IS THERE MORE THAN ONE LOGIC OF CAUSAL INQUIRY?

MOVING BEYOND THE QUAL/QUANT DEBATE

John Gerring
Department of Political Science
Boston University
232 Bay State Road
Boston MA 02215
Phone: 617-353-2756
Fax: 617-353-5508
jgerring@bu.edu

Craig W. Thomas
Daniel J. Evans School of Public Affairs
University of Washington
205 Parrington Hall, Box 353055
Seattle, WA 98195-3055
thomasc@uwashington.edu
Phone: 206-221-3669

Approximate Word Count: 6,749
ABSTRACT

This paper argues that the underlying reason for the persistence of qualitative and quantitative techniques of causal analysis lies in the comparability of observations. Quantitative observations can be arrayed in dataset formats because they are readily comparable. Qualitative observations, by contrast, are not comparable to one another, and therefore cannot be analyzed in a dataset format. Traditional distinctions between qualitative and quantitative methods – such as small-N versus large-N, narrative versus statistical analysis – are, to some extent, simply epiphenomena of the more fundamental distinction between comparable and non-comparable observations. Analyzing non-comparable observations requires a different logic of causal analysis that does not conform to standard notions of formal methodological rigor. Nevertheless, we show that the inferences drawn from informal analyses of non-comparable observations may be more secure in some instances than inferences based on quantitative methods.
The distinction between qualitative and quantitative methods has long vexed the disciplines of social science. In contrast to the presumed unity and coherence of natural science, knowledge in these fields seems less cumulative. Often, scholars reared in different methodological traditions do not read – much less integrate – each other’s work. Often, subjects are broached in such disparate terms that the results seem incommensurable.

Despite recent moves toward better integration of methods and approaches (e.g., Brady & Collier 2004; Mahoney 2000; Ragin 2000; see also Political Analysis 14:3, Summer 2006), the disciplines of political science and sociology (not to mention anthropology) often give the appearance of a large banquet with “separate tables.”

Among these divisions perhaps the most recalcitrant and the most consequential is the division between quantitative and qualitative methods, a division that harkens back to the early twentieth century (Glassner & Moreno 1989). Today, these camps are ensconced in separate sections of the American Political Science Association and the American Sociological Association, separate training camps and methods courses for graduate students, and, in some instances, separate journals for published research.

What is the basis of these divisions? Do qualitative and quantitative methods follow divergent logics of inquiry? Work emanating from the quantitative camp has generally denied that such differences exist, or has questioned whether such differences as do exist can be justified methodologically (Blalock 1982, 1989; Friedman 1953; Goldthorpe 2000; King, Keohane & Verba 1994; Lazarsfeld & Rosenberg 1955; Lieberson 1985; Wilson 1998). Studies emanating from the qualitative camp demur (e.g., Becker 1996; Bennett & Elman 2006; Brady & Collier 2004; Bryman 1984; Caporaso 1995; George & Bennett 2005; Glassner & Moreno 1989; Glaser & Strauss 1967; Mahoney & Goertz 2006). But there is no
consensus within the qualitative camp upon what, precisely, differentiates the qualitative enterprise.

In this paper, we survey the various rationales that have been offered for the existence of “two logics” (qual and quant) of causal analysis. We argue that these rationales have overlooked a fundamental difference between quantitative and qualitative research practices. Specifically, we argue that distinctively qualitative techniques of causal analysis have persisted – in the face of the proliferation of an ever-expanding menu of statistical techniques – because, in many research contexts, available observations are not comparable to one another. Hence, they cannot be analyzed in a standard dataset format. In these circumstances, a qualitative approach to causal analysis is difficult to avoid, and virtually impossible to improve upon. We argue, therefore, that the traditional contrasts drawn between qualitative and quantitative methods – small-N versus large-N, narrative versus statistical analysis, and so forth – are, to some extent (though of course not exclusively), epiphenomena of this fundamental distinction between comparable and non-comparable observations. Whether this distinction constitutes a difference in underlying causal “logic” is another matter, a discussion deferred to the conclusion of this paper.

