous types of inferential methods for estimating causal effects, but some methods are very close to those used in the qualitative methods literature. In this section, we discuss one key method — matching — that can easily be described as both a quantitative and qualitative method. In fact, modern matching methods are highly intuitive and nonparametric, and require no linearity or normality assumptions (Ho et al., 2007). Both camps would benefit from learning more about these methods, as they can be thought of as formalizations of some ideas of causal inference in the qualitative methods literature, and the formalizations and the specific methods can be used to construct better quantitative and qualitative research designs. In fact, all the formalizations that use regression for intuition in King, Keohane and Verba (1994) can be described in the context of matching without regression assumptions.

At its simplest, the idea of matching closely parallels the potential outcomes that need to be estimated. Suppose we are interested in the effect of the form of government in a city, such as an elected mayor vs. a professional city manager, on some measure of social capital. We all know that merely comparing the amount of social capital in cities with elected mayors to the amount in cities with a professional manager may lead to omitted variable bias. For example, suppose that city managers are more prevalent in one region of the country that also happens to have more social capital well before the form of government was decided. Because region is causally prior to the form of government, and may also affect social capital after taking into account the form of government, omitted variable bias is likely.

To avoid this problem via matching is easy. For simplicity, consider this causal effect only for cities with an elected mayor. The causal effect is then defined as the amount of social capital observed (i.e., with an elected mayor) minus the amount of social capital that would have been observed in those same cities if the cities instead had a professional manager. This second potential outcome is not observed and so we estimate it by finding, for each city with an elected mayor, a real city with a professional manager that is matched on region (i.e., is from the same region). We then take the difference in social capital for these two cities, repeat the same matching procedure for each of the cities with elected mayors, and then average the estimated causal effects.

Of course, matching, like regression and all other quantitative and qualitative methods used to estimate causal effects from nonexperimental data, requires that the investigator identify all the potentially confounding variables to control for, not merely one convenient variable as in our example. As always, the potential confounders include all variables which meet three conditions: they are causally prior to the treatment, related to the treatment, and affect the outcome variable after controlling for the treatment; other variables can

be ignored. Avoiding omitted variable bias is a difficult problem in observational data, but it is of course well-known throughout the discipline and so we at least have ways of thinking about and attacking it. Matching and other approaches do not enable researchers to sidestep omitted variable bias, only to avoid making all but the most minimal additional assumptions after the omitted variables are identified and measured.

Thus, this simple version of matching and Mills' method of difference are almost the same procedure. The difference comes when we realize that the two groups of cities differ on more grounds than just region. In fact, its easy to think of dozens and dozens of potential omitted variables in this example that are causally prior to the form of government and may affect the degree of social capital. As a result, finding even two cities that differ in their form of government but are exactly matched on all these other variables quickly becomes impossible. But that is exactly where modern matching methods start to differ from existing methods and begin to make a real difference, since they enable one to match approximately in the best possible ways on numerous variables. Matching methods are also not merely a quantitative analogue to qualitative research, since they can be used quite directly in qualitative research for selecting cases, designing research strategies, and highlighting unacceptable heterogeneity.

Similarly, quantitative scholars also have much to learn from qualitative uses of matching: In many quantitative data sets the identity of each subject is known (such as in comparative politics where the unit is the country or country-year or in international relations where the unit is the conflict or dyad-year). This means that a considerable amount of nonnumerical information exists in quantitative data sets that can be used to improve the quality of matching algorithms. Even a simple step of listing the matches for readers to see will often highlight some matches that can be improved. For this, the best way forward is for quantitative researchers to add some of this type of qualitative research to their repertoire.

So regression is not remotely synonymous with causal inference. Indeed, even in those cases when regression is used with matching, matching makes causal inferences much less sensitive to the assumptions underlying regression and other forms of statistical modeling. It would seem that both qualitative and quantitative scholars can benefit greatly from these methods.

5 Selection

5.1 The Absurdity of Random Selection for Case Study Designs

Randomness has an odd status in the qualitative methods literature. It is after all an extremely powerful concept, possibly the most important development in research methods in the last hundred years. Scholars in the qualitative methods literature invoke the idea in discussing the selection of cases, but almost all invocations are irrelevant or unhelpful. Fortunately, those who apply qualitative methods in empirical research seem to know intuitively that random selection is inappropriate for research designs with a small number of observations, and indeed it is. Here's why.

