Political Research Quarterly
http://prq.sagepub.com/

How Good Is Good Enough? A Multidimensional, Best-Possible Standard for Research Design

John Gerring

Political Research Quarterly 2011 64: 625 originally published online 16 August 2010
DOI: 10.1177/1065912910361221

Political Research Quarterly
64(3) 625–636
© 2011 University of Utah
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1065912910361221
http://prq.sagepub.com

How Good Is Good Enough? A Multidimensional, Best-Possible Standard for Research Design

John Gerring[1]

Abstract

Recent years have seen a shift in methodological emphasis from the observable properties of a sample to its unobservable properties, that is, judgments about the process by which the data were generated. Considerations of research design have moved front and center. This article attempts to bridge discussions of experimental and quasi-experimental data and of quantitative and qualitative approaches, so as to provide a unified framework for understanding research design in causal analysis. Specifically, the author argues that all research designs aim to satisfy certain fundamental criteria, applicable across methods and across fields. These criteria are understood as desirable, ceteris paribus, and as matters of degree. The implications of this framework for methodological standards in the social sciences are taken up in the final section of the article. There, the author argues for a *best-possible* standard of proof that judges overall methodological adequacy in light of other possible research designs that might be applied to a particular research question.

Keywords

research design, quasi-experiment, natural experiment, qualitative methods

> If you insist on strict proof (or strict disproof) in the empirical sciences, you will never benefit from experience, and never learn from it how wrong you are.
>
> —Karl Popper (1934/1968, 50)

> Our problem as methodologists is to define our course between the extremes of inert skepticism and naïve credulity . . .
>
> —Donald Campbell (1988, 361)

Methodological standards are central to the working of any scientific discipline. Yet the problem of establishing and maintaining standards seems to pose greater practical difficulties in the social sciences than in the natural sciences. Scholars in astronomy, biology, chemistry, computer science, earth sciences, engineering, and physics are less conflicted over what constitutes a finding than their counterparts in anthropology, economics, political science, sociology, and related fields.[1]

What is a reasonable methodological standard for analyzing causal inferences in social science? When is a method of analysis good, or at least good enough? Various approaches to this question may be derived from work in philosophy of science and history of science. However, work emanating from these fields tends to float hazily above the surface of social science praxis. My interest here is in standards as they are applied, or might be applied, to workaday decisions such as those facing reviewers and editors. When should we accept or reject a manuscript, and what are plausible methodological grounds for doing so?

Recent years have seen a dramatic shift in methodological emphasis from the observable properties of a sample to its unobservable properties, in particular, judgments about the process by which the data was generated, which I shall refer to as issues of *research design*. However, we have not fully come to grips with the diversity of research design issues relevant to causal inference. In this article, I propose a unified and relatively comprehensive framework to summarize these research design considerations—a common set of criteria that should

[1]Department of Political Science, Boston University, Boston, MA, USA

Corresponding Author:
John Gerring, Department of Political Science,
Boston University, 232 Bay State Rd, Boston MA 02215, USA
Email: jgerring@bu.edu

apply across experimental and nonexperimental settings and to research in quantitative and qualitative modes.

My second objective is to argue against the application of a binary threshold for scientific adequacy. I shall argue that methodological criteria are multiple, diverse, and prone to methodological trade-offs and practical constraints. Most importantly, different substantive research questions require different methods and result in different overall levels of uncertainty. Some things are easier to study than others. In the concluding section of the article, I lay out the argument for a "best-possible" approach to methodological adequacy.

Note that space constraints compel a terse presentation of complex issues in this article. Readers may consult cited studies for more in-depth discussion of examples and potential solutions to the methodological problems discussed below (see also Gerring 2011).

Methodological Standards

The purpose of a research design is to test a hypothesis. In this article, I shall assume that the hypothesis of interest is causal in nature and that the principal focus of research is on the measurement of a causal effect, that is, *X*'s effect on *Y* across a population of cases. It is important to recognize that causal investigations may have other objectives as well. They may be focused on explaining a single outcome rather than a class of outcomes. They may attempt to explain *all* the causes of a class of outcomes rather than focusing on a single causal factor. They may be directed toward the investigation of causal mechanisms or establishing the boundaries of an inference. Even so, the average effect of *X* on *Y* across a sample of cases (the average treatment effect, or ATE) is still the most common objective, and likely to remain so. It thus serves as a common point of reference in the following discussion.

In measuring a causal effect, research designs (and associated data analyses) strive for *validity* and *precision*. These desiderata come into play at the level of the studied sample (internal validity/precision) and for a larger unstudied population (external validity/precision). Again, it is important to acknowledge that there are other objectives. However, these are usually subsidiary to validity and precision.

A traditional approach to methodological goodness rests on the observable properties of the sample, as revealed by *t*-statistics, model-fit statistics, robustness tests, and model diagnostics. The general intuition is that a test should be sufficiently precise to reveal the relationship between *X* and *Y*; otherwise, it does not meet the standards of scientific rigor. If uncertainty about a proposition, as understood through probabilities generated by the logic of repeated sampling, is greater than 5 percent, or 10 percent, then it does not pass the bar of truth. This perspective is evident in many older econometrics texts (e.g., Blalock 1979).