**Where are the Apples and Oranges?**

Before beginning, it will be helpful to clarify the meaning of two key terms. “Qualitative” is understood as a study, or a proposition lying within a study, for which the evidence is presented and evaluated largely in prose. “Quantitative” will refer to a study, or a proposition lying within a study, that relies largely on statistical analysis. Thus, when we say that a study is qualitative or quantitative we are referring to a predominant tendency – the main evidence for the main argument. It is true that many studies rely on a mixture of
evidentiary types. Even so, most are readily classifiable in one category or the other. In any case, the key point of our discussion rests on distinguishing among logics of evidence. Whether, or to what extent, these two logics are combined in a single book or article is not directly germane to the argument. Our point is that they are methodologically distinct.

In order to satisfy an argument for methodological distinctiveness it must be demonstrated that differences between qualitative and quantitative logics of research are more than aesthetic preferences. They must also be methodologically justifiable. If this claim cannot be sustained, then it would appear that the mainstream methodological perspective is probably correct: there is only one logic of inquiry for causal analysis; and it is, presumably, that which is practiced by the dominant (quantitative) faction in social science today.

We begin with a brief discussion of the literature in qualitative methods that has accumulated in recent years within the disciplines of political science and sociology. This literature is organized around a variety of overlapping terms including process tracing, causal-process observations, pattern-matching, causal-chain explanation, colligation, congruence method, genetic explanation, narrative explanation, sequential explanation, and interpretation/hermeneutics. Revealed in this diverse body of work are a series of oppositions intended to explain and (at least implicitly) to justify the qual/quant division in social science research. They are, as follows: a) the size of the sample (small-N versus large-N), b) the objective of the study (elucidating mechanisms versus covariational patterns), and c) the style of evidence-gathering (contextualized versus decontextualized).

The first distinction is arguably the most common: qualitative work is based on small-N samples, while quantitative work is large-N. Indeed, large samples are typically analyzed with quantitative techniques (such as statistical correlations), while small samples are typically examined with qualitative techniques (such as narrative case studies). However,
the choice of a small-sample research design is not one that has any intrinsic methodological merit (unless the small sample design allows for more information to be collected for each case study and this extra information substantially strengthens the research design). Small samples are not (by any methodological logic) preferable to large samples. Less is not more. Moreover, it may be questioned whether the logic of analysis employed in small-sample research is really different than that employed in large-sample research, a matter to which we return below. Thus, this initial claim for separate logics is not sustained by the distinction between small-N and large-N samples.

Qualitative work is sometimes identified with the study of causal mechanisms – the causal pathways that lead from a particular causal factor (X) to a particular outcome (Y) (Seawright & Collier 2004: 177-8; George & Bennett 2005). While this may be true as a matter of emphasis it is certainly not true as a matter of definition. All causal analysis strives to identify a set of causal mechanisms; this is the near-consensus among recent work on causation (Gerring 2005). While quantitative studies typically focus on X/Y covariational patterns rather than causal mechanisms, some quantitative studies are more successful than qualitative studies in achieving insights into causal mechanisms. This is often the case where pathways are easily measured across multiple cases, or through multiple observations within a single case (as is done with interrupted time-series analyses such as ARIMA, where interventions, such as new public policies, are understood to be the causal mechanism). Moreover, all qualitative studies of causal relationships presume a covariational pattern between X and Y. Even if such studies focus on only a single example of that relationship (a single case), they are never purely “mechanistic.” Indeed, a causal mechanism has no meaning in isolation from the covariation between an exogenous causal factor (X), or factors, and an outcome (Y), or outcomes. This argument for two logics based on causal
mechanisms versus covariational analysis is clearly false (or must be understood as a matter of emphasis rather than a matter of classification).

Qualitative work has also been identified with a “holistic” style of research that emphasizes “contextual” factors in a causal relationship (Seawright & Collier 2004: 177-8). While this suggests an intuitive distinction vis-à-vis quantitative work, there is a serious question about how these two concepts (“holistic” and “contextual”) should be understood and about their relevance to the analytic process. If a contextual factor is extraneous to the causal relationship of interest then presumably it is extraneous to the argument; as such, it cannot be considered a justifiable logic of causal analysis. If a contextual factor is instrumental to the causal relationship of interest then it is relevant; however, there is no reason to presume that it cannot be modeled in a quantitative format. All background causal factors are, in this sense, contextual. (If these contextual factors are understood as unique features of each unit under analysis then they may be modeled as “fixed” effects.) Thus, it might be said that both quantitative and qualitative analysis is contextual.

Although we do not have space for a longer review (of what is, it should be noted, quite a large literature), it would appear that some of the most frequently vetted arguments for a distinctively qualitative methodology do not pan out. A distinctive approach to social science cannot be founded on the size of a sample, a mechanistic objective, or a contextualized analysis of evidence. Either there is something else driving the distinction between qualitative and quantitative forms of causal analysis or we shall be compelled to concede that the venerable distinction cannot be justified, and – by extension -- should not be maintained.
**Two Logics of Analysis**

Let us begin by characterizing what will be called the “standardized” social science research design. Here, the analyst enlists a number of comparable observations in order to tease out the relationship between a causal factor of interest (X) and an outcome (Y). Comparable observations are understood to be comparable relative to the causal hypothesis of interest; they are unit homogeneous. The resulting research design might be spatial or temporal (or both), within-case or across-case, experimental or observational. It may pertain to causal mechanisms lying between X and Y or to inputs and outputs (measures of X and Y). Most importantly, it may be small-N or large-N, and hence may comport with either a quantitative or qualitative analysis. Millian research designs are an example of the latter. They build on a small-N sample of comparable observations, understood to reflect most-similar or most-different cases. No statistical analysis is required – though, if the sample were enlarged, it could easily be performed.

A second logic of analysis is unique to qualitative analysis. It cannot be handled in a large-N framework precisely because the pieces of evidence are non-comparable to one another – apples and oranges, as the saying goes. Because the subject matter is complicated – it is, at once, tediously familiar and strangely exotic – we begin with a series of examples. These examples are chosen to illustrate diverse types of research designs drawing on non-comparable bits of evidence and to clarify the contrast with the standardized research design template. They are also well known to many readers, having been the subject of previous methodological discussion, and thus serve a heuristic function.

As an initial example, let us consider Pressman and Wildavsky’s (1973) study of policy implementation. The authors follow the implementation of a federal bill, passed in 1966, to construct an airport hangar, a marine terminal, a 30-acre industrial park, and an
access road to a major coliseum in the city of Oakland, California. The authors point out that this represents free money for a depressed urban region. There is every reason to assume that these projects will benefit the community and every reason – at least from an abstract public interest perspective – to suppose that the programs will be speedily implemented. Yet, three years later, progress was agonizingly slow and few projects had actually been completed. The explanation provided by the authors rests upon the bureaucratic complexities of the American polity. Pressman and Wildavsky show that these small and relatively specific tasks undertaken by the federal government necessitated the cooperation of seven federal agencies (the Economic Development Administration [EDA] of the Department of Commerce, the Seattle Regional Office of the EDA, the Oakland Office of the EDA, the General Accounting Office, HEW, the Department of Labor, and the Navy), three local agencies (the Mayor of Oakland, the city council, and the port of Oakland), and four private groups (World Airways Company, Oakland business leaders, Oakland black leaders, and conservation and environmental groups). These fourteen governmental and private entities had to agree on at least seventy important decisions in order to implement a law initially passed in Washington. James Q. Wilson observes, “It is rarely possible to get independent organizations to agree by ‘issuing orders’; it is never possible to do so when they belong to legally distinct levels of government.” The plausible counterfactual is that with a unitary system of government, these tasks would have been accomplished in a more efficient and expeditious fashion.

This renowned study rests largely upon a demonstration of proximal relationships between key actors. The authors show, for example, that there is resistance to federal directives from local political leaders who have their own agendas and often do not see eye-to-eye with the Washington bureaucrats assigned to implement the construction of public
works projects in Oakland. Interviews with these actors, as well as their own public statements, bolster the counterfactual reasoning of the book. These local actors have different perspectives because they have different constituencies, different organizational norms, and consequently different incentive structures. All of this, including their ability to resist federal directives, may be considered as a product of the constitutional and statutory structure of the American polity. Without federalism, and without the local bureaucratic independence and multiplication of agencies with overlapping jurisdictions that stems from a federal constitution, things would have been very different. Thus, a series of idiosyncratic observations is enlisted to demonstrate a macro-causal claim applying not just to the US at-large but to democratic polities everywhere.

As a second example, consider Theda Skocpol’s theory of social revolution, which hinges critically upon the breakdown of the French state in the decades leading up to 1789. James Mahoney diagrams this element of the argument in meticulous detail, identifying three general causal factors – agrarian backwardness, international pressure, and state autonomy – which are, in turn, broken down into thirty-seven discrete steps. The entire argument is reproduced in Figure 1. For our purposes, what is noteworthy is that the evidence for each step in this causal chain is unique, which is to say that the evidence mustered to prove (1) is different in character from the evidence adduced for step (2), and so forth all the way down the line. Each is a separate argument, nested within a larger argument about the causes of state breakdown in France in 1789. And this, in turn, is nested within a larger argument about social revolution in the modern world. Mahoney (1999: 1168) points out that Skocpol’s overall theory is rendered more plausible by her ability “to order numerous idiosyncratic features of French, Russian, and Chinese history into meaningful accounts of unfolding processes that are consistent with a broader, overarching macrocausal argument.”
These carefully constructed narratives allow for an account of causal mechanisms that simply would not be possible were Skocpol restricted to quantitative techniques.
Figure 1:
Skocpol’s Explanation of Breakdown of the French State (1789)

As a final example, consider Henry Brady’s (2004: 269-70) reflections on his study of the Florida election results in the 2000 presidential election. In the wake of this close election at least one commentator – John Lott (2000), an economist – suggested that because several networks called the state for Gore prior to a closing of the polls in the Panhandle section of the state, this might have discouraged Republican voters from going to the polls, and therefore might have affected the margin (which was razor thin and bitterly contested in the several months following the election). Lott reaches his conclusions on the basis of a regression analysis of turnout in all sixty-seven Florida counties over the course of four presidential elections, with a collection of controls (including fixed year and county effects).

Brady is unconvinced by the method, and the results. Instead, he stitches together isolated pieces of evidence in an “ad hoc” fashion. He begins with the timing of the media calls – ten minutes before the closing of the polls in the Panhandle. “If we assume that voters go to the polls at an even rate throughout the day,” Brady continues, “then only 1/72^n (ten minutes over twelve hours) of the [379,000 eligible voters in the panhandle] had not yet voted when the media call was made.” This is probably a reasonable assumption. (“Interviews with Florida election officials and a review of media reports suggest that, typically, no rush to the polls occurs at the end of the day in the panhandle.”) This means that “only 4,200 people could have been swayed by the media call of the election, if they heard it.” He then proceeds to estimate how many of these 4,200 might have heard the media calls, how many of these who heard it were inclined to vote for Bush, and how any of these might have been swayed, by the announcement, to go to the polls in the closing minutes of the day. Brady (2004: 269-70) concludes: “the approximate upper bound for
Bush’s vote loss was 224 and...the actual vote loss was probably closer to somewhere between 28 and 56 votes.”

Brady’s conclusions rest not on a formal research design but rather on isolated observations (both qualitative and quantitative) combined with deductive inferences. How many voters “had not yet voted when the media called the election for Gore? How many of these voters heard the call? Of these, how many decided not to vote? And of those who decided not to vote, how many would have voted for Bush?” (Brady 2004: 269). This is a very different kind of study than the foregoing studies of implementation and revolution. Yet, the approach has certain similarities insofar as Brady and colleagues enlist a diverse set of individual observations that do not fall neatly into a standardized research design, and they rely heavily on inferential reasoning about specific causal pathways to link these diverse observations to reach their conclusions.

**Comparable and Non-Comparable Observations**

What is distinctive about the foregoing examples is that researchers rely on pieces of evidence – which may be plentiful or minimal – that are not comparable to each other. Each observation is relevant to the central argument (they are not random), but the observations represent different populations (of people or things). They are more correctly understood as a series of N=1 samples from different populations, which means the observations are not readily comparable with one another. Brady’s observation about the timing of the call – ten minutes before the closing of the poll – is followed by a second observation, the total number of people who voted on that day, and a third and a fourth. Although the procedure seems messy, we are convinced by its conclusions. Thus, it seems reasonable to suppose that, in some circumstances at least, causal inferences based on non-comparable
observations are more scientific than inferences based on comparable observations, even though the “method” borders on the ineffable. Our confidence rests on specific propositions and specific observations; it is, in this sense, ad hoc. Indeed, there appears to be little one can say, in general, about the research designs employed by Pressman and Wildavsky, Skocpol, and Brady. While other methods can be understood according to their covariational properties, non-comparable observations invoke a more complex logic, one analogous to detective work, legal briefs, journalism, and traditional historical accounts. The analyst seeks to make sense of a congeries of disparate evidence, some of which may shed light on a single event or decision. The research question is always singular, though the ramifications of the answer may be generalizable. Who shot JFK? Why did the US invade Iraq? What caused the outbreak of World War One? Non-comparable observations are, by definition, case-based. There is no covariational analysis (other than counterfactual analysis), because there is no variation in a sample of N=1.

Note that the explanation drawn from an account based on non-comparable observations may be quite general. Skocpol’s explanatory sketch enlists the minutiae of French history to demonstrate a much larger, macro-theoretical account pertaining to all countries that are wealthy and independent (non-colonies) in the modern era. Even so, the use of non-comparable observations is limited to the explanation of a single case – or, more accurately, to one case at a time.9

Note also that the descriptive statement contained within a non-comparable observation may be either qualitative or quantitative. Indeed, all three of the examples discussed above involve evidence that is both quantitative and qualitative. However, because each quantitative observation is quite different from other quantitative (and qualitative) observations they do not collectively constitute a sample from a single population. Each
observation is sampled from a different population. This means that each quantitative observation is qualitatively different. It is thus the non-comparability of adjacent observations, not the nature of individual observations that differentiates qualitative and quantitative modes of causal analysis.

Note, finally, that because each observation is qualitatively different from the next, the total number of observations in a study is indeterminate. Indeed, the cumulative number of observations may be quite large. However, because these observations are not well defined, it is difficult to say exactly how many there are. Non-comparable observations are, by definition, difficult to count. In an effort to count, one may resort to lists of what appear to be discrete pieces of evidence. This approximates the numbering systems employed in legal briefs (e.g., “there are fifteen reasons why X is unlikely to have killed Y”). But lists can always be composed in multiple ways, and each individual argument carries a different weight in the researcher’s overall assessment. So the total number of observations remains an open question. We do not know, and by the nature of the analysis cannot know, precisely how many observations are present in the studies by Pressman and Wildavsky, Skocpol, and Brady, or in similar accounts such as Richard Fenno’s Homestyle (Fenno 1978), Herbert Kaufman’s The Forest Ranger (Kaufman 1960) or Clifford Geertz’s Negara (Geertz 1980).

Non-comparable observations are not different examples of the same thing; they are different things. Consequently, it is not clear where one observation ends and another begins. They flow seamlessly together. We cannot re-read the foregoing studies with the aid of a calculator and hope to discover the total number of observations, nor would we gain any analytic leverage by so doing. Quantitative researchers are inclined to assume that if observations cannot be counted they must not be there, or – more charitably – that there must be very few of them. Qualitative researchers may insist that they have many “rich”
observations at their disposal, which provide them with the opportunity for thick description or contextual analysis. But they are unable to say, precisely, how many observations they have, or where these observations are, or how many observations are needed for thick description or contextual analysis. Indeed, the observations themselves remain undefined.

This ambiguity is not necessarily troublesome, for the number of non-comparable observations in a study does not bear directly on the study’s usefulness or truthfulness. While the number of observations in a sample drawn from a well-defined population contains information directly relevant to any causal inferences that might be drawn from that sample, the number of non-comparable observations in a study (assuming one could estimate the N) has no obvious relevance to causal inferences that might be drawn from that study. Consider that if it was merely quantity that mattered we might conclude that longer studies, which presumably contain more observations, are more reliable or valid than shorter studies. Yet, it is laughable to assert that long books are more convincing than short books. It is the quality of the observations and how they are analyzed, not the quantity of observations, that is relevant in evaluating the truth-claims of a study based (at least in part) on non-comparable observations. Indeed, the non-comparable observations drawn upon in a given study are unlikely to be of equal importance, so merely counting them gives no indication of their overall significance in a causal argument.

Thus, the N=1 designation that we have attached to non-comparable observations should not be understood as pejorative. In some circumstances, one lonely observation (qualitative or quantitative) is sufficient to prove an inference. This is quite common, for example, when the author is attempting to reject a necessary or sufficient condition. If we are inquiring into the cause of Joe’s demise, and we know that he was shot at close range, we can eliminate suspects who were not in the general vicinity. One observation – say, a single
frame from a videotape in a surveillance camera – is sufficient to provide conclusive proof that a suspect was not, in fact, the killer, even though the evidence is neither quantitative nor comparable to other pieces of evidence.

**STANDARDIZED AND NON-STANDARDIZED RESEARCH DESIGNS**

Now, we return to our initial contrast between standardized and non-standardized research designs. It will be seen that the first involves a well-defined population, a sample resting within the population, and a set of observations (within the sample) that are treated as comparable (in ways that might be relevant to the causal proposition) and independent (thus providing independent evidence pertaining to the causal proposition). The number of observations may be large (allowing statistical analysis) or small (allowing for Mill-ean analysis). The second logic of inferential analysis is (for lack of a better term) *non-standardized*. Here, the population of the inference is limited to a single case, and the sample consists of observations that are not comparable to one another.

We do not want to give the impression that standardized research designs are always superior to approaches that employ non-comparable observations. Indeed, one of the foregoing examples, which pits Henry Brady’s non-comparable observations against the large-N statistical analysis conducted by John Lott, is a good case for the superiority of an informal method over a formal one. The reader can probably think of many others. The point is simply that *if* a viable formal research design can be constructed, it behooves the researcher to do so. Non-comparable observations are useful wherever a formal research design does not pass the laugh test or where it is insufficient, by itself, to prove an inference. Indeed, one way to think about non-comparable observations is as a cross-check, a triangulation, that can be – and ought to be – applied to all results gained through formal
methodological approaches. Studies based on a formal research design will sometimes note parenthetically that the account is consistent with “anecdotal” or “narrative” evidence, i.e., with evidence that falls outside the formal research design. It makes sense of the statements made by the actors, of their plausible motives, and so forth. This is often extremely important evidence and deserves a more respectful label than ‘anecdotal’ and a more revealing label than ‘narrative’ (what is the evidentiary status of a narrative?). To say that a method is non-standardized or informal is not to say that the evidence drawn from that method is weak or peripheral to the point at issue. It is to say only that the information cannot be (or need not be) understood as comparable observations.

A good example of non-comparable observations as an adjunct mode of causal analysis to formal research designs appears in a recent paper that examines the behavior of the U.S. Federal Reserve during the Great Depression. The central question is whether the Fed was constrained to adopt tight monetary policies because any deviation from this standard would have led to a loss of confidence in the nation’s commitment to the gold standard (i.e., an expectation of a general devaluation), and hence to a general panic (Eichengreen 1992). To test this proposition, Chang-Tai Hsieh and Christina Romer examine an incident in monetary policy during the spring of 1932, when the Federal Reserve embarked on a brief program of rapid monetary expansion. “In just fourteen weeks,” the authors note, “the Federal Reserve purchased $936 million worth of U.S. government securities, more than doubling its holdings of government debt” (Hsieh & Romer 2001: 2). To determine whether the Fed’s actions fostered investor insecurity Hsieh and Romer (2001: 4) track the forward dollar exchange rate during the spring of 1932, which is then compared to the spot rate, using “a measure of expected dollar devaluation relative to the currencies of four countries widely thought to have been firmly attached to gold during this period.”
Finding no such devaluation, they conclude that Eichengreen’s (1992) theory is false – investor confidence could not have constrained the Fed’s actions during the Great Depression.

However, this conclusion would be questionable were it not bolstered by additional evidence bearing on the likely motivations of the officials of the Federal Reserve at the time. To shed light on this matter, the authors survey the *Commercial and Financial Chronicle* (a widely read professional journal, presumably representative of the banking community) and other documentary evidence. They find that “the leaders of the Federal Reserve . . . expressed little concern about a loss of credibility. Indeed, they took gold outflows to be a sign that expansionary open market operations were needed, not as a sign of trouble” (Hsieh & Romer 2001: 2). Thus, the adjunct evidence provided by non-comparable observations are instrumental in helping the authors disconfirm a reigning theory. Moreover, this evidence also sheds light on a new theory about Fed behavior during this critical era.

Our reading of the Federal Reserve records suggests that a misguided model of the economy, together with infighting among the twelve Federal Reserve banks, accounts for the end of concerted action. The Federal Reserve stopped largely because it thought it had accomplished its goal and because it was difficult to achieve consensus among the twelve Federal Reserve banks (Hsieh & Romer 2001: 3).

This interpretation would not be possible (or at least would be highly suspect) without the adjunct evidence provided by non-comparable observations.

Granted, the non-comparable observations enlisted by Hsieh and Romer might have been converted into standardized (comparable) observations arrayed in a dataset format. For example, the authors might have conducted a content analysis of *Commercial and Financial*...
Chronicle and/or of Federal Reserve records. This would have required coding sentences (or some other linguistic unit) according to whether they registered anxiety about a loss of credibility. Here, the sentence becomes the unit of analysis and the number of sentences comprises the total N (the number of comparable observations) in a formal research design. If documents rather than sentences are the unit of analysis, then the number of comparable observations would likely be smaller, perhaps so small as to obviate statistical analysis. One would still have comparable bits of evidence, but the bits would be much larger, and less numerous.

In principle, it is always possible to convert non-comparable observations into standardized research designs, by increasing the number of observations from specific populations. Non-comparable bits of evidence can be transformed into comparable bits of evidence – i.e., standardized “observations” – simply by getting more bits of similar evidence and coding them according to type. However, it may not be possible to do so in practice, given that comparable observations may not be available (without creating counterfactuals), and that it may be difficult to code observations that are not readily comparable to one another. Moreover, there may be little advantage in doing so. In the previous example, it is not clear that anything would be gained from this sort of formalization. If there is, as the authors claim, no evidence whatsoever of credibility anxieties in the documentary evidence, then the reader is not likely to be more convinced by an elaborate counting exercise (coded as 0, 0, 0, 0, 0, …). More useful, we think, are specific examples of what leaders of the Fed actually said, as provided by the authors of this study.

Sometimes standardization is useful, and sometimes it is not. Recall that the problem of infinite causal regress forces the analyst to decide at what point she will cease investigating prior causes. The conventional answer is, whenever additional investigation
would be superfluous, i.e., whenever a formal research design would end up telling us what we already are inclined to believe (with a fair degree of certitude). Often, non-comparable observations have this character. If an argument is obvious from one observation it is redundant to collect further observations (whether comparable or not). Larger samples of comparable observations are extraneous if non-comparable observations will do the job.

This brings us to a final characteristic of non-comparable observations, as employed in causal analysis: they lean heavily on general assumptions about the world, which may be highly theoretical (nomothetic “laws”) or pre-theoretical (“common sense”). Precisely because of the paucity of evidence provided by a non-comparable observation, the researcher must assume a great deal about how the world works. The non-comparable observation makes sense only because (if) it fits snugly within a comprehensible universe of causal relations.

We do not wish to imply that non-comparable evidence is “shaky”; indeed, much of it is quite matter-of-fact and close to the ground. Our point is simply that these facts are comprehensible only when they can be ordered, categorized, “narrativized,” and this in turn rests upon a broad set of assumptions about the world. The contrast with a standardized research design, if well-constructed, is revelatory. Where there is a manipulated treatment and a control, a priori assumptions about the world are minimized. There is not much to intuit in reaching conclusions about whether \( X_1 \) causes \( Y \) (though the question of why \( X_1 \) causes \( Y \) – the question of causal mechanisms – is often more complex).

However, as one moves away from the experimental ideal, one perforce leans more heavily on background assumptions about the way the world works. Insofar as these assumptions provide “priors” against which subsequent pieces of evidence can be evaluated, the analysis of non-comparable observations takes on a Bayesian flavor.¹⁰ But the
importance of contextual knowledge about the world also extends to other features of the analysis, e.g., the identification of viable alternatives (what, under the circumstances, were the options?), the playing out of various scenarios (counterfactual logic), and so forth. The insufficiencies of the formal research design must be compensated by natural wisdom -- an intuitive “feel” of a situation, usually gained through many years’ experience in that area, be it a foreign country, a historical era, or a medical specialty.

To be sure, background knowledge informs all causal analysis. Even so, it is usually more prominently on display where some portion of the evidence derives from non-comparable observations, for each observation must be separately evaluated. In some sense, each non-comparable observation may be considered a separate research design, thus requiring a different set of background assumptions.

CONCLUSIONS

We have argued that the nub of the difference between qualitative and quantitative modes of causal analysis lies in the comparability of adjacent observations. Comparable observations provide evidence that is amenable to standardized dataset formats, and hence to statistical analysis. Even where it is employed in small-N (“Mill-ian”) analyses, the logic of analysis is quite similar. If a causal relationship is present X should vary with Y across a set of observations – observed temporally and/or spatially.

Non-comparable observations must be approached one-by-one, because each is sampled from a different population. Here, one enters a different logic of causal analysis, with certain characteristic features, as enumerated above. The key point is that these characteristic features arise from the core problem of how to put together a causal story where the individual bits of evidence are, at least to a certain extent, unique. As we have
seen, many observations (either qualitative or quantitative) are typically enlisted, each making a slightly different point, but all related to some overall argument (i.e., the primary inference). Since the observations are not comparable to one another, the presentation is delivered in prose (or “narrative analysis” [Mahoney 1999]). However, it is the absence of comparability among adjacent observations – not the use of prose or the number of observations – that makes this approach so distinctive, and so mysterious.

Non-comparable observations are the hallmark of most qualitative methods – e.g., process tracing, causal-process observations, pattern-matching, causal-chain explanation, colligation, congruence method, genetic explanation, narrative explanation, sequential explanation, and so forth. Studies of this sort do not conform to usual notions of methodological rigor because most elements of a formal research design are absent. There is, for example, no formally defined sample of observations, or populations from which each observation is sampled. Moreover, the techniques for making causal inferences that link non-comparable observations into a causal chain are often not explicitly stated. Consequently, studies that rely on non-comparable observations give the impression of being informal, ad hoc – one damn observation after another. However, we have shown that the wayward reputation of analysis based on non-comparable observations is only partially deserved. The reason is that in some circumstances it is not necessary to enlist a large number of comparable observations in order to prove a point, and in other circumstances it is impossible to do so.
BIBLIOGRAPHY


Theory and Society 20: 405-453.


Philosophy of Science. Notre Dame, IN: University of Notre Dame Press.


Mahoney and D. Rueschemeyer (eds) Comparative Historical Analysis in the Social Sciences, pp.

Howson, C. and P. Urbach. (1989) Scientific Reasoning: The Bayesian Approach. La Salle, IL:

Open Court.

Expectations in the 1932 Monetary Expansion’, NBER Working Paper No. W8113
(February).

Hopkins University Press.


University of California Press.

31-56.
*Philadelphia Inquirer* (November 14): 23A.


Mahoney, J. and D. Rueschemeyer (eds). (2003) *Comparative Historical Analysis in the Social 
Sciences*. Cambridge: Cambridge University Press.

Mahoney, J. and G. Goertz. (2006) ‘A Tale of Two Cultures: Contrasting Quantitative and 
Qualitative Research’, *Political Analysis* 14(3): 227-249.


Press.

Berkeley: University of California.


Making’, in *Advances in Information Processing in Organizations* Vol. 2, pp. 59-64. Greenwich, 
Connecticut: JAI Press Inc.

University Press.

International Studies, occasional paper no. 6.


1 Almond (1990), Gutting (1980). While the presumed unity of natural science is a bit of a
caricature, it is nonetheless true that the natural sciences are a good deal more unified than their
social science cousins.

2 We do not intend to downplay other methodological divisions within the social sciences,
e.g., between experimental and observational research designs (Freedman 1991), between frequentist
and Bayesian analysis (Howson & Urbach 1989), or among different views of causation (Brady 2002;
Winship & Sobel 2004), and so forth. Even so, the division between qualitative and quantitative
approaches seems more widespread, more longstanding, and perhaps more recalcitrant.

3 We shall assume throughout this paper that the purpose of an analysis is causal, rather than
descriptive. Descriptive statements may also be classified as qualitative (the mountain is high) or
quantitative (the mountain is 3,400 feet high). But this is not our interest here.

Note that many of the arguments vetted in these works repeat venerable themes going back a century
or more (Snow 1959/1993).

5 This characterization is based loosely on a reading of standard methodological texts across
a range of disciplines (e.g., Babbie 1983; Greene 2002; King, Keohane, & Verba 1994; Shively 1990).


7 Wilson (1992: 69). For further discussion of methodological issues in implementation
research see Goggin (1986).


9 If several cases are included in a given study, the non-comparable observations employed
in each case study are independent of each other. Indeed, several of the foregoing examples might
better be classified as single-outcome studies, as explored in the epilogue, rather than case studies. It
is not clear what the extinction of the dinosaurs, or of the 2000 Florida election, is a “case of.” But
for present purposes, this issue of classification is extraneous. Non-comparable observations work
identically in these two contexts because their employment in the context of a case study is to help
explain one particular outcome.

10 See George & Bennett (2005), Gill, Sabin & Schmid (2005), and previous discussion.