What randomness does in experimental research is make it possible to select cases, and assign values of the treatment variable, in a manner that protects us, not only from confounding variables we know about, but also from all confounding variables that we do not know about. That it is possible to do something to protect us from unknown evils is a surprising and stunningly important development.

To be more specific, one can make relatively automatic inferences with (1) random

selection of observations from a known population, (2) random assignment of values of the key causal variable, and (3) a large n. All three must be present or the study becomes an essentially observational study with all the assumptions and potential problems of any such research.

But given the benefits of relatively automatic assumption-free inferences, who wouldn't want to jump on the randomness bandwagon? The answer is anyone who can select only a very small set of cases. King, Keohane and Verba (1994, p.124-8) give an example with three observations where random selection gives you the wrong answer two-thirds of the time. Indeed, this is quite general: if you can only afford to select a small number of cases random selection will often be worse than selection by other means. As Section 5.2 explains, if you can only collect a very small number of cases, then any method of selection, including randomness, will not let you generalize with any useful level of certainty. At that point, random selection is besides the point. You either need more information or should focus on the cases at hand.

When collecting more information is infeasible, as it typically is for qualitative case study researchers, go where the information is: within your cases. When it is possible to collect more cases or information at the same level of analysis, then a stratified design will avoid some of the variability induced by randomness. The idea of stratification is to select cases within categories of one or more pre-treatment explanatory variables ("pre-treatment" refers to explanatory variables that are causally prior to your key causal variable). For example, if you select countries randomly, you could in your particular sample happen to get only democratic countries, but if you stratify on democracy, your sample will include exactly as many democratic and nondemocratic countries as you choose ex ante. Stratification is far more efficient strategy when feasible (see Imai, King and Stuart, 2008). In fact, stratification can be thought of as matching (as per Section 4.3) prior to treatment assignment.

5.2 Worry Less about Selecting Cases, More about Analyzing Them

The problem of selection is crucial in all research, but scholars in the qualitative methods literature tend to focus too little on selection *within* case studies and at least relatively too much on which cases to study. That is, they write about how to select (for example) their five case studies, but sometimes seem to act as if anything goes (or perhaps "the researcher knows best") when it comes to collecting information within their cases (see Duneier, 2008). Both of course are crucially important and can determine the answers to the questions posed. But qualitative researchers should *not* feel guilty when selecting a small sample in a nonrandom or unrepresentative way. Inferences in this situation will necessarily be limited to the sample rather than some broader population, but learning something about some part of the world in some prescribed period of time is often valuable in and of itself. This is worth recognizing especially in qualitative research where understanding unique events or processes is often a more desirable goal than generalization to unrelated cases (Mahoney and Goertz, 2006, Sec. 1). It certainly is in numerous other areas, especially when — as with the selection of cases but not necessarily the rich information within cases — samples are limited to small numbers.

Indeed, the vast majority of research of all types — quantitative and qualitative, observational and experimental, and in every field of study — selects its major units of analysis based on convenience, personal interest, and available resources. There is of course always some attention to issues of generalization, but the main goal of much research is typically to obtain some useful information about some sample. Even the most expensive,

high quality, randomized clinical trials in medicine usually select experimental subjects based on who happens into a particular hospital, answers a newspaper ad, is willing to be enrolled in a study, and impresses a researcher into thinking that they will respond to the treatment. Most observational researchers collect whatever data are convenient. In comparative politics studies, if more countries were available, or additional data were easy to collect for more years, they would almost certainly be included. Even sample surveys do not involve truly random selection from known, enumerated populations, although at least they often do attempt it. In most other research, scholars do not even try. Instead, they focus on learning what they can about the observations at hand, and there should be no shame in doing so.

Statistical researchers often explicitly attempt to estimate only the *sample average treatment effect* — the effect of some cause in the sample at hand — rather than the *population average treatment effect* — the causal effect in some specified broader population (Imai, King and Stuart, 2008). Making inferences about the former is often considerably easier, less uncertain, and more well-defined than the latter. If other researchers make sample-based inferences about different areas or time periods, then the collective product from all these studies can be considerably greater than any one person's research, which of course is the signal contribution, if not the definition, of a scientific community. And whether or not you are interested in the population of all major international crises or all terrorist incidents, a study that only tells us about why the Cuban Missile Crisis or the 9/11 terrorist incidents happened can still be exceptionally important. There's no reason to feel inferior about the lack of representativeness of your study: if the small number of cases you have collected are truly important, then they are not representative almost by definition.

Moreover, even if it were possible to collect a truly random sample of a very small number of observations, random selection might well not be advisable. After all, a small n means that the uncertainty due solely to sampling variability would be huge. Even without measurement error or any other inferential problem, we would not learn very much. This of course does not mean that qualitative studies with few cases are meaningless. On the contrary, it just means that the vast majority of the information learned in them is within cases about the sample treatment effect rather than at the case-level about population treatment effects.

Thus, once one narrows the inferential target to features of their particular cases, they must be exceptionally careful to select information to be measured and recorded in a systematic and unbiased way. Here — within cases — is where the vast majority of thought about selection issues should be put by qualitative researchers. If that doesn't happen, then not only might you lose the ability to generalize to some other population; you might lose the ability to learn even about your cases. At that point, there is little sense in having conducted the study at all. And *within cases, all the same issues of selection bias apply that are now focused on case selection for inferring to a larger population.* So there should be considerable effort devoted to convincing oneself and readers that the method of selecting information within cases is representative of those cases. It is *not* sufficient to say that you went and interviewed several people without clarifying how they were selected, or that you wandered around or "soaked and poked" without delinating how you chose your path for wanding, where you soaked, and what the process was by which you poked. The same goes for the rules by which you might conduct participant observation: the research is not about you; its about the the world you are studying which requires understanding the process by which you selected evidence within your cases. Haphazard selection within cases is not necessarily representative of that case. Readers need documented evidence

supporting exactly how the within-case information was collected, how it might be thought of as representative of these cases, and what might have been missed.

For these reasons, if you can only collect a few cases to study, then you should consider it reasonable to go ahead and select the cases that interest you, or those which will help you answer or generate your theoretical questions, or those which are most important or which would best inform the literature. Do not expect a small number of cases to be representative of a much larger population, but don't worry about it too much either. But be absolutely sure to get your cases right and to plan, understand, execute, and communicate the process by which your data were collected.

5.3 How and Why to Select on Y

"Do not select on the dependent variable" is a well known rule in the quantitative and qualitative methods literatures. Its also appropriate since selecting on Y can bias descriptive and causal inferences (King, Keohane and Verba, 1994, ch.4). However, the same rule causes great consternation among case study qualitative researchers who understandably do not want to miss the Cuban Missile Crisis when selecting their five crises to study or World War II when selecting international conflicts. So what gives? Do they really need to select cases without regard to the most important issues they care about?

Qualitative researchers would be well advised to learn more about *case-control studies*. This data collection design enables you to select on your outcome variable, such as some number of crises and the same number of non-crises. Then the usual rule that prohibits one from selecting on the dependent variable does not apply *if* you do not select on the explanatory variable and you correct in a particular way. (If you select cases based on the values of both your dependent and explanatory variables, nothing is left to learn from the data, so you must be sure not to select on both.) If you select your data on Y rather than randomly, on X, or in some other way, you must apply a case-control correction, such as described in King and Zeng (2001). In quantitative research, these methods can save 99% or more of data collection costs — so that for example one need not measure a new variable for all pairs of countries for all years in the last century (with an n of approximately 1 million); instead it is possible to get approximately the same empirical result by collecting only the 1,000 or so international conflicts and a set of 1,000 or 2,000 non-conflictual dyads. The same issue applies in qualitative research, even if enumerating all possible examples of the cases of interest is not possible.

Qualitative researchers will of course not be able to make this correction quantitatively, but understanding the simple statistical correction will make it easy to follow the direction and general magnitude of the bias and what can be learned from this data collection design. It is not difficult, and it can save enormous data collection resources when the events of interest are rare.

5.4 Formal Methods to Avoid Conceptual Stretching

With all the effort spent on increasing the number of observable implications of a theory, its worth pausing to ensure that each observation is closely related to the theory at issue. Including observations only vaguely related to the theory or causal inference is wasteful of time and resources at best and can induce bias at worst. In the qualitative literature, the term *conceptual stretching* refers to "the distortion that occurs when a concept does not fit the new cases" (Collier and Mahon, 1993, p.845). To avoid conceptual stretching, qualitative scholars attempt to select cases that fit their categories and carefully adjust their categories or concepts to fit their cases. Conceptual stretching is not only important

in designing research but is also at center stage in a large intradiscplinary dispute generated by severe criticism qualitative scholars have levied at cross-national statistical studies over the last half-century ("no branch of political science has been in more extreme ferment than comparative politics during the last fifteen years"; see LaPalombara 1968, p.52 and Girosi and King 2008, Section 1.5). Since these early days, qualitative scholars have branched out from area studies to be more comparative in more careful ways, and quantitative scholars have developed methods to avoid observations that do not belong in the same data. The connections between these two approaches are not well known in the two literature but deserve to be.