Recent years have seen a subtle—but cumulatively quite dramatic—shift in emphasis among methodologists from questions of data analysis (the observable qualities of a sample) to questions of research design (often unobservable). Rather than asking what we can learn from a sample once the data is in—the *ex post* approach to methodology—methodologists now advise researchers to do whatever they can, *ex ante*, to obtain a suitable sample of observations for analysis. Likewise, in evaluating a study (conducted by someone else), their attention is likely to focus on the data generation process (DGP) that produced the sample, not just the method of data analysis. "Design trumps analysis," in the much-quoted words of Donald Rubin (2008; see also Rosenbaum 1999; Sekhon 2009; Shadish and Cook 1999, 294).

In particular, contemporary methodologists pay attention to the *experimental* or *quasi-experimental* qualities of a sample. Strictly interpreted, the experimental standard presumes that only those research designs with randomized (and perhaps explicitly manipulated) treatments ought to be given the scientific imprimatur.[2] Loosely interpreted, the quasi-experimental standard may be met so long as treatments are as-if randomized ("exogenous") and/or selection effects are corrected by appropriate statistical procedures, for example, instrumental variables, matching estimators, and the like (Berk 2005; Dunning 2008; Holland 1986; Imbens and Wooldridge 2009; Morgan and Winship 2007; Robinson, McNulty, and Krasno 2009; Sekhon 2009; Shadish, Cook, and Campbell 2002).

The strict experimental standard—only explicitly randomized treatments—is clear and operational. We know it when we see it. However, if adopted, this standard would exclude from the rubric of science a large majority of the work currently conducted in the social sciences. I shall assume that few social scientists are willing to contemplate such an extreme step.

The loose interpretation—quasi- or natural experiments—while more inclusive, is also considerably more ambiguous. It is rarely clear, for example, when circumstances of "conditional independence" or "ignorability" have been achieved with observational data. While we may agree that all causal models should be adequately specified (all problems of identification overcome), these are not matters that can usually be tested empirically. They rest, instead, on a delicate and often quite elaborate skein of assumptions, for example, about the actual method of assignment, the generalizability of the treatment, the level of interference between units, the relative success of instruments or matching estimators in purging potential biases in the sample, and so forth. Robinson, McNulty, and Krasno (2009, 348) comment, "The distinction between a natural experiment and a merely correlational study is neither analytically nor quantitatively demonstrable but is instead a judgment call made by the scholar who is attempting to show the validity and importance of her findings and by the community of scholars that reads and evaluates her research."[3] Needless to say, it is difficult to set acceptable levels of validity and precision—such that work in a scientific mode can be distinguished from work in an unscientific mode—when the bar consists of a series of assumptions about how closely the chosen model replicates the true (ontological) reality.

Another complication is that scholars in this tradition do not always agree on what this benchmark should be—even if they could measure and test it. Researchers and methodologists abide by rather different "minimum" standards of what might qualify a study as quasi-experimental (scientific). Indeed, the quasi-experimental benchmark is rarely made explicit and is therefore applied in a largely ad hoc manner. In this respect, standards of validity and precision are a bit like the layman's version of aesthetics: largely a matter of taste.

Goodness in Research Design

The first goal of this article is to more fully elaborate the elements of "methodological goodness" associated with the laboratory experiment. What are these research design components, and how do they apply across experimental and nonexperimental settings? The second goal of this article is to reach beyond the experimental template to encompass other research design goals and constraints that have little or nothing to do with controlled settings, manipulated treatments, and randomization.

I will argue that methodological criteria pertaining to research design fall into six general categories: *theoretical fit, cumulation, the treatment, the outcome, the sample, and practical constraints.* Each category is associated. . . . Each category is associated with a set of specific criteria, as laid out in Table 1. These are the factors, I argue, that rightly differentiate good and bad research designs.

It should be understood that the framework is focused narrowly on questions of research design *after* a theory and hypothesis has been formulated. It does not reflect criteria that rightly apply to the formation of a causal theory. Sometimes, the most interesting theories are the most difficult to test. However, we are not concerned here with the broader utility of a theory but rather, more narrowly, with its testability.

Theoretical Fit

Since the purpose of an empirical analysis is to shed light on a causal theory, it follows that the closer the alignment between the research design and the theory being tested, the stronger the argument. Four issues bear on the theoretical fit of a research design: *construct validity*, *severity*, *partition*, and the *elimination of rival hypotheses*. All may be considered aspects of a general scientific ideal known as the *crucial* (or critical) *test*, which decisively confirms a theory while disconfirming rival theories (Eckstein 1975).

Table 1. Research Design Tasks and Criteria

1. Theoretical fit
   (a) Does the research design provide an appropriate test for the inference (construct validity)? (b) Is the test easy or hard (severity)? (c) Is the research design segregated from the argument under investigation (partition)? (d) Are alternative explanations ruled out (elimination)?
2. Cumulation
   (a) Is the research design standardized with other similar research on the topic? (b) Does it replicate extant findings and facilitate future replications by other scholars? (c) Are procedures transparent?
3. Treatment
   Is *X* (a) varying, (b) simple, (c) discrete, (d) uniform, (e) evenly distributed, (f) strong, and (g) proximate to *Y*?
4. Outcome
   Is *Y* (a) varying, or at least free to vary?
5. Sample
   Are the chosen observations (a) representative, (b) large in number, (c) at the principal level of analysis, (d) independent (of one another), and (e) causally comparable?
6. Practical considerations
   What (a) pragmatic, (b) logistical, or (c) ethical concerns apply to the construction of a research design?

Construct validity. This term refers to the faithfulness of a research design to the theory under investigation. This includes *concept validity*—the connection between a key concept and the chosen indicators—and also basic assumptions or interpretations of the theory. If a research design deviates significantly from the theory—involving, let us say, questionable assumptions about the theory or building on peripheral elements of the theory—then the theory can scarcely be proven or disproven, for the research design does not bear centrally upon it. By the same token, if a researcher chooses a hypothesis that lies at the core of a theory, the research design has greater relevance.

Severity. Some empirical tests are easy, requiring little of a theory to clear the hurdle (which may be formalized in a statistical test). Other empirical tests are demanding, requiring a great deal of a theory. Other things being equal, we are more likely to believe that a theory is true if it has passed a demanding empirical test, where any potential biases are stacked against the theory in question. *Severe* (aka crucial, demanding, difficult, hard, risky) tests are more conclusive than easy tests (Eckstein 1975; Popper 1934/1968). Severity is a product of several factors: the riskiness of the predictions made by a theory, the cut-off point established for success (high or low), and any other potential biases that might affect the outcome. Of course such biases are never desirable; but they are less problematic if their directionality can be accurately anticipated and if the theory still passes the test.

Partition. Falsifiability is also enhanced insofar as an argument can be effectively isolated, or *partitioned*, from the procedure of empirical analysis. This reduces the possibility that a theory might be adjusted, post hoc, to accommodate negative findings. It also reduces the temptation to construct arguments closely modeled on a particular empirical setting ("curve-fitting"), or research designs whose purpose is to prove (rather than test) a given argument. Ideally—at least for purposes of appraisal—the construction of an argument should be considered a separate step from the testing of that same argument. This may be understood in three ways: as a separation in time (some time should separate the construction of a hypothesis and the testing of that hypothesis), as a separation of data (a hypothesis should be constructed with evidence separate from that which is employed to test the hypothesis), and as a state of mind (theory construction requires creativity, ad hoc adjustments, and nurturing; theory testing requires adherence to rigid a priori procedures and an attitude of skepticism).[4]

Elimination of rival hypotheses. A good research design allows one to prove the main hypothesis, while rejecting "plausible rival hypotheses."[5] Arguably, these are twin sides of the same goal. One can hardly prove hypothesis A if the empirical evidence of one's research is also consistent with hypothesis B. Additionally, the role of elimination in causal assessment stems from a basic set of assumptions about the world. Consider that there are a limited number of causal factors that may influence phenomenon *Y*. Suppose further that we are able to intuit—from common sense and from extant work on a subject—what these might be. This provides a list of suspects, that is, possible causes. In this context, it should be clear why—although we might test A by itself—we shall be more convinced of our conclusions (and shall know much more about the phenomenon of interest) if we are also able to refute other hypotheses about *Y*. Insofar as a research design is constructed to show that A—rather than B, C, and D—caused *Y*, we shall be more inclined to accept this finding. Causal assessment is always to some extent a matter of comparison (Feyerabend 1963; Harman 1965; Miller 1987), and a research design that allows us to effectively compare the truth-value of rival theories is generally a better research design.

Cumulation

Science is not a solitary venture; it is, indeed, best conceptualized as a collaborative project among researchers working in a particular area. This means that a research design's utility is partly a product of its methodological fit with extant work. Three elements facilitate cumulation: the *standardization* of procedures across studies, the *replication* of results, and the *transparency* of procedures.

Standardization. If there is a usual way of investigating a particular research question, this should slavishly imitated, at least as a point of departure. The standardization of research designs allows findings from diverse studies to cumulate. A thousand studies of the same subject—no matter how impeccable their claims to internal validity—will make only a small contribution to the growth of knowledge if they are designed in ad hoc (and hence incommensurable) ways.

Naturally, each study must have some element of originality (even if it is only the application of an old idea to a new location). And there may be more than one way to approach a subject; often, there are benefits from triangulation (aka multimethod research). *Y*et the benefits of alternative approaches are assessable only insofar as a study can be measured by a common yardstick provided by extant work on a subject. The call for standardization should not be confused with a call for theoretical timidity or methodological monism. It is, rather, a call for a more organized approach to knowledge gathering (Berk 2005; Berk et al. 1992; Bloom, Hill, and Riccio 2002). Similarly, one would not want standardization to impede potential improvements in research design. The point is that the new method should be compared with—benchmarked against—an established method. If the new method can be shown to provide a substantial improvement relative to the old method, it should, in turn, become the industry standard.

Replication. Another way that scientific activity relates to a community of scholars is through the replication of results. Research on a topic typically begins by replicating key findings related to that research. This will help clarify the validity of the chosen research design, not to mention the validity of the previous finding. This is the initial replication. Other replications occur after a study has been completed (King 1995). To facilitate replication, a research design must be conducted in such a way that future scholars can reproduce its results. Note that replicability is simply a method of checking a study's internal validity. It is also a means of testing—and where necessary, reevaluating—a study's external validity. The broader meaning of replication refers to the ability of future researchers to replicate a research design in a new context or with a new source of data. This provides a way for future research to build on what has already been accomplished, as discussed above.

Transparency. Standardization and replication are possible only insofar as procedures employed in empirical analyses are transparent. One cannot standardize or replicate what is ambiguous. It is common in natural sciences—but again, not in the social sciences—for researchers to maintain a laboratory notebook, in which a close record is kept of how an empirical analysis unfolds. While it may not be necessary to record every specification test, it should at least be possible for future scholars to see which tests were conducted, in what order, and with what implications for the theory. By contrast, if scholars see only the final product of a piece of research (which may have unfolded over many years) it is more difficult to render judgment on its truth value. One fears, in particular, that the final data tables may contain the one set of tests that culminated in "positive" (i.e., theoretically and statistically significant) results, ignoring hundreds of prior tests in which the null hypothesis could not be rejected. Granted, the achievement of full transparency imposes costs on researchers, mostly in the form of time and effort (since the posting of notebooks is essentially costless). And it does not entirely solve problems of accountability. We shall never know if all procedures and results were faithfully recorded. However, the institution of a transparency regime is a precondition of greater accountability and may in time enhance the validity and precision of empirical analysis in the social sciences.

The Treatment

For purposes of testing, a good treatment ought to be (1) varying, (2) simple, (3) discrete, (4) uniform, (5) evenly distributed, (6) strong, and (7) proximate to Y.

Variation. Empirical evidence of causal relationships is largely covariational in nature. There are numerous near-synonyms for this basic idea, including *correlation*, *association*, *constant conjunction* (Hume), *concomitant variation* (Mill), and *congruity* (Bennett).[6] Whatever the precise nature of the relationship, X and Y must display some covariational pattern—at least hypothetically. Without it, causation cannot be at work. Empirical covariation is thus appropriately regarded as a necessary (though by no means sufficient) condition of a causal relationship.

Variation in X—the explanatory variable of interest—is especially crucial. If we have no such variation, our analysis must take the form of a counterfactual thought experiment in which such variation is imagined—a much weaker research design in most settings (see Fearon 1991; Gerring 2007; Lebow 2007). By variation, we mean not only that X has maximum *range* of variation but also that it embodies maximum *dispersion* (around the mean). These twin concepts are captured in the statistical concept of *variance*. A good research design should embody observations with maximum variance on X. Although the concept of variance expresses more precisely what we mean to say, it has become common to refer to this issue of research design as one of "variation." In the interests of terminological consistency, I adopt that usage here.

Simple. A simple treatment involves only two conditions: a treatment condition ($X = 1$) and a control condition ($X = 0$). Complexity, by contrast, may mean many things. It may involve different levels (doses) of a single treatment, different treatments altogether (each of which exhibits some degree of variation), combinations of different treatments, treatments differing in timing or in some other contextual factor, and so forth. The point is that each variation in treatment must be regarded as a distinct hypothesis, and the more hypotheses there are, the harder it will be to achieve causal inference. At the very least, a larger sample will be required, perhaps coupled with stronger assumptions about the shape of the data generation process. Thus, although some theories require complex treatments (it is not always in the power of the researcher to simplify the treatment), we must remain cognizant of the costs imposed by complexity. Ceteris paribus, simplicity is desirable. Alternately stated, a research design should not be any more complex than it needs to be.

Discrete. The discreteness of a treatment largely determines the ease with which causal relations will be observable. A discrete treatment is potentially observable prior to treatment and after treatment, allowing for pretests and posttests. A discrete treatment is also potentially observable across the treatment group (where $X = 1$) and the control group (where $X = 0$). If, however, the treatment is nondiscrete, there is no apparent baseline against which the effect of the treatment can be compared. As it happens, causal factors of theoretical interest are often matters of degree and therefore do not constitute discrete treatments in the real world (unmanipulated by the researcher). What is messy about observational data is not simply the nonrandomized treatment (as discussed below) but also the nature of the treatment itself.

Uniformity. In order to test the impact of a causal factor it is essential that the intervention be relatively uniform across the chosen units. If the treatment is binary (0/1) or multichotomous (0/1/2/…), then achieving uniformity is a simple matter of making sure that doses are correct and that a "3" administered to one unit is the same as a "3" administered to another unit. Note that insofar as a treatment is non-uniform its causal effect will be impossible to interpret.

Even distribution. The distribution of values for X also affect our ability to test X's impact on Y. Ideally, one would like to have a fairly even distribution of actual values across the various possible values for X. If the treatment is simple (0/1) then there should be an equal number of cases in which $X = 0$ and $X = 1$. If the treatment is continuous, then each portion of the potential distribution of X should be represented (more or less equally) by actual cases. Where this is not so, we face a problem of "missing values" or "empty cells" (Rosenbaum 1984). Let us say that X varies in principle from 0 to 10, but we have no cases where $7 < X < 9$. In this situation, we cannot directly test X's effect on Y for situations in which X lies between 7 and 9. We can of course intuit these relationships from cases where $0 < X < 7$ and $9 < X < 10$. But we cannot really know for sure. Likewise, if we have only a few cases that exemplify rare values for X (in this instance, $7 < X < 9$), this range is testable but is also liable to threats from stochastic error.

Strength. A strong signal is always easier to detect than a weak signal. Thus, it is helpful if the treatment chosen for empirical testing has a (putatively) strong effect on Y. Miniscule causal effects are likely to result in a failure to reject the null hypothesis, even if (some form of) the hypothesis is true.[7]

Proximity. To observe X's effect on Y, it is helpful if the treatment lies fairly close (in causal distance) to the chosen outcome. In this fashion, limits are placed on the number of possible confounders and the temporal scope of the investigation. Note that a lengthy waiting period between intervention and outcome means that, unless the units under investigation can be isolated for long periods in laboratory conditions, the causal relationship will be subject to all manner of posttreatment threats to inference, as discussed below. Moreover, a close proximity between X and Y means that the causal mechanism(s) running from X to Y is likely to be easier to observe and to identify.

The Outcome

With respect to the outcome of a hypothesis, the main requirement is that Y be free to vary. Otherwise, there is no point in conducting a test. That said, the problem of "variation in Y" is more relevant to observational settings than to experimental settings. In observational settings, it is highly desirable to identify settings in which Y actually does vary (rather than is simply free to vary, from some hypothetical perspective). Here, we cannot manipulate X, and we may have much less assurance about Y's capacity to respond to a given change in X. Confounding factors are legion, and a stable outcome could therefore be credited to many things. Thus, we are reassured if we observe some variation in Y, whether or not it was caused by X. This does not mean that we learn nothing from a situation in which X changes but Y does not. It means that we learn more—much more—from a setting in which there is variation on both X and Y.

The problem of Y variation is most apparent when it is extremely rare, for example, with binary outcomes like war, revolution, or school shootings. To test a particular causal hypothesis relating to these outcomes (in a nonexperimental context), we must identify a sample within which there are some examples of these rare events; otherwise, there is no variation on Y to explain. Although "selecting on the dependent variable" may introduce biases into the sample (relative to the population), methods have been developed to provide unbiased estimates from samples that have an estimable bias (known variously as case-control sampling, response-based sampling, or choice-based sampling) (Breslow 1996; Manski 1995, chap. 4). Also, it should be noted that this supposed sin is often mislabeled when applied to case study designs (Gerring 2007). As with X variation, Y variation refers to the maximum range of variation (minimum to maximum) in an outcome but also to maximum dispersion around the mean. These twin concepts are captured in the statistical concept of *variance*.

The Sample

Ideally, observations chosen for inclusion in a sample should be (1) representative, (2) large in number (N), (3) at the principal level of analysis, (4) independent, and (5) comparable.[8]

Representativeness. Perhaps the most important grounds for drawing conclusions about the external validity of a proposition rests on the representativeness of the chosen sample. Is the sample similar to the population in ways that might affect the relationship of X to Y? Regardless of the technique of case selection (e.g., random, purposive, ad hoc), the concern for representativeness must occupy any scholar who wishes to generalize her or his results.

Large-N. More observations are better than fewer; hence, a larger sample ("N") is superior to a smaller sample—all other things being equal. This is fairly commonsensical. The same logic that compels us to provide empirical support for our beliefs also motivates us to accumulate multiple observations. More specifically, a larger sample may clarify what the correlational relationship between X and Y is. The more observations one has, the less indeterminacy there is with respect to X's possible relationship to Y, and the less results are subject to stochastic variation and measurement error. Additionally, it will be easier to attain a sample that is representative of a larger population (because methods of random sampling can be applied). A large sample may also help in specifying a hypothesis—the envisioned positive and negative outcomes, a set of cases which the proposition is intended to explain (a context or contrast space), and operational definitions of the foregoing. If there is only one observation, or multiple observations drawn from one unit, these tasks become rather ambiguous. The problem, briefly stated, is that with a narrow empirical ambit the researcher is faced with an overabundance of ways to operationalize a given hypothesis.[9]

Level of analysis. To be sure, increasing the *N* does not always enhance the quality of a research design. Not all observations are equally useful. Usually, observations are most helpful in elucidating causal relationships when they are situated at the same level of analysis as the main hypothesis. One often faces difficulties if one attempts to explain the activity of a particular kind of unit by examining units at a higher, or lower, level of analysis (Lieberson 1985, chap. 5). This is not to say that observations at different levels of analysis are unhelpful in ascertaining causal mechanisms and in confirming other aspects of a theory. Often, they are. The point is simply that observations lying at the same level of analysis as the theory are generally more revealing. Certainly, they are more indicative of the size and shape of the main causal effect.

Independence. Useful observations should be independent of one another. This refers to the separateness of observations such that each observation provides new and valuable evidence of a causal relationship. Violations of independence, that is, dependence, may be introduced by *serial* (temporal) *autocorrelation*, where one observation of a given unit depends upon a previous observation of that unit; or by *spatial correlation*, where the observations obtained from one unit depend upon observations taken from another unit. Prior to treatment, commonalities among units may create a clustering of attributes ("correlated groups") that violate the independence of each unit. After treatment is administered one often faces a problem of *interference*, when units "contaminate" each other (Rosenbaum 2007).

Violations of independence, like most other criteria, are usually matters of degree and may not disqualify an observation or a unit from inclusion. However, it is likely to lend a false sense of precision to a causal analysis by virtue of artificially inflating the number of observations in the sample. More troubling, it may introduce a problem of bias by creating noncomparabilities among the sample, an issue addressed below.

Comparability. Closely related—indeed, often intermingled—with the criterion of independence is the final criterion of a good sample, causal comparability. This means that the expected value of *Y* for a given value of *X* should be the same across the studied units throughout the period of analysis.[10] If they are, we can say that a group of units is causally comparable, or equivalent, with respect to a given hypothesis. A minimal understanding of this criterion requires only that units be comparable to one another *on average*, which is to say that a large error rate across units is satisfactory as long as its distribution is centered on the true mean (i.e., as long as the error is random). A maximal understanding of this criterion, sometimes expressed as *unit homogeneity*, is that units should evidence *identical* responses of *Y* to a given value of *X* across units. The latter ideal is rarely, if ever, realized in the world of social science (and perhaps not even in the world of natural science). However, the minimal definition seems too minimal. After all, noncomparabilities are always somewhat problematic (at the very least, they introduce extraneous noise). Thus, we shall regard this desideratum as a matter of degrees.

Some noncomparabilities introduce noise, or random error, into the analysis. A second sort of noncomparability is correlated with the treatment and is therefore regarded as a source of systematic error, or bias—a *confounder*. This is the more serious problem associated with the issue of causal comparability.

When the empirical comparison is spatial (across cases), comparability may be achieved at the beginning of an analysis by randomly assigning the treatment across cases—the experimental technique. However, since most experiments (and most nonexperimental studies) extend over a period of time, well beyond the initial assignment of treatment, we must also consider additional confounders that may arise prior to the final posttest. Random assignment is a necessary, but by no means sufficient, condition for valid causal inference. Violations of posttreatment comparability may be introduced by a variety of threats, for which a highly specialized vocabulary has evolved (e.g., noncompliance, contamination, reputation effects, experimenter effects, testing effects, history, instrumentation effects, and so forth).

Practical Constraints

Often, one chooses a research design because it is more convenient for one to do so, or perhaps because it is impossible to do otherwise (Barrett and Cason 1997; Lieberman, Howard, and Lynch 2004; Van Evera 1997). For example, we may lack the language skills to study something else. Political or cultural barriers may prevent us from gathering additional information. Evidence itself may be scarce. Funding opportunities may be limited. And of course time is always limited. Ethical considerations may also constrain our ability to develop a convincing solution to problems of causal inference.[13] Practical constraints are not methodological in the usual sense of the term. However, they often prove decisive in crafting the methodology of a given study and in judging its overall quality.

A Best-Possible Standard

I began by reviewing two approaches to methodological adequacy—one based on the observable properties of a sample and another based on the presumed data generating process (DGP). Let us call the first "statistical" and the second "experimental" (or quasi-experimental). I argued that neither manages to erect a defensible and operational standard by which to distinguish sound and unsound research.

Some readers may find this argument anodyne. Arguably, it was not really the intention of writers in these traditions to erect a systematic and comprehensive standard by which to separate good and bad research. Perhaps it would be more correct to say that these traditions elucidate *certain features* relevant to judging the methodological quality of a study. In this light, the present work is rightly viewed as an extension, rather than a refutation, of prior efforts.

I want to push the pluralistic implications of this endeavor a bit further. Much of the angst behind the current *Methodenstreit* in the social sciences arises from the tacit belief that there exists one uniform standard of methodological adequacy that can be applied to all work within a given discipline, or even across disciplines. This species of methodological monism serves to engender feelings of insufficiency on the part of many practitioners, who do not measure up to the rigid strictures of this standard of truth. It also encourages methodological faddism, exemplified by the current raft of studies employing natural experiments (often not very experimental), instrumental variables (rarely with good instruments), or matching estimators (often failing to fully resolve the assignment problem). These are wonderful—and often useful—methodological tricks, but they are frequently applied inappropriately or given an overly optimistic interpretation.

In contrast to the single-standard view, I argue that methodological criteria applicable to research designs in social science are multiple and diverse. Six general areas, and multiple associated criteria (summarized in Table 1), were discussed briefly in the preceding section. Even this lengthy compendium may not be entirely comprehensive. (For example, it omits issues of conceptualization and measurement, which presumably undergird all attempts at causal inference.) Indeed, one might question whether a truly comprehensive checklist of methodological criteria applying to research design could ever be developed, given that the subject itself is difficult to bound. In this light, the argument for pluralism is proven, a fortiori.

In sum, the construction of a research design is a *complex* and *multidimensional* task. Moreover, the diverse methodological criteria encompassed in the framework are often in conflict with one another. Attaining one research design goal may affect the attainment of other goals. Trade-offs are endemic. This means that every dimension listed in Table 1 must be understood with a ceteris paribus caveat.

This means that the goals of research design in causal assessment are considerably more complicated than most extant accounts suggest. While there are shared standards (Brady and Collier 2004), as exemplified in Table 1, there is no single threshold that would allow us to distinguish work that is methodologically adequate ("science") from work that is methodologically inadequate ("nonscience"). The search for a simple and parsimonious formula (on the order of *t*-statistics) by which to evaluate methodological goodness is doomed.

In its stead, I want to propose a *best-possible* approach to methodological adequacy. According to this standard, researchers are obliged to maximize adequacy across the multiple dimensions identified in Table 1, reconciling potential conflicts wherever possible. There is no absolute threshold that must be crossed in order for a study to be regarded as scientific (i.e., worthy of publication in a scientific journal). Rather, the standard is comparative—relative to all possible research designs that might be devised to address the chosen research question.

This allows for research flowing from methods that lie far from the experimental ideal to enter the social science pantheon without shame or derogation—but only if no better expedient can be found. Just as we honor classicists, astronomers, and theoretical physicists (despite the speculative nature of their trades), we should also honor those social scientists who labor to reach causal conclusions on the basis of sketchy data, if no plausible means can be identified to improve the quality of the data.

Studies based on weak evidence must answer a very difficult, but absolutely essential, question: could the research design be significantly improved upon? What is achievable, *under the circumstances*? If a research ideal is entirely out of reach—by virtue of lack of data, lack of funding sources, lack of cooperation on the part of relevant authorities, or ethical considerations—it is pointless to admonish an author for failing to achieve it. Perfection becomes the enemy of scientific advance. We must guard against the possibility that work adding value to what we already know about a given subject might be rejected even when no better approach is forthcoming. Standards must be realistic.

If, on the other hand, a better approach to a given subject can be envisioned, and the costs are not too great, a work based on weak data is rightly criticized and perhaps ultimately rejected (as unscientific). We must guard against the possibility that second-best research designs might drive out first-best research designs simply because the former are more familiar to scholars or marginally easier to implement. Mediocrity should not be the enemy of excellence.

It is time to acknowledge that work on the frontiers of social science, in common with work on the frontiers of natural science, is prone to a great deal of uncertainty. Moreover, there is a great deal of *variation* in uncertainty across subjects in the social sciences. The causes and effects of democratization (Geddes 2007) or the causes of economic growth (Kenny and Williams 2000) will never be known with the same degree of precision as the effect of de-worming on school attendance (Miguel and Kremer 2004). This means that uncertainty could be vanquished, but only at the cost of forgoing research on many subjects generally regarded as theoretically important or policy relevant.

The practical solution is to embrace the uncertainty of our enterprise, honestly and explicitly. This does not mean that we should celebrate uncertainty. It means that uncertainty should be minimized but that binary thresholds (e.g., 95 percent) should be eschewed. While a 95 percent threshold of certainty is perfectly appropriate for some questions (e.g., those that can be approached experimentally), it is probably unrealistic for many others. Unfortunately, in the latter case we continue to demand the same level of statistical uncertainty (though scholars are aware that *p*-values mean something quite different if derived from nonexperimental data).

By way of acknowledging the uncertain status of much of the evidence in social science, scholars have developed a number of framing devices. They may introduce a statistical finding as “descriptive” rather than causal (even when the theoretical motivation of the analysis is causal). They may frame the evidence as a series of “stylized facts”—consistent with a theory, but by no means conclusive. And so forth. These rhetorical devices serve a worthy purpose insofar as they overcome a rigid and unproductive dichotomy (science/nonscience). However, their status as tools of causal inference are highly ambiguous (what is “descriptive” evidence, or “stylized facts”?). These terms, and countless others of a similar nature, serve to smuggle into the causal narrative pieces of evidence that have no generally recognized scientific foundation.

It should not be necessary to play games with words to introduce evidence relevant to solving a particular problem. This is why it is so important that a diverse array of observational methods be recognized as legitimate. Of course, this does not mean that we should regard them as equivalent to experimental data. What it means, rather, is that the researcher should remain open to all the available evidence, weighing each according to its relative import to the question at hand.

For each research design, we must ask whether the choices made by a researcher fully exploit the opportunities that the data present, or whether better data could have been collected. The framework developed in this article is intended to offer an encompassing menu upon which to judge whether all issues of research design have been fairly considered. It is to be hoped that this realistic and inclusive standard of estimating *X*’s effect on *Y* will obviate some of the posturing and prevarication that accompanies publication in top social science venues, where researchers often feel compelled to pretend they have attained the highest standards of truth, regardless of realities on the ground. Weaknesses in design and analysis should be openly acknowledged rather than hidden in footnotes or obscured in jargon and endless statistical tables. These elements of uncertainty should not preclude publication in top journals—unless of course better methods are possible.

This is important not just as a matter of intellectual honesty but also for the long-run development of the social sciences. Note that the cumulation of knowledge in a field probably depends as much on methodological transparency as on statistically significant results. We have a fighting chance of reaching consensus on the causal impact of a difficult question such as school vouchers if scholars are scrupulous in reporting the strengths and weaknesses of each piece of research and if each result is accompanied by an overall estimate of uncertainty (taking all factors into consideration). By contrast, there is little prospect of attaining consensus if each study strives for some statistically significant finding, downplaying threats to inference and remaining silent on statistically insignificant results (a point made nicely by Gerber, Green, and Nickerson 2001).

Standards of scholarship need to be adapted so as to structure the incentives of scholars in appropriate ways. It is to be hoped that a flexible and multidimensional standard, understandable in relation to other potential research designs that might be applied to the same problem, will serve that goal. A truly multidimensional approach to testing offers no quick and easy test on the order of *t*-statistics to demarcate science from quackery. The hard task of vetting remains hard. And it remains hard precisely because there are so many divergent goals of social science, so many dimensions of goodness to consider, and no necessary and sufficient conditions that apply universally.

Acknowledgments

I am grateful to comments and suggestions received from Andy Harris and Evan Lieberman.

Declaration of Conflicting Interest

The author declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author received no financial support for the research and/or authorship of this article.

Notes

1. A test of this hypothesis might compare the variance across reviews received for a random sample of manuscript submissions to top journals in natural science and social science fields. I suspect that the variance is considerably higher, for example, among reviewers for the *American Political Science Review* than among reviewers for *Nature* or *Science*. See also Samuelson (1959, 189).
2. This seems to be the standard advocated by Gerber, Green, and Kaplan (2004).
3. On this point, see also Dunning (2008) and Pearl (2009).
4. King, Keohane, and Verba (1994) advise: "Ad hoc adjustments in a theory that does not fit existing data must be used rarely" (p. 21). "Always . . . avoid using the same data to evaluate the theory [you] used to develop it" (p. 46). Original data can be reused "as long as the implication does not 'come out of' the data but is a hypothesis independently suggested by the theory or a different data set" (p. 30). See also Eckstein (1992, 266), Friedman (1953).
5. The term is credited to Campbell and Stanley (1963). See also Rindskopf (2000). The method of elimination was first articulated as a "method of residues" by J. S. Mill (1843/1872).
6. See Bennett (1999), Hume (1960, 219), Marini and Singer (1988), Neuman (1997, 50), Mill (1843/1872, 263). These terms are not exactly synonymous (no two terms are). Yet the differences among them are so slight as to hinder, rather than enhance, clarity.
7. Of course, a weak signal can be compensated by other research design virtues, for example, sensitive instrumentation (leading to high precision and low measurement error), few confounders, and a large sample of observations. This criterion, like others, is accompanied by a ceteris paribus caveat.
8. Note that our concern here is with the qualities of a sample as initially chosen for analysis, that is, the process of sample selection (sometimes referred to as case selection). It should be kept in mind that problems of representativeness and comparability may also be introduced at a later stage of the analysis, that is, once one considers the nature of the treatment. These issues are addressed in the following section.
9. There is one exception to the large-*N* criterion: the rejection of invariant causal arguments. Where *X* is understood as necessary and/or sufficient for *Y* or a particular causal mechanism is stipulated as integral to a theory, a single observation that contradicts the posited features may allow the researcher to falsify a proposition with only a single observation (Dion 1998). This is a rare circumstance, to be sure, for there are few deterministic arguments in the social sciences. Moreover, even where arguments appear to be deterministic, there is usually some theoretical and empirical wiggle room—as one can see with the various interpretations of the "democratic peace" hypothesis that have surfaced in recent years (Elman 1997).
10. Strictly speaking, a single observation cannot be causally comparable to another because a single observation does not register variation between *X* and *Y*. Causal comparability is the attribute of a set of observations, sometimes understood as a case or unit.
11. See George and McKeown (1985, 34ff) and Goldstone (1997). The idea of process tracing is also similar to judgments about *context*, which often play an important role in causal inference (Goodin and Tilly 2006). When invoking "context," one is invoking an idea of how *X* influences *Y*—or not—within a particular setting. Thus, I treat the broad category of contextual evidence as a species of process tracing.
12. The literature on pattern matching, which derives from early work by Donald Campbell (1966), is sparse—reflecting its peripheral status. Trochim (1989) is one of the few writers to subsequently address the question in a sustained way (see also George and Bennett 2005). However, it might be argued that these techniques are employed quite frequently, if not always explicitly, as an auxiliary means of testing causal propositions.
13. Kelman (1982) offers general reflections on research ethics in the social sciences. Mazur (2007) and Sales and Felkman (2000) discuss research on human subjects. Paluck (2008) and Wood (2006) investigate ethical dilemmas of field research, with special focus on areas of intense conflict.

References

Barrett, Christopher, and Jeffery Cason. 1997. *Overseas research*. Baltimore: Johns Hopkins University Press.

Bennett, Andrew. 1999. Causal inference in case studies: From Mill's methods to causal mechanisms. Paper presented at the annual meetings of the American Political Science Association, Atlanta, GA, September.

Berk, Richard A. 2005. Randomized experiments as the bronze standard. Unpublished manuscript.

Berk, Richard A., A. Campbell, R. Klapp, and Bruce Western. 1992. The differential deterrent effects of an arrest in incidents of domestic violence: A Bayesian analysis of four randomized field experiments. *American Sociological Review* 5:689-708.

Blalock, Hubert M., Jr. 1979. *Social statistics*. 2nd ed. New York: McGraw-Hill.

Bloom, H. S., C. J. Hill, and J. A. Riccio. 2002. Linking program implementation and effectiveness: Lessons from a pooled sample of welfare-to-work experiments. *Journal of Policy Analysis and Management* 22:551-75.

Brady, Henry E., and David Collier, eds. 2004. *Rethinking social inquiry: Diverse tools, shared standards*. Lanham, MD: Rowman & Littlefield.

Breslow, N. E. 1996. Statistics in epidemiology: The case-control study. *Journal of the American Statistical Association* 91:14-28.

Campbell, Donald T. 1966. Pattern matching as an essential in distal knowing. In *The psychology of Egon Brunswick*, ed. K. R. Hammond. New York: Holt, Rinehart and Winston.

Campbell, Donald T. 1988. *Methodology and epistemology for social science*. Edited by E. Samuel Overman. Chicago: University of Chicago Press.

Campbell, Donald T., and Julian Stanley. 1963. *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Dion, Douglas. 1998. Evidence and inference in the comparative case study. *Comparative Politics* 30 (2): 127-45.

Dunning, Thad. 2008. Improving causal inference: Strengths and limitations of natural experiments. *Political Research Quarterly* 61:282-93.

Eckstein, Harry. 1975. Case studies and theory in political science. In *Handbook of political science*, vol. 7, *Political science: Scope and theory*, ed. Fred I. Greenstein and Nelson W. Polsby, 94-137. Reading, MA: Addison-Wesley.

Eckstein, Harry. 1992. *Regarding politics: Essays on political theory, stability, and change*. Berkeley: University of California Press.

Elman, Miriam Fendius. 1997. *Paths to peace: Is democracy the answer?* Cambridge, MA: MIT Press.

Fearon, James. 1991. Counter factuals and hypothesis testing in political science. *World Politics* 43 (January): 169-95.

Feyerabend, Paul. 1963. How to be a good empiricist: A plea for tolerance in matters epistemological. *Philosophy of Science: The Delaware Seminar* 2:3-39.

Friedman, Milton. 1953. The methodology of positive economics. In *Essays in positive economics*, 3-43. Chicago: University of Chicago Press.

Geddes, Barbara. 2007. What causes democratization? In *The Oxford handbook of comparative politics*, ed. Carles Boix and Susan Stokes, 317-39. Oxford: Oxford University Press.

George, Alexander L., and Andrew Bennett. 2005. *Case studies and theory development*. Cambridge, MA: MIT Press.

George, Alexander L., and Timothy J. McKeