

Memo for Workshop on Interdisciplinary Standards for Systematic Qualitative Research

May 19-20, 2005
National Science Foundation
Draft: 5/09/05

-- John Gerring --

For some time, methods have been associated with quantitative methods. Thankfully, this has begun to change. Yet, there is still very little training for non-quantitative styles of research – 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 resistance on the part of established methods journals (e.g., *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 social science methods. This should involve the expansion of IQRM/CQRM, the creation of new 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).

I. 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.

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 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* (aka scope, range, 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 design refers to the method used to prove or demonstrate an empirical argument. Again, I shall focus on causal propositions, with the understanding that the following criteria apply, with modifications, to research that is descriptive or predictive in nature. Research designs aim for: 1) *plentitude* (multiple instances of a phenomenon, understood as cases or observations [“N”]), 2) *comparability* (equivalence or unit homogeneity among chosen observations), 3) *independence* (chosen observations must be regarded as independent instances or examples of whatever phenomenon is of interest; the problem of autocorrelation or contamination), 4) *representativeness* (the degree to which the chosen observations may be said to represent a broader population; external validity), 5) *variation* (the chosen observations should provide variation along key dimensions, though not along other dimensions), 6) *replicability* (the research design should be replicable), 7) *process-tracing* (the chosen observations should provide evidence of the causal pathway connecting X and Y), and 8) *comparison* (the research design should allow the researcher to test alternate causal hypotheses).

It must be stressed that the foregoing criteria -- applying to concepts, propositions, and research designs -- 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.

II. An Experimental Framework

Since research design is, arguably, the most important element of qualitative methods I want to present a more detailed treatment of this issue (more detailed, that is, than the eight criteria set forth in the previous section). It is helpful, I think, to view the question of research design through the logic of the classic experiment, 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): 1) the experiment with control, 2) the experiment without control, 3) the observational study with control and intervention, 4) the observational study with intervention but without control, 5) the observational study with control and no intervention, and 6) the observational study without control or intervention, as illustrated in Table 1.

For heuristic purposes, let us suppose that the main research question concerns whether the change from a first-past-the-post (FPP) electoral system to a list-proportional (list-PR) electoral system moderates inter-ethnic hostility in a polity with high existing levels of ethnic conflict. Let us also assume that we 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 study, the propensity to ethnic conflict. With this set-up, how might one apply the six foregoing designs?

In an experimental research design with control (#1) the researcher might choose two communities that are similar in all respects including their employment of a majoritarian electoral system and their relatively high levels of inter-ethnic hostility. She would then administer an electoral system change in one of these communities, holding the other constant. The final step is to compare the results to see if there is a difference over time between treatment and control groups.

In an experimental research design without control (#2) the researcher follows the same procedure, but without the control group. Consequently, her judgment of results rests solely on a before/after comparison of inter-ethnic conflict.

The observational study with control (#3) is identical to the first research design 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 observational study without control (#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 observational study with control and no intervention (#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.

Finally, in an observational study without control or intervention (#6) 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 measured or simply assumed. 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. 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, we shall assume that the intervention is of a dichotomous nature (present/absent, high/low, on /off), but one 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 “+” thus 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. An experiment, for example, might involve 1,000 treatments of the same case. Observations are thus understood to occur temporally within each case ($t_1, t_2, t_3, \dots t_n$).

Although the number of cases in these examples is limited to one or two, research designs may incorporate any number of cases. Each case listed in Table 1 may therefore be understood to represent a treatment or control *group*, or a range of continuous variation across multiple cases. (The terms “case” and “group” are used interchangeably in this discussion.)

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 experimental settings a control group is sometimes redundant, in which circumstance research design #1 is no better than research design #2. Sometimes, the existence of an observable intervention is more important than the existence of a real control, in which case research design #4 may be preferable to research design #5. Evidently, the four features that define this typology do not exhaust the features of a good research design. However, in most social-science 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 hierarchy accurately reflects the desirability of various possible research designs. 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. 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?

Table 1:
A Comprehensive Typology of Research Designs

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

EXPERIMENTAL . . .		t_1	t_2	
1. w/ control	Treatment	Y: - ?	X ₁ : - +	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.
	Control	Y: - ?	X ₁ : - -	
		t_1	t_2	
2. no control	Treatment	Y: - ?	X ₁ : - +	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 ₂ : - -		
OBSERVATIONAL . . .		t_1	t_2	
3. w/ control	Treatment	Y: - ?	X ₁ : - +	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.)
	Control	Y: - ?	X ₁ : - -	
4. no control	Treatment	Y: - ?	X ₁ : - +	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 ₂ : - -		
5. w/ control, no intervention	Treatment	Y: ?	X ₁ : +	Two similar communities, one of which has FPP and the other list-PR. Ethnic hostility is compared in both communities. (Identical to #3 except there is no observable intervention.)
	Control	Y: ?	X ₁ : -	
6. no control or intervention	Control	Y: ?	X ₁ : -	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 ₂ : -		

Cases:

Treatment = with intervention
Control = without intervention

Variables:

Y = outcome
X₁ = independent variable of interest
X₂ = 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 (-)

III. Qualitative and Quantitative

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 “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 reach for generality (breadth) – 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 facilitate 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 selfconscious 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.¹

¹ 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
<i>Primary evidence:</i>	Large-N sample	Small-N sample	Disparate N=1 observations
<i>Adjacent obs:</i>	Comparable	Comparable	Non-comparable
<i>Total number of obs:</i>	Large	Small	Indeterminate
<i>Measurement (along relevant dimensions):</i>	Necessary	Unnecessary	Contingent
<i>Presentation:</i>	Rectangular dataset with prose	Table or figure with prose	Prose
<i>Analytic technique:</i>	“Quantitative”: Statistics, Boolean algebra	“Qualitative”: Most-similar, Most-different	“Qualitative”: Processual, Deductive
<i>Covariation:</i>	Real	Real	Real or imagined
<i>Stability, replicability:</i>	High	Moderate	Low
<i>Familiar labels:</i>	Statistics, QCA	Comparative, Comparative-historical, Small-N cross-case study	Historical, Narrative, Ethnographic, Legal, Journalistic, Single-case study

Mathematics. The Mathematical genre will be familiar to most readers because it is represented by hundreds of methods textbooks and courses. Here, the analysis is conducted upon a large sample of 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.

However, my 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 such non-comparabilities can be corrected for by the modeling procedure (e.g., by weighting, 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)

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 I proceed.

Comparative Methods. The two most familiar Comparative methods are *most-similar* analysis and *most-different* analysis, both of which can be traced back to J.S. Mill’s nineteenth-century classic, *System of Logic* (first published in 1834). 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. There are many varieties of small-N cross-case comparisons, and they go by many names – e.g., “extreme,” “deviant,” “influential,” “crucial,” “diverse,” and so forth. For our purposes, it is important that the cross-case component of the analysis be fairly explicit; there must be a recognizable sample (or population) which the chosen cases are analyzed against. In other words, there must be significant cross-case variation and this variation must comprise an important element of the overall analysis. Comparative-historical work is similar to the foregoing except that the analysis also incorporates a significant over-time component. Cases are thus examined spatially and temporally (in this respect, comparative-historical analysis mirrors time-series cross-sectional research designs). The temporal analysis usually includes a change in one or more of the key variables, thus introducing an intervention (or treatment) into the analysis. I appropriate the general term Comparative to refer to small-N case comparisons with or without a temporal component.

Comparative methods, like Mathematical methods, are based upon a relatively stable sample of comparable cases, if not observations within those 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. Even so, it is usually possible to identify the larger sample and the smaller sub-samples and they usually remain constant over the course of the investigation.

Because Comparative methods employ cases that are highly comparable to one another these cases may be represented in a standard, rectangular dataset where the various dimensions of each case are represented by a variable. Yet, because there are relatively few observations (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. Thus, the most important difference between Mathematical methods and Comparative methods is that the latter employs small samples and may therefore be analyzed without the assistance of formal mathematical models.

Process-tracing. Process-tracing refers here to any method in which the researcher analyzes a series of noncomparable observations occurring within a single case. Process-tracing studies typically consist of many (qualitative and quantitative) observations, each making a slightly different point, but presumably – if well put-together – related to some overall argument, the primary inference. Since the evidence is not comparable, the presentation is delivered in prose. However, it is the absence of comparability among adjacent observations – not use of prose -- that makes this approach so distinctive, and so mysterious. Process-tracing methods do not conform to standard notions of methodological rigor. They have no “research design,” in the usual sense of the term. There is no formally defined sample, and the population of the inference may also be somewhat hazy. Consequently, Process-tracing studies give the impression of being informal, ad hoc – one [!@#%\$] observation after another. This reputation is only partially deserved, in our opinion. It is true, one must acknowledge, but it is not necessarily remediable. And it does *not* mean that inferences drawn from Process-tracing evidence are necessarily less secure than inferences based on Mathematical or Comparative analysis. They may, or they may not be. In short, there is sometimes a strong argument for the employment of non-comparable (N=1) observations in social science. This section will show why, and to this extent may be read as a defense of Process-tracing methods (though that is not our primary purpose).

Because Process-tracing methods are not well-understood – indeed, there is considerable disagreement over how to define the term – I begin with an extend example. This example is drawn from Henry Brady’s reflections on his own study (in tandem with a team of methodologists) of the Florida election results in the 2000 presidential election.² In the wake of this close election a number of commentators 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). 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 [eligible voters in the panhandle – 379,000] 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

² Henry E. Brady, “Data-Set Observations versus Causal-Process Observations: The 2000 U.S. Presidential Election.” In Henry E. Brady and David Collier (eds), *Rethinking Social Inquiry: Diverse Tools, Shared Standards* (Lanham: Rowman & Littlefield, 2004) 269-70.

“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 “mushy.”

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 pieces of evidence. 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. 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 (either 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 to single-case studies (but not, as I have noted, to small- N cross-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 has entered a different methodological context. [Hm...]

It is important to note that the individual pieces of evidence enlisted in a Process-tracing case study may be either qualitative or quantitative. Brady, for example, employs a good deal of quantitative evidence. However, because each quantitative observation is quite different to the others, it does not constitute, collectively, a “sample.” Each quantitative observation is qualitatively different. The ambiguity encountered here is rooted in our ambiguous use of the terms “quantitative” and “qualitative,” which may refer to an individual observation (contrast “The population of Ghana is approximately 21 million” with “The population of Ghana is moderately large”) or to a mode of analysis (contrast a survey research design intended to measure the population of Ghana with a speculative approach to this question). Thus, when I classify Process-tracing as a qualitative technique in Table 1 I am referring to the mode of analysis, not the actual bits of evidence employed. Again, the reader will note that 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. The point is driven home when one considers that Mathematical analysis is often conducted on the basis of observations that are, individually, qualitative – e.g., qualitative assessments of a color, a situation, or degrees of corruption (to take three quite different examples). Words can be used in Mathematical analyses, just as numbers can be employed in Process-tracing analyses. The key issue, for present purposes, is whether the words or numbers are comparable to other words and numbers employed in the same study.

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” (I use this term advisedly) shifts from observation to observation. Because of this, I refer to samples of 1, or $N=1$ observations. A careful reader might object that the notion of an “observation” implies the

existence of other comparable observations, in which case there can be no such thing as an isolated observation. Even so, I find no better way to express the empirical components of Process-tracing research. Regardless of what one chooses to call it, there will be no disagreement on the basic point: samples, populations, and sampling techniques are not well specified. They are shifting sands.

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 will be difficult to say exactly how many. Non-comparable observations are, by definition, difficult to count. Recall, from our previous discussion, that the act of counting presumes comparability among the things that are 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 Pressman and Wildavsky's *Implementation*, Kaufman's *The Forest Ranger*, and Geertz's *Negara*. 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, Geertz, Kaufman, or Pressman and Wildavsky with the aid of a calculator and hope to discover their true N. 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 insist that they have a many “rich” observations at their disposal, though they are unable to say, precisely, how many they are or where they are. Indeed, they remain undefined.

This ambiguity is not, in my opinion, troublesome, for the N of a Process-tracing-method study does not bear directly on the usefulness or truthfulness of that study. While the N of a sample contains information directly relevant to any inferences that might be drawn from that sample, the N of a set of singular observations (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 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 I 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 video tape 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. If several cases are analyzed, the researcher has switched to a different style of analysis, one in which there is a specifiable sample (either large- or small-N).

What is it, then, that makes a Process-tracing study convincing? (Evidently, not all such studies are.) What is the “method” of the Process-tracing method? A fundamentally puzzling aspect of the Process-tracing method is that it rests, at times, on extremely proximate evidence (evidence 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.

Conclusions. Evidently, there is a certain degree of genre-crossing in all social science work. Studies may employ Mathematical, Comparative, and Process-tracing approaches. Indeed, when studies move across levels of analysis it is common for authors to adopt different approaches; different methods may be appropriate, for example, when analyzing cross-case and within-case evidence. However, at any given point in the narrative the author must employ one of these three methods. In this respect, the classification of Table 2 is mutually exclusive and exhaustive. Moreover, most studies can be readily classified according to the *predominant* mode of analysis that the author employs. (This usually corresponds to the principal unit of analysis.) It is only a small exaggeration of reality to suggest that most studies in the social sciences today fit one of these three types (even if they are not *exclusively* in one box).

In certain respects, this tripartite typology respects the traditional distinction between quantitative and qualitative work. As we have seen, the Mathematical genre is more or less equivalent to the usual meaning of the term “quantitative,” and the other two categories are fairly classified as “qualitative.” What this conventional dichotomy misses is the important distinction between work based on small samples (Comparative) and work based on N=1 samples (Process-tracing). The first is usually associated with studies that employ a variant of Mill’s most-similar or most-different research design; the second is associated with studies that provide a narrative-based analysis of a single case. This distinction, I believe, is much more portentous than the more commonly cited qualitative/quantitative distinction. Arguably, the most important division in social science work today is not between quant and qual, but rather between work that relies on samples comprised of comparable observations (which may be large- or small-N) and work based on evidence drawn from a concatenation of disparate observations. “Process-tracing” remains a much-invoked, but highly cryptic, designation precisely because its sample is undefined. This is not necessarily a weakness; in many instances, it is a necessity, as is clear from the previous discussion.

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.*

- 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* 98:2 (May) 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* 15:2 (April) 201-32.
- Gerring, John and Rose McDermott. 2005. “Experiments and Observations: Towards a Unified Framework of Research Design.” In process.