For example, King and Zeng (2006, 2007) proved mathematically that when a statistical quantity of interest is far from the data, inferences are more *model-dependent*, which means that small, indefensible changes in statistical assumptions can lead to unacceptably large differences in empirical conclusions. This proof was designed for quantitative work, but it also applies directly to the problem of conceptual stretching in qualitative work. It gives some precision to the qualitative notion that the farther you stretch a concept by applying it to new cases from distant (conceptual) lands, the more untenable are the assumptions that would need to be defended and justified in order to shore up a claim that one's inferences are still valid.

Building on these results, statistical methods have now been developed to evaluate how far a counterfactual or other target of inference is from some or all of one's data, and in some cases they indicate exactly how far is too far (Ho et al., 2007; King and Zeng, 2007). Quantitative scholars now routinely either change the quantity of interest to one that is possible to learn about without much model dependence or they prune their data set of observations that cause the model dependence. In some situations, this pruning process changes the quantity of interest in a similar manner as qualitative scholars do when they adjust the data and concepts to fit each other. It seems that much benefit would accrue to both the qualitative literature on conceptual stretching and the quantitative literature on model-dependence if methodologically oriented scholars, and applied researchers, in both fields became better aware of the parallel and complementary developments in the two areas.

5.5 Documenting Selection Rules

Even more important than unbiased selection methods is documenting, and making available to readers, exactly how you gathered the data you are offering as evidence. In this way, if you collect data in a biased way, all is not lost: you can still correct and produce unbiased inferences.

For example, if you are interviewing current and former administration officials, we know that their answers will tend to be biased in ways that favor themselves and their administrations. If in analyzing or discussing the results of the interviews you do not reveal which administration each respondent worked for, then there is an obvious source of inferential bias. If, instead, we merely made available this information, then at least some of the bias could be corrected. Of course, by paying attention to the relationship between where people sit and where they stand, we do this almost automatically all the time. But its only possible to correct bias if we know the process by which the data were collected. And whether the analysis was qualitative or quantitative is irrelevant; the same issues apply.

This point is very general: A valid inference cannot be made from a given set of data without knowing the whole chain of evidence leading from the world to the qualitative

conclusions in final written work. Any break in the chain of evidence, any missing piece of information about how the subjects were chosen, data were collected, or information was unearthed, can lead to large biases for a given inferential target.

Quantitative researchers are now well aware of this issue and have been moving to fix it in institutional ways. The replication or data sharing movement has spread throughout the physical, natural, and social sciences. This movement supports what is know as "The *replication standard* [which] holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author" (King, 1995). Some journals ask authors to include sufficient detail in their published work about the source of data and selection criteria. In other types of scholarship, page limitations make this infeasible, and so many journals require, and others strongly encourage, scholars to create and distribute "replication data sets" along with every published article (Fienberg, Martin and Straf, 1985; King, 2007). Many of these replication data sets are frequently analyzed by other scholars; and students, even in class papers, are able to get right to the cutting edge of academic research via replication assignments and often publish new articles themselves (King, 2006).

Of course, putting together replication data sets for qualitative research projects can be much more difficult than uploading a numerical data set, some documentation, and computer code. Many of the advantages of exploring an archive, conducting an ethnography, or doing participant observation come from learning the most interesting questions to ask or from discovering sources of information not foreseen ex ante. These difficulties may explain why qualitative researchers seem to devote less (and generally insufficient) time and effort to making public the chain of evidence for their studies. And few make available replication data sets.

Yet, qualitative researchers can easily make some types of field notes, oral interviews, and videotapes available, and well established procedures now exist for selectively sharing this type of information with researchers, no matter how confidential, proprietary, sensitive, or secret it may be (e.g., `http://thedata.org`). Professional archives have been established to accept and permanently preserve this information (such as ESDS Qualidata and the Murray Research Archive).

Thus, making qualitative data available certainly is possible, although the deep, diverse, and extensive nature of evidence in many qualitative studies makes preparing a complete replication data set difficult. Even so, making the effort in your case, developing facilities to help researchers make their data available, and teaching students to do so is crucial for two fundamental reasons. First, and most obviously, research for which the chain of evidence is irreparably broken is itself irreparably flawed: if readers do not know the process by which the researcher came to observe the data presented, they cannot assess for themselves whether to believe the conclusions drawn on the basis of it. And trust, without possibility of verification, is not a serious foundation for a scientific enterprise.

Second, developing easier procedures for researchers to make qualitative data available is crucial, not only for qualitative researchers but also for quantitative researchers. Since all quantitative studies are also qualitative to some degree, all social scientists could use a great deal of help in learning how to record, document, systematize, index, and preserve the qualitative information associated with their studies, and qualitative researchers are well positioned to lead this effort.

In fact, technology is turning many types of quantitative efforts into much richer and more qualitative studies. For example, imagine continuous time location information from cell phones linked to survey responses, information about proximity with other cell phone

users, recorded voice-to-text and text analytic data, photographs taken from the camera phone, emails and text messages sent and received, and phone calls made. The deep contextual information available from this type of information, and many others emerging like it, is in great need of study from the perspective of replication, data sharing, and preservation — as well as asking the right questions and collecting the most useful evidence. For another example, sample surveys can record interviews, and we can preserve the audiotape. Even the most basic cross-national comparative quantitative data summaries have behind them numerous qualitative stories not fully documented about how each datum was created and measured. Qualitative and quantitative researchers need to understand how to best document, preserve, and distribute all these and other types of qualitative information, as all form crucial parts of the chain of evidence for social science analyses.

6 Concluding Remarks

The issues and errors we highlight here in the qualitative methods literature, and in the uses of qualitative methods in the substantive applied fields of political science, all have analogues in quantitative political science. All the issues we address are readily fixed by understanding the theory of inference underlying all our work, and in most cases by connecting work in qualitative methods with better developed quantitative methods or by work in quantitative methods with better developed qualitative approaches. We have touched on issues of path dependence, case selection, selection bias, observable implications, obstacles to inference, causality, selection, theory formation, implications of bias, the relationship between qualitative and quantitative methods, and several others. But these are only examples. Many other such issues remain, and we encourage scholars to find these, highlight the similarities and connections across our disparate fields, and to continue to improve empirical research throughout our discipline.

References

Adcock, Robert and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3, September):529–546.

Ayres, Ian. 2007. *Supercrunchers*. New York: Random House.

Beck, Nathaniel and Jonathan Katz. 1996. "Nuisance vs. Substance: Specifying and Estimating Time-Series-Cross-Section Model." *Political Analysis* VI:1–36.

Braumoeller, Bear F. 2003. "Causal Complexity and the Study of Politics." *Political Analysis* 11:209–233.

Braumoeller, Bear and Gary Goertz. 2000. "The Methodology of Necessary Conditions." *American Journal of Political Science* 44(4, October):844–858.

Campbell, James E. 2005. "Introduction: Assessments of the 2004 Presidential Vote Forecasts." *PS: Political Science & Politics* 38:23–24.

Collier, David and James E. Mahon, Jr. 1993. "Conceptual 'Stretching' Revisited." *American Political Science Review* 87(4, December):845–855.

Duneier, Mitchell. 2008. How Not to Lie with Ethnography. Technical report Princeton University.

Elman, Colin. 2005. "Explanatory Typologies in Qualitative Studies of International Politics." *International Organization* 59(2, spring):293–326.

Fienberg, Stephen E., Margaret E. Martin and Miron L. Straf. 1985. *Sharing Research Data*. National Academy Press.

Gelman, Andrew and Gary King. 1993. "Why are American Presidential Election Campaign Polls so Variable when Votes are so Predictable?" *British Journal of Political Science* 23(1, October):409–451. http://gking.harvard.edu/files/abs/variable-abs.shtml.

George, A.L. and A. Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Mit Press.

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

Gill, Jeff. 2008. *Bayesian Methods: A Social and Behavioral Sciences Approach, 2nd edition*. Chapman & Hall/CRC.

Girosi, Federico and Gary King. 2008. *Demographic Forecasting*. Princeton: Princeton University Press. http://gking.harvard.edu/files/smooth/.

Glynn, Adam and Kevin Quinn. 2008. Non-parametric Mechanisms and Causal Modeling. Technical report Harvard.

Goertz, Gary. 2003. The Substantive Importance of Necessary Condition Hypotheses. In *Necessary Conditions: Theory, Methodology, and Applications*, ed. Gary Goertz and Harvey Starr. Lanham, MD: Rowman & Littlefield.

Goldthorpe, J.H. 2001. "Causation, Statistics, and Sociology." *European Sociological Review* 17(1):1–20.

Granato, Jim and Frank Scioli. 2004. "Puzzles, Proverbs, and Omega Matrices: The Scientific and Social Significance of Empirical Implications of Theoretical Models (EITM)." *Perspectives on Politics* 2(02):313–323.

Grove, William M. 2005. "Clinical Versus Statistical Prediction: The Contribution of Paul E. Meehl." *Journal of Clinical Psychology* 61(10):1233–1243.

Hall, Peter A. 2006. "Systematic Process Analysis: When and How to Use It." *European Management Review* 3(1, Spring):24–31.

Hall, Peter A. 2009, forthcoming. Path Dependence. In *The Future of Political Science: 100 Perspectives*, ed. Gary King, Kay Scholzman and Norman Nie. Routledge.

Hamilton, James Douglas. 1994. *Time Series Analysis*. Princeton: Princeton University Press.

Heckman, James J. 2008. Econometric Causality. Technical Report 13934 National Bureau of Economic Research Cambridge, MA: . http://www.nber.org/papers/w13934.

Ho, Daniel, Kosuke Imai, Gary King and Elizabeth Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15:199–236. http://gking.harvard.edu/files/abs/matchp-abs.shtml.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.

Huff, Darrell. 1954. *How to Lie With Statistics*. New York: WW Norton & Company.

Imai, Kosuke, Gary King and Elizabeth Stuart. 2008. "Misunderstandings Among Experimentalists and Observationalists about Causal Inference." *Journal of the Royal Statistical Society, Series A* 171, part 2:481–502. http://gking.harvard.edu/files/abs/matchse-abs.shtml.

Imai, Kosuke, Gary King and Olivia Lau. 2007. "Toward A Common Framework for Statistical Analysis and Development.". http://gking.harvard.edu/files/abs/z-abs.shtml.

King, Gary. 1995. "Replication, Replication." *PS: Political Science and Politics* 28(3, September):443–499. http://gking.harvard.edu/files/abs/replication-abs.shtml.

King, Gary. 2006. "Publication, Publication." *PS: Political Science and Politics* 39(01, January):119–125. http://gking.harvard.edu/files/abs/paperspub-abs.shtml.

King, Gary. 2007. "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing." *Sociological Methods and Research* 36(2):173–199.

http://gking.harvard.edu/files/abs/dvn-abs.shtml.

King, Gary and Langche Zeng. 2001. "Explaining Rare Events in International Relations." *International Organization* 55(3, Summer):693–715. http://gking.harvard.edu/files/abs/baby0s-abs.shtml.

King, Gary and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2):131–159. http://gking.harvard.edu/files/abs/counterft-abs.shtml.

King, Gary and Langche Zeng. 2007. "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference." *International Studies Quarterly* (March):183–210. http://gking.harvard.edu/files/abs/counterf-abs.shtml.

King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press. http://www.pupress.princeton.edu/titles/5458.html.

King, Gary and Ying Lu. 2008. "Verbal Autopsy Methods with Multiple Causes of Death." *Statistical Science* 23(1):78–91. http://gking.harvard.edu/files/abs/vamc-abs.shtml.

LaPalombara, Joseph. 1968. "Macrotheories and Microapplications in Comparative Politics: A Widening Chasm." *Comparative Politics* (October):52–78.

Lieberman, Evan S. 2005. "Nested Analysis as a Mixed-Method Strategy for Comparative Research." *American Political Science Review* 99(3, August):435–452.

Little, Daniel. 1991. *Varieties of Social Explanation: An Introduction to the Philosophy of Social Science*. Westview Press.

Mahoney, James. 2008. "Toward a Unified Theory of Causality." *Comparative Political studies* 41(4/5, April/May):412–436.

Mahoney, James and Gary Goertz. 2006. "A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research." *Political Analysis* 14(3):227–249.

Martin, A.D., K.M. Quinn, T.W. Ruger and P.T. Kim. 2004. "Competing Approaches to Predicting Supreme Court Decision Making." *Perspectives on Politics* 2(04):761–767.

Meehl, Paul E. 1954. *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis: University of Minnesota Press.

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

Pierson, Paul. 2000. "Increasing Returns, Path Dependence, and the Study of Politics." *American Political Science Review* (June):251–268.

Steele, J. Michael. 2005. "Darrell Huff and Fifty Years of *How to Lie With Statistics*." *Statistical Science* 20(3):205–209.

Tetlock, Philip E. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton: Princeton University Press.