What Standards Are (or Might be) Shared?

John Gerring
Department of Political Science
Boston University

For some time, methods have been associated with quantitative methods. Thankfully, this has begun to change. Yet, there is still very little space for non-quantitative styles of research in the discipline of political science – relative, that is, to what is available for quantitative research. Witness: the contrast between "Arizona" (IQRM), with its annual contingent of 80+ students, and "Michigan" (ICPSR), with its hundreds of annual participants. Witness: the relative paucity of non-quantitative methods courses at the graduate level. Witness: the ongoing resistance on the part of established journals (e.g., *American Journal of Political Science, Political Analysis*) to the inclusion of work by qualitative methodologists. Witness: the general confusion and consternation about what constitutes a solid, methodologically defensible, qualitative study.

My first recommendations are therefore quite simple, though rather difficult to implement: greater attention should be paid to the qualitative aspects of political science methods. This should involve the expansion of IQRM/CQRM, the creation of new (required) courses in graduate programs, greater openness on the part of editors and reviewers to qualitative methods, and more explicit care to these matters on the part of researchers in the field.

Yet, all of this presumes an important precondition: that a field can be created, or is in the process of creation, which is sufficiently explicit about methodological criteria and where sufficient consensus exists such that these norms – whatever they may be – can be taught, understood, and respected. This, I take it, is the central goal of NSF's current initiative.

Where shall we look for these cross-disciplinary, cross-subfield criteria? Here, I provide a brief and necessarily schematic treatment of arguments pursued at length elsewhere (see References).

Towards Common Criteria for Social Science Work

I have argued that the work of social science is usefully divided into three inter-dependent tasks: *concept formation*, *propositions*, and *research design*. Each of these tasks responds to a somewhat different set of demands. Thus, the vast and complex subject of social science methodology may be conceptualized as a set of discrete tasks and their attendant criteria.

It must be stressed that the following criteria are understood as general goals, not as necessary conditions. They are always applicable, but not always fully achievable. Indeed, the process of conducting research usually involves *tradeoffs* among these three tasks and their attendant criteria. It should also be noted that this framework explicitly excludes research issues pertaining to practical or logistical issues – e.g., funding, time, expertise, availability of data, and so forth. Practical matters are important, to be sure; but they are not methodological issues per se.

Concepts. Concepts answer the *what?* question. In order to talk about anything at all one must call it by a name. Since some names are better than others, and some definitions better than others, we cannot escape the problem of concept formation. Adequacy in concept formation obliges one to consider eight criteria more or less simultaneously: 1) *coherence*, 2) *operationalization*, 3) *validity*, 4) *field utility*, 5) *resonance*, 6) *contextual range*, 7) *parsimony*, and 8) *analytic/empirical utility*. Juggling these criteria successfully is the art of forming good concepts.

Propositions. Propositions involve the formulation of empirical statements about the phenomenal world. (Arguments, hypotheses, explanations, and inferences are all 'propositions' in the broad sense that I employ this term.) Propositions can be classified as descriptive, predictive, or causal. Causal propositions – the most complex, methodologically speaking – are subject to the following criteria: 1) *specification* (clarification of the range of outcomes under investigation, the set of cases to which the proposition refers, the resolution of any internal contradictions, and the operationalization of all key terms), 2) *precision*, 3) *breadth* (i.e., scope, range, and generality), 4) *boundedness* (the establishment of a logical and theoretically defensible set of boundaries for the proposition; that which it covers and that which it does not), 5) *completeness* (the degree of variance explained by the proposition), 6) *parsimony*, 7) *differentiation* (is X differentiable from Y), 8) *priority* (the causal distance between X and Y), 9) *independence* (the extent to which X is exogenous relative to Y), 10) *contingency* (the identification of a causal factor that is contingent, relative to what may be considered the normal course of events), 11) *mechanism* (the causal path connecting X and Y), 12) *analytic utility* (the extent to which a proposition accords with what we know about the world, including commonsense and theoretical knowledge), 13) *intelligibility*, 14) *relevance* (societal significance), 15) *innovation* (novelty), and 16) *comparison* (is the favored X better – along these various dimensions – than other possible Xs?). A good causal argument is well-specified, precise, broad, bounded, and so forth. (Descriptive and predictive proposition can also be understood in terms of these general criteria, though not all of these sixteen dimensions apply, or they apply somewhat differently.)

Research Designs in Causal Inference. The fundamental problem of causal inference is that we cannot re-run history to see what effects X actually had on Y in a particular case. At an ontological level, this problem is unsolvable. However, we have various ways of reducing this uncertainty such that causal inference becomes possible and plausible.

There are two dimensions upon which causal effects may be observed, the temporal and the spatial. Temporal effects may be observed directly when an intervention occurs: X intervenes upon Y and we observe any change in Y that may follow. Here, the "control" is the pre-intervention state of Y; what Y was prior to the intervention (a state that we presume would remain the same, or whose trend would remain constant, in the absence of an intervention). Spatial effects may be observed directly when two phenomena are similar enough to be understood as examples (cases) of the same thing. Ideally, they are similar in all respects but one – the causal factor of interest. In this situation, the "control" is the case without the intervention.

Experimental research designs usually achieve variation through time and across space, thus maximizing leverage into the fundamental problem of causal inference. They also minimize ceteris paribus assumptions, inherent in all causal analysis. First, because the intervention is manipulated by the researcher it is unlikely to be correlated with other things that might influence the outcome of interest. Thus, any changes in Y may be interpreted as the product of X and only X, other factors being held constant. (Natural interventions are likely to be accompanied by other factors that violate the ceteris paribus assumption.) Second, treatment and control cases are identical in all respects (except the intervention itself) that might affect the causal inference in question. This is usually achieved by a randomization of treatment and control groups. (However, randomization is not viewed here as a definitional attribute of experimental research designs.) Finally, the treatment and control groups are *isolated* from each other, preventing spatial contamination. This, again, means that the ceteris paribus assumption inherent in all causal inference is relatively safe. The control may be understood as reflecting a vision of reality as it would have been without the specified intervention.

Let us now reconstruct the logic of within-case research design through the logic of the classic experiment, which I shall define – stipulatively – as characterized by a manipulated intervention (the treatment) and a suitably matched control group. This suggests three parameters upon which all research designs may be evaluated: whether there is change in the status of the key causal variable during the period under observation (an intervention); whether this intervention is manipulated or not (i.e., whether the study is experimental or observational); and whether there is a well-matched control group. The intersection of these three dimensions produces a six-fold typology (not all logically conceivable cells are relevant), which I shall label as follows: 1) the Classic Experiment, 2) the Experimental Intervention, 3) the Natural Experiment, 4) the Natural Intervention, 5) the Natural Comparison, and 6) the Counterfactual, as described in Table 1. (Note that these terms carry a more specific meaning than they do in ordinary social science discourse; hence, the importance of capitalization.)

In order to familiarize ourselves with the differences among these six paradigmatic research designs I begin with a series of scenarios built around a central (hypothetical) research question: Does the change from a first-past-the-post (FPP) electoral system to a list-proportional (list-PR) electoral system moderate inter-ethnic hostility in a polity with high levels of ethnic conflict? I shall assume that one can effectively measure inter-ethnic hostility through a series of polls administered to a random sample (or panel) of respondents at regular intervals throughout the research period. This measures the outcome of our hypothetical study, the propensity to ethnic conflict. With this set-up, how might one apply the six foregoing designs?

In its simplest form, the Classic Experiment (#1) would proceed by the selection of two communities that are similar in all respects including the employment of a majoritarian electoral system and relatively high levels of inter-ethnic hostility. The researcher would then administer an electoral system change in one of these communities, holding the other constant. The final step would be to compare the results to see if there is a difference over time between treatment and control groups.

An Experimental Intervention (#2) would follow the same procedure, but without the control group. Consequently, the researcher's judgment of results would rest solely on a before/after comparison of inter-ethnic conflict in the community that underwent change in their electoral system.

A Natural Experiment (#3) is identical to the Classic Experiment except that the researcher is now operating in a non-experimental setting. This means that she must find two communities that are similar in all respects including the employment of a majoritarian electoral system and relatively high levels of inter-ethnic hostility, one of which changes its electoral system from majoritarian to proportional. She may then compare results across the two communities.

The Natural Intervention (#4) replicates the conditions of the second research design but in a non-experimental setting. That is, the researcher observes a community with a majoritarian electoral system and high levels of inter-ethnic hostility that undergoes an electoral system change to PR, comparing results before and after the intervention.

The Natural Comparison (#5) is identical to the third research design except that in this instance there is no intervention. Here, the researcher searches for two communities similar in all respects including the employment of a majoritarian electoral system and relatively high levels of inter-ethnic hostility. One employs a majoritarian electoral system and the other a proportional electoral system. This spatial variation on the key variable forms the crux of causal inference, but is not observable through time.

In a Counterfactual research design (#6), finally, the researcher observes a community with a majoritarian electoral system and high levels of inter-ethnic hostility that does *not* undergo an electoral system change to PR. Since there is no observable change over time in the key variable of interest,

her only leverage on this question is the counterfactual: what would have happened if this country had reformed its electoral system?

The essential properties of these six research designs are illustrated in Table 1, where Y refers to the outcome of concern, X_1 marks the independent variable of interest, and X_2 represents a vector of controls (other relevant exogenous factors that might influence the relationship between X_1 and Y). These controls may be directly measured or simply assumed (as they often are in randomized experiments). The initial value of X_1 is denoted "-" and a change of status as "+." The vector of controls, by definition, remains constant. A question mark indicates that the value of the dependent variable is the major objective of the analysis. Observations are taken before (t_1) and after (t_2) an intervention and are thus equivalent to pre- and post- tests.

Interventions may be manipulated (experimental) or natural (observational), as noted in Table 1. Note also that the nature of an intervention may be sudden or slow, dramatic or miniscule, dichotomous or continuous, and the effects of that intervention may be immediate or lagged. For ease of discussion, I shall assume that the intervention is of a dichotomous nature (present/absent, high/low, on/off), but the reader should keep in mind that the actual research situation may be more variegated. Thus, I use the term intervention (aka "event" or "stimulus") in the broadest possible sense, indicating any sort of change in trend in the key independent variable, X_1. It should be underlined that the absence of an intervention does not mean that a case does not change over time; it means simply that it does not experience a change of *trend*. Any evaluation of an intervention involves an estimate of the baseline – what value a case would have had without the intervention. A "+" therefore indicates a change in this baseline trend.

Because interventions may be multiple or continuous within a single case it follows that the number of temporal observations within a given case may also be extended indefinitely. This might involve a very long period of time (e.g., centuries) or multiple observations taken over a short period of time. Observations are thus understood to occur temporally within each case (t_1, t_2, t_3, . . . t_n).

Although the number of cases in the following examples varies, and is sometimes limited to one or two, research designs may incorporate any number of cases. In the previous example, each respondent to the survey of inter-ethnic conflict is understood as a case; there is evidently no limit, a priori, to the number of respondents that might be polled. Thus, the designations "treatment" and "control" in Table 1 may be understood to refer to individual cases *or* to groups of cases. (In this paper, the terms "case" and "group" will be used interchangeably.)[50]

Finally, the classical division of an experiment into two groups – a treatment and control – may be varied. There may, indeed, be a much larger number of groups, each receiving a slightly different treatment. At the limit, the treatments may be so variegated, and so numerous, as to defy a simple division into groups. Here, the researcher may choose to model the treatment in a general format -- usually a standard mathematical algorithm, which may be linear or non-linear. In this fashion, experiments merge with statistics. (Note, once again, the softness of the boundaries.)

In numbering these research designs (#1-6) I intend to indicate a gradual "falling away" from the experimental ideal. However, it would be incorrect to assume that a higher number necessarily indicates an inferior research design. In particular, it should be underlined that my discussion focuses mostly on issues of internal validity; often, the search for greater external validity leads to the adoption of an observational research design. Evidently, the three dimensions that define this typology do not exhaust the features of a good research design (Gerring 2001; 2006: chs 4-5). However, in most social-science

50 One obvious drawback to a very small sample is that one cannot randomize the selection of treatment and control cases. If there is only one control case, or several, it makes no sense to select it randomly (Gerring 2006: ch 5). Here, the case-selection procedure should follow the *most-similar* design.

research settings, and with a strong ceteris paribus caveat – i.e., when the chosen cases are equally representative (of some population), when the interventions are the same, and when other factors that might affect the results are held constant – the researcher will usually find that this numbering system accurately reflects the preferred research design. The six-part typology is intended to simplify the field of choices, expose the full range of options, and clarify the methodological issues attached to each one. Although initially the presentation may seem a trifle abstract it is hoped that after rehearsing numerous examples these models will begin to seem second-nature. It is also hoped that they will help the reader to craft her research and explain her choices in the simplest and clearest fashion possible. To reiterate, the essential questions are a) how experimental is your research design and b) in what specific ways does it deviate from the experimental ideal?[51]

51 There is one missing ingredient in this six-part typology. It concerns situations in which relevant observations are not comparable to one another and hence cannot be arrayed in a typical (large-N or small-N) research design. These sorts of observations have been referred to as causal-process (Brady 2004) or process-tracing (Gerring and Thomas 2005) observations.

Table 1. A Comprehensive Typology of Research Designs

Hypothesis: A change from FPP to list-PR mitigates ethnic hostility.

			t_1	t_2	
EXPERIMENTAL . . .					
1. Classic Experiment	Treatment	Y	--	?	Two similar communities with FPP electoral systems and high ethnic hostility, one of which is induced to change from FPP to list-PR. Ethnic hostility is compared in both communities before and after the intervention.
		X_1	--	+	
		X_2	--	--	
	Control	Y	--	?	
		X_1	--	--	
		X_2	--	--	
			t_1	t_2	
2. Experimental Intervention	Treatment	Y	--	?	A community with a FPP electoral system and high ethnic hostility is induced to change from FPP to list-PR. Ethnic hostility is compared before and after the intervention (identical to #1 except there is no control case).
		X_1	--	+	
		X_2	--	--	
OBSERVATIONAL . . .					
			t_1	t_2	
3. Natural Experiment	Treatment	Y	--	?	Two similar communities with FPP electoral systems and high ethnic hostility, one of which changes from FPP to list-PR. Ethnic hostility is compared in both communities before and after the intervention (identical to #1 except that treatment is not manipulated).
		X_1	--	+	
		X_2	--	--	
	Control	Y	--	?	
		X_1	--	--	
		X_2	--	--	
			t_1	t_2	
4. Natural Intervention	Treatment	Y	--	?	A community with a FPP electoral system and high ethnic hostility changes to list-PR. Ethnic hostility is compared before and after the intervention (identical to #2 except the intervention is not manipulated).
		X_1	--	+	
		X_2	--	--	
			t_1	t_2	
5. Natural Comparison	Treatment	Y		?	Two similar communities, one of which has PFF and the other list-PR. Ethnic hostility is compared in both communities (identical to #3 except there is no observable intervention).
		X_1		+	
		X_2		--	
	Control	Y		?	
		X_1		--	
		X_2		--	
			t_1	t_2	
6. Counterfactual	Control	Y		?	A community with a FPP electoral system and high ethnic hostility is considered, by counterfactual thought-experiment, to undergo a change to list- PR (identical to #4 except there is no treatment case).
		X_1		--	
		X_2		--	

Cases:
Treatment = with intervention
Control = without intervention

Variables:
Y = outcome
X_1 = independent variable of interest
X_2 = a vector of controls

Observations:
t_1 = pre-test (before intervention)
t_2 = post-test (after intervention)

Cells:
| = intervention
- = stasis (no change in status of variable)
+ = change (variable changes value or trend alters)
? = the main empirical finding: Y changes (+) or does not (-)

Three Genres of Causal Analysis

I agree with naturalists such as King, Keohane, and Verba that there is – or at least ought to be – one logic of inference that unites qualitative and quantitative work. I do not want to see the development of a separate and independent "qualitative methodology," in other words. Nor do I believe that this is possible or likely so long as we retain sight of the scientific ideal. If knowledge is to be systematic, parsimonious, cumulative, and replicable, if it is to extend to causal as well as descriptive inference, and if it is to strive for generality – if all of these scientific goals are to be respected then it makes no sense to develop separate fiefdoms for qualitative and quantitative methods. Both should speak to one another. And in order to facilitative this cross-field communication we need a common logic of inference.

That said, I also agree with the critics of DSI and other naturalistically-inclined methodologists: the current mainstream view of methods is often too narrow, too constraining, defining out much of what is now regarded as sound (and scientific) practice on the qualitative side of the ledger. This oversight is not, I think, malicious. My impression is that quantitative methodologists simply do not understand what constitutes a non-mathematical approach to empirical knowledge. Nor, for that matter, do most scholars who perform qualitative work. They conduct research on an intuitive level, but without the self-conscious tools of a "methodology." Indeed, they are often openly contemptuous of any attempt to intellectualize and systematize the work of scholarship. So it is a misunderstanding that – appropriately, in view of my thesis – crosses the qualitative/quantitative boundary.

What, then, *is* the qualitative/quantitative distinction? I would argue that it is best understood as derivative of an underlying methodological issue that remains obscured in most discussions. In my view, it is all about data *comparability*. Quantitative work presumes a high level of comparability among observations (pieces of evidence); qualitative work presumes a low level of comparability. This is the principal methodological justification for doing work that is quantitative or qualitative.

Accordingly, the methodological issues faced by research designs employed in causal analysis are recognizable by the number of comparable observations that lie within each "sample." Three broad categories are distinguishable: large-N samples, small-N samples, and samples of 1. This provides the empirical foundation and methodological rationale for three well-established styles of empirical research: 1) *Mathematical*, 2) *Comparative*, and 3) *Process-tracing*.

Table 2 illustrates the defining features of these genres, most of which follow, more or less ineluctably, from differences in sample size. Since these are extraordinarily broad groupings, encompassing all disciplines in the social sciences, and since the categories themselves are internally diverse, it seems appropriate to refer to them as methodological *genres*. In any case, it should be clear that when speaking about "Mathematical methods" or "Comparative methods" we are speaking about a diverse set of approaches.[52]

52 It should be clarified, finally, that this tripartite typology refers to methods of *data analysis*, not to methods of case selection or data generation. Prior to data analysis, we assume that researchers have carefully selected cases (either randomly or purposefully), and that researchers have generated data appropriately (either by experimental manipulation or some natural process). This data may contain quasi-experimental characteristics or it may be far from the experimental ideal. Data analysis may be conducted across cases or within cases. For our purposes, these issues are extraneous, though by no means trivial. In by-passing them I do not intend to downplay them. My intention, rather, is to focus narrowly on what analysts do with data once cases have been chosen, the data has been generated, and the relevant observations have been defined. This topic, I believe, is much less well understood.

Table 2. Three Genres of Causal Analysis

	Mathematical	**Comparative**	**Process Tracing**
Individual obs:	Quantitative	Quant or Qual	Quant or Qual
Groups of obs:	Large-N sample (comparable)	Small-N sample (comparable)	Disparate N=1 observations (non-comparable)
Total number of obs:	Large	Small	Indeterminate
Presentation of obs:	Rectangular dataset	Table or prose	Prose
Analytic technique:	Statistics, Boolean algebra	Most-similar, Most-different	Processual, Counterfactual, Pattern-matching Highly deductive
Covariation:	Real	Real	Real and imagined
Stability, replicability:	High	Moderate	Low
Familiar labels:	Statistics, QCA	Comparative, Comparative-historical, Small-N cross-case study	Historical, Narrative, Ethnographic, Legal, Journalistic, Single-case study

Mathematical Methods. The Mathematical genre will be familiar to most readers because it is represented by hundreds of methods textbooks and courses. Here, the analysis is typically conducted upon a large sample of highly comparable observations contained in a standard rectangular dataset, using some mathematical algorithm to establish covariational patterns within the sample. For better or worse, this is the standard template upon which contemporary understandings of research design in the social sciences is based. For some, it appears to be the *sine qua non* of social science research (Beck 2004; Blalock 1982, 1989; Goldthorpe 2000; King, Keohane, Verba 1994; Lieberson 1985; for general discussion see Brady and Collier 2004).

Our use of the term "Mathematical" does not presuppose any particular assumptions about how this analysis is carried out. If statistical, the model may be linear or non-linear, additive or non-additive, static or dynamic, probabilistic or deterministic (i.e., employing necessary causal factors), and so forth. The only assumption that statistical models must make is that the observations are *comparable* to one another – or, if they are not, that non-comparabilities can be corrected for by the modeling procedure (e.g., by weighting techniques, selection procedures, matching cases, and so forth). For statisticians, the assumption of unit homogeneity is paramount. It should be clear that the same requirements apply whether the observations are defined spatially (a cross-sectional research design), temporally (a time-series research design), or both (a time-series cross-section research design). By extension, the same requirements apply whether the analysis is probabilistic ("statistics") or deterministic (as in some versions of Qualitative Comparative Analysis [Ragin 1987, 2000]).

As a rule, Mathematical work employs a sample that remains fairly stable throughout the course of a single study. Granted, researchers may exclude or down-weight outliers and high-leverage observations, and they may conduct sub-sample analyses. They may even interrogate different datasets in the course of a longer study, or recode the sample to conduct sensitivity analyses. However, in all these situations there is a relatively explicit and well-defined sample that contains the evidentiary basis for causal inference. The importance of this issue will become apparent as we proceed.

Comparative Methods. The two most familiar Comparative methods are *most-similar* analysis (a.k.a method of agreement) analysis and *most-different* analysis (aka method of difference), both of which can be traced back to J.S. Mill's nineteenth-century classic, *System of Logic* (1834/1872). In most-similar analysis, cases are chosen so as to be similar on all irrelevant dimensions and dissimilar on both the hypothesized causal factor and the outcome of interest. In most-different analysis, cases are chosen to maximize difference among the cases on all causal factors (except one), while maintaining similarity on the outcome. The most-similar research design is more common, and probably more well-grounded, than the most-different research design (Gerring 2006: ch 5; Seawright and Gerring 2005).

The details of these research designs are not important. What is important is that the cross-case component of the analysis be fairly explicit. There must be a recognizable sample within which the chosen cases are analyzed. In other words, there must be significant cross-case variation and this variation must comprise an important element of the overall analysis. This is the "comparative method" as it has become known within the subfield of comparative politics (Collier 1993).[53] "Comparative-historical" work is similar to the foregoing except that the analysis also incorporates a significant over-time component (Mahoney and Rueschemeyer 2003).

53 My use of the term Comparative includes what Mahoney (1999) labels "nominal comparison" and "ordinal comparison," but not what he labels "narrative analysis," which I incorporate under Process Tracing below.

Cases are thus examined spatially and temporally, and the temporal analysis usually includes a change in one or more of the key variables, thus introducing an intervention ("treatment") into the analysis.[54]

Comparative methods, like Mathematical methods, are based upon a relatively stable sample of comparable cases. Granted, there are likely to be some shifts in focus over the course of a longer study. Sometimes, a researcher will choose to focus on a series of nested sub-samples, e.g., paired comparisons (Collier and Collier 1991). The small size of the sample means that any change in the chosen cases will have a substantial impact on the sample, and perhaps on the findings of the study. Ceteris paribus, small samples are less stable than large samples.

Because Comparative methods must employ cases that are fairly comparable to one another, they may be represented in a standard, rectangular dataset where the various dimensions of each case are represented by discrete variables. Yet, because there are relatively few cases (by definition), it is rare to see a dataset presentation of the evidence. Instead, scholars typically rely on small tables, 2x2 matrices, simple diagrams, or prose.

The most important difference between Mathematical methods and Comparative methods is that the latter employs small samples that may be analyzed without the assistance of interval scales and formal mathematical models. This does not *preclude* the use of mathematical models (e.g., Houser and Freeman 2001), or of algorithms to assign precise weightings to "synthetic" cases (Abadie and Gardeazabal 2003). However, it is not the usual mode of procedure. Indeed, statistics are relatively powerless when faced with samples of a dozen or less. A simple bivariate analysis may be conducted, but this does not go much further than what could be observed visually in a table or a scatterplot diagram.

Another difference with the Mathematical framework is that Comparative methods presuppose a fairly simple coding of variables, usually in a dichotomous manner. Similarities and differences across cases must be clear and distinct, otherwise they cannot be interpreted (due to the small-N problem). Thus, continuous variables are usually dichotomized into high/low, present/absent, strong/weak, and so forth. Simple coding schemes, and the absence of probability distributions, impose a deterministic logic on Comparative methods, such that causal factors (or combinations of factors) must be understood as necessary, sufficient, or necessary and sufficient. Deterministic assumptions may also be employed in Mathematical methods, particularly Boolean methods, but they are not de rigueur in statistical methods. Moreover, the smaller the sample size, the more difficult it is to incorporate continuous causal factors and probabilistic logic if firm conclusions are to be reached.

Process Tracing Methods. Process Tracing, in our lexicon, refers to any method in which the researcher analyzes a series of noncomparable observations occurring within a single case.[55] Studies that employ Process Tracing typically consist of many observations (either qualitative or quantitative), each making a slightly different point, but all related to some overall argument (i.e., the primary inference).

54 My discussion thus far has approached Comparative methods according to the primary unit of analysis, usually referred to as a "case" (a spatially and temporally delimited unit that lies at the same level of analysis as the principal inference). To be sure, this genre of work may also exploit *within-case* variation, which might be large-N (e.g., a mass survey of individual respondents or a time-series analysis of some process), small-N (e.g., a comparison among a half-dozen regional units), or a series of N=1 observations (e.g., a study of a particular decision or set of decisions within the executive). In short, the within-case components of Comparative methods are indeterminate; they may be Mathematical, Comparative, or Process Tracing. The fact that a single study may employ more than one method is not disturbing; as we observed, a change in an author's level of analysis *often* corresponds to a change in research design. In short, the same tripartite typology is applicable at any single level of analysis; but it does not always apply across-the-board to all levels of analysis in a given study.

55 The term "process tracing" is ambiguous, having been appropriated for a variety of uses. For some writers, it refers to any investigation (qualitative or quantitative) into causal mechanisms (George and Bennett 2005). There is, to be sure, a strong affinity between this technique, as we describe it, and a researcher's insight into causal paths. However, it may be a mistake to *define* process tracing as the search for causal mechanisms. After all, this is also an objective of Mathematical and Comparative studies. In short, while Process-Tracing methods give more attention to causal mechanisms, this should not be considered a defining feature. Our definition of process tracing might also be labeled "causal-process" observations (following Brady and Collier 2004), or alternatively, colligation, narrative explanation, pattern-matching, sequential explanation, genetic explanation, and causal chain explanation. For general discussion, see Brady (2004), George and Bennett (2005: ch 8), Little (1995: 43-4), Scriven (1976), Seawright and Collier (2004), Tarrow (1995: 472). For examples, see Geddes (2003: ch 2), George and Smoke (1974), Goldstone (2003: 50-1), George and Bennett (2005: appendix).

Since the observations are not comparable with one another, the presentation is delivered in prose – or what Mahoney (1999) labels "narrative analysis." However, it is the absence of comparability among adjacent observations – not the use of prose (or narrative) – that makes this approach so distinctive, and so mysterious. Process-Tracing methods do not conform to standard notions of methodological rigor because most elements of a "research design," in the usual sense of the term, are absent. There is, for example, no formally defined sample of observations, as with Mathematical and Comparative methods. Moreover, the methods for making causal inferences that link observations into a causal chain are often not explicitly stated. Consequently, Process-Tracing studies give the impression of being informal, ad hoc -- one damn observation after another.

The skepticism of mainstream methodologists is not difficult to comprehend. William Riker (1985: 62-3; see also Beck 2004) regards process-tracing as "scientifically impossible." Tracing a process, and imposing a pattern is, of course, no more and no less than writing history. Although some nineteenth-century historians claimed to be scientific, such a claim has seldom been put forward in this century until now, when it rises up, camouflaged, in social science. There was good reason for abandoning the claim: Historical explanation is genetic. It interprets cause as no more than temporal sequence, which, in the philosophy of science, is precisely what has long been denounced as inadequate. Causality in science is a necessary and sufficient condition; and, although temporal sequence is one of several necessary conditions, it is not sufficient. . . Process-tracing of the history of an event, even the comparison of several traced processes, does not give one generalizations or theory. However, we shall argue that the wayward reputation of Process Tracing is only partially deserved. Indeed, inferences drawn from Process-Tracing methods may be more secure, at least in some instances, than inferences based on Mathematical or Comparative methods. Thus, there are strong arguments for the employment of non-comparable (N=1) observations in social science.

We begin with an extended example drawn from Henry Brady's (2004: 269-70) reflections on his study (in tandem with a team of methodologists) of the Florida election results in the 2000 presidential election. In the wake of this close election at least one commentator 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 after the election) (Lott ??). In order to address the question, Brady stitches together isolated pieces of evidence in an inferential chain. 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 1/72nd (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 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 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 the sort of detective work that fuels the typical Process-Tracing study, and it is not a sort that can be represented in a rectangular dataset. The reason is that the myriad pieces of evidence are not comparable to each other. They all support the central argument – they are not "random" – but they do not comprise observations in a larger sample. They are more correctly understood as a series of N=1 (one-shot) observations – or perhaps the more ambiguous phrase "pieces of evidence" is appropriate. In any case, Brady's observation about the timing of the call – ten minutes before the closing of the poll – is followed by a second piece of evidence, the total number of people who voted on that day, and a third and a fourth. It would be impossible to string these together into a large, or even moderately-sized, sample, because each element is disparate. Being disparate, they cannot be counted. While the analytic procedure seems messy, we are convinced by its conclusions – more convinced, indeed, than by the large-N analysis that Brady is arguing against (in which . . .). Thus, it seems reasonable to suppose that, in some circumstances at least, Process Tracing is more scientific than sample-based inferences, even though its method is difficult to describe.

This is the conundrum of Process-Tracing research. We are often convinced by the results, but we cannot explain – at least not in any generalizable, formal fashion – why. Our confidence appears to rest on highly specific propositions and highly specific observations. There is little we can say, in general, about "Brady's research design" or other Process-Tracing research designs. It is no surprise that Process Tracing receives little or no attention from traditional methods texts, structured as they are around the quantitative template (e.g., King, Keohane, and Verba 1994). These methods texts do not tell us why a great deal of research in the social sciences, including a good deal of case study research, succeeds or fails.

While sample-based methods (both Comparative and Mathematical) can be understood according to their covariational properties, Process-Tracing methods invoke a more complex logic, one that is analogous to detective work, legal briefs, journalism, traditional historical accounts, and single-case studies. The analyst seeks to make sense of a congeries of disparate evidence, some of which may explain 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? Process-Tracing methods are, by definition, case-based. If a researcher begins to draw comparisons with other assassinations or other wars, then she is using (at least implicitly) a Comparative method, which means that all the standards of rigor for Comparative methods pertain and the researcher is entering a different methodological context.

It is important to note that the observations enlisted in a Process-Tracing case study may be either qualitative or quantitative. Brady employs a good deal of quantitative evidence. However, because each quantitative observation is quite different from the others they do not collectively constitute a sample. Each observation is sampled from a different population. This means that each quantitative observation is qualitatively different. Again, it is the comparability of adjacent observations, and the number of those observations, not the nature of the observations, that define a study as Mathematical, Comparative, or Process Tracing.

Note also that because each observation is qualitatively different from the next, the entire set of observations in a Process-Tracing study is indeterminate and unstable. The "sample" (we use this term advisedly) shifts from observation to observation. Because of this, we refer to samples of 1, or N=1 observations (of which there may be many in a single case study). A careful reader might object that the notion of an "observation" implies the existence of other comparable observations in a larger population. We accept that this is true for most observations. The issue is not whether comparable observations exist, but rather whether those other observations are considered (i.e., sampled and analyzed) in the case study. If they are not considered, then we have a set of N=1 observations. Regardless of how carefully

one seeks to define these things, there should be no disagreement on our basic point that samples, populations, and sampling techniques are not well specified in Process-Tracing methods. If they are well specified, then we are working in the realm of Comparative or Mathematical methods.

There may be *many* non-comparable observations in a single Process-Tracing study, so the cumulative number of observations could 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. Recall, from our previous discussion, that the act of counting presumes comparability among the things being counted. Process-Tracing evidence lacks this quality; this is why it is resistant to the N question. In an effort to count, one may of course resort to lists of what appear to be distinct pieces of evidence. This approximates the numbering systems commonly employed in legal briefs. But lists can always be composed in multiple ways, 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 studies such as Fenno's *Homestyle* (1978), Kaufman's *The Forest Ranger* (1960), Geertz's *Negara* (1980), and Pressman and Wildavsky's *Implementation* (1973). Process-Tracing observations are not different examples of the same thing; they are, instead, *different things*. Consequently, it is not clear where one observation ends and another begins. They flow seamlessly together. Thus, we cannot re-read Fenno, Kaufman, Geertz, or Pressman and Wildavsky with the aid of a calculator and hope to discover their true N, nor would we gain much – if 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; 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 analysis. Indeed, the observations themselves remain undefined.

This ambiguity is not in our opinion troublesome, for the number of observations in a Process-Tracing study does not bear directly on the usefulness or truthfulness of that study. While the number of observations in a sample drawn from a well-defined population contains information directly relevant to any inferences that might be drawn from that sample, the number of observations in a Process-Tracing study (assuming one could estimate their number) has no obvious relevance to inferences that might be drawn from that study. Consider that if it was merely quantity that mattered we might safely 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 quite evidently 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 Process-Tracing study.

Thus, the N=1 designation that we have attached to Process-Tracing evidence 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 – "I saw Peter at the supermarket" – is sufficient to provide fairly conclusive proof (provided, of course, that the witness is reliable). Better yet would be a videotape of the suspect at the supermarket from a surveillance camera. This would be conclusive evidence to falsify a hypothesis (in this case, Peter shot Joe), even though it is not quantitative or comparable evidence.

Process-Tracing methods apply only to situations in which the researcher is attempting to reconstruct a sequence of events occurring within a single case – i.e., a relatively bounded unit such

as a nation, family, legislature, or decision-making unit. That case may be quite broad, and might even encompass the whole world, but it must be understood as a single unit, for purposes of the analysis. All Process-Tracing methods are inherently within-case analysis. If several cases are analyzed, the researcher has either switched to a different style of analysis or adopted an additional style of analysis, one in which there is a specifiable sample (either large-N or small-N). The researcher may, for example, have begun with a Process-Tracing analysis within one case study, and later switched levels of analysis by comparing that case study with other case studies using a Comparative method.

What is it, then, that makes a Process-Tracing study convincing or unconvincing? What are the methods within this genre of causal analysis? A fundamentally puzzling aspect of the Process-Tracing method is that it rests, at times, on extremely proximate evidence (observations lying close to the "scene of the crime"), and at other times on extremely general assumptions about the theory at hand or the way the world works. Process Tracing thus lies at both extremes of the inductive-deductive spectrum. Sample-based studies, by contrast, generally require fewer deductive assumptions and, at the same time, are more removed from the facts of the case. The extreme quality of Process Tracing – which bounces back and forth from Big Theory to detailed observation – contributes to its "unstable" reputation. However, there are good reasons for this back-and-forth.

Broadly, we may distinguish among two styles of Process-Tracing research; one is *exploratory* and the other *confirmatory* (Gerring 2001: ch ?). In an exploratory mode, the researcher seeks to discover what went on in a specific context without any strong theoretical preconceptions. The question "What happened?" is asked in an opened-ended fashion. While this may seem removed from the deductive mode of inquiry that we have described, in fact it relies heavily on an understanding (theoretical or pre-theoretical) of the way the world works. In order to demonstrate a causal relationship from the mass of evidence at hand it is necessary to provide a reconstruction of the event under slightly different (imaginary) circumstances. One must construct valid "what if?" scenarios. The method of Process Tracing is thus linked to what has come to be known as the counterfactual thought-experiment (cites). There is simply no other way that the tracing of a single process through time can make causal claims – since, by definition, there are no "real" (actually existing) contrasting cases. Note that if there are other cases, and if these cases are brought into the analysis, then the researcher has transitioned into either a Mathematical or Comparative mode of analysis (depending upon the number of comparison-cases she is considering and her mode of examination). Process Tracing is limited to a single thread of occurrences. To be sure, the fact that these occurrences can be interpreted at all is courtesy of the analyst's general assumptions about how the world works (or how this particular part of the world works). This is why general knowledge – even if it is not specific to a particular theory – counts heavily in all Process-Tracing studies. The conjunction of general and specific knowledge is nicely brought out in Clayton Roberts's (1996: 66) description of process tracing as "the minute tracing of the explanatory narrative to the point where the events to be explained are microscopic and the covering laws correspondingly more certain." While we hesitate to invoke the rather controversial notion of a covering law, we hold, with Roberts, that Process Tracing conjoins highly specific and highly general observations.

Confirmatory Process Tracing also relies on imaginary counterfactuals, and also combines the general and the specific. The difference is that here a theory, rather than one's general knowledge of the world, is instrumental in identifying relevant factuals and counterfactuals. This style of Process Tracing sometimes goes under the label of "pattern-matching." Here, a theory "generates predictions or expectations on dozens of other aspects of the [subject at hand], and [the writer] does not retain the theory unless most of these are also confirmed. In some sense, he has tested the theory with degrees of freedom coming from the multiple implications of any one theory" (Campbell 1975/1988: 380; see

also Scriven 1976). An exploratory study asks "What happened?" A pattern-matching investigation inquires, first, "What should have happened if Theory X is true?" and, second, "Did that predicted course of action actually occur?" To be sure, in practice researchers often blend these two closely related techniques. A researcher may start inductively, but find herself with several weak links in the causal chain. To bolster these links, she might turn to pattern-matching, using hypotheses drawn from theories (i.e., covering laws) to make the causal inferences for those links.

References

NB: The foregoing discussion draws from the following works, as well as from work by other scholars (e.g., Andrew Bennett, Henry Brady, David Collier, Colin Elman, Jim Mahoney) and discussions with many friends and associates.

Abadie, Alberto and Javier Gardeazabal. (2003). "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* (March): 113-32.

Beck, Nathaniel. (2004). "Is Causal-Process Observation an Oxymoron? A Comment on Brady and Collier (eds.), 'Rethinking Social Inquiry.'" *Ms.*

Blalock, Hubert M., Jr. (1982). *Conceptualization and Measurement in the Social Sciences.* Beverly Hills: Sage.

Blalock, Hubert M, Jr. (1989). "The Real and Unrealized Contributions of Quantitative Sociology." *American Sociological Review,* 54: 447-60.

Brady, Henry E. (2004). "Data-Set Observations versus Causal-Process Observations: The 2000 U.S. Presidential Election." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards,* edited by H. E. Brady and D. Collier. Lanham: Rowman & Littlefield.

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

Campbell, Donald T. (1975/1988). "'Degrees of Freedom' and the Case Study." In *Methodology and Epistemology for Social Science*, edited by E. Samuel Overman. Chicago: University of Chicago Press.

Collier, David. (1993). "The Comparative Method." In *Political Science: The State of the Discipline II*, edited by A.W. Finifter. Washington, DC: American Political Science Association.

Collier, Ruth B. and David Collier. (1991). *Shaping the Political Arena: Critical Junctures, the Labor Movement, and Regime Dynamics in Latin America*. Princeton: Princeton University Press.

Fenno, Richard F., Jr. (1978). *Home Style: House Members in their Districts.* Boston: Little, Brown.

Geddes, Barbara. (2003). *Paradigms and Sand Castles: Theory Building and Research Design in Comparative Politics*. Ann Arbor: University of Michigan Press.

Geertz, Clifford. (1980). *Negara: The Theatre State in Nineteenth-Century Bali*. Princeton: Princeton University Press.

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

George, Alexander L. and Richard Smoke. (1974). *Deterrence in American Foreign Policy: Theory and Practice*. New York: Columbia University Press.

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

Gerring, John. (2004). "What is a Case Study and What is it Good For?" *American Political Science Review,* 98: 341-54.

Gerring, John. (2005). "Causation: A Unified Framework for the Social Sciences." *Journal of Theoretical Politics* 17:2 (April).

Gerring, John. (2006). *Case Study Research: Principles and Practices.* Cambridge: Cambridge University Press (forthcoming).

Gerring, John and Craig Thomas. (2005). "What is 'Qualitative' Evidence?: When Counting Doesn't Add Up." *In process.*

Gerring, John and Jason Seawright. (2005). "Selecting Cases in Case Study Research: A Menu of Options." *In process.*

Gerring, John and Paul A. Barresi. (2003). "Putting Ordinary Language to Work: A Min-Max Strategy of Concept Formation in the Social Sciences." *Journal of Theoretical Politics,* 15: 201-32.

Gerring, John and Rose McDermott. (2005). "Experiments and Observations: Towards a Unified Framework of Research Design." *In process*.

Goldstone, Jack A. (2003). "Comparative Historical Analysis and Knowledge Accumulation in the Study of Revolutions." In *Comparative Historical Analysis in the Social Sciences*, edited by J. Mahoney and D, Rueschemeyer. Cambridge: Cambridge University Press.

Goldthorpe, John H. (2000). *On Sociology: Numbers, Narratives, and the Integration of Research and Theory.* Oxford: Oxford University Press.

Houser, Daniel and John Freeman. (2001). "Economic Consequences of Political Approval Management in Comparative Perspective." *Ms.*.

Kaufman, Herbert. (1960). *The Forest Ranger: A Study in Administrative Behavior*. Baltimore: Johns Hopkins University Press.

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

Lieberson, Stanley. (1985). *Making it Count: The Improvement of Social Research and Theory*. Berkeley: University of California Press.

Little, Daniel. (1995). "Causal Explanation in the Social Sciences." *Southern Journal of Philosophy,* 34: 31-56.

Mahoney, James. (1999). "Nominal, Ordinal, and Narrative Appraisal in Macro-Causal Analysis." *American Journal of Sociology,* 104: 1154-96.

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

Mill, John Stuart. (1843/1872). *System of Logic, 8th ed*. London: Longmans, Green.

Pressman, Jeffrey L. and Aaron Wildavsky. (1973). *Implementation*. Berkeley: University of California Press.

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

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

Riker, William H. (1985). "Comments on 'Case Studies and Theories of Organizational Decision Making." *Advances in Information Processing in Organizations*, 2: 59-64.

Roberts, Clayton. (1996). *The Logic of Historical Explanation*. University Park: Pennsylvania State University Press.

Seawright, Jason and David Collier. (2004). "Glossary." Pp. 273-313 in *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, edited H. E. Brady and D. Collier. Lanham: Rowman & Littlefield.

Scriven, Michael. (1976). "Maximizing the Power of Causal Investigations: The Modus Operandi Method." Pp. 101-18 in *Evaluation Studies Review Annual*, edited by G.V. Glass. Beverly Hills: Sage.

Tarrow, Sidney. (1995). "Bridging the Quantitative-Qualitative Divide in Political Science." *American Political Science Review,* 89: 471-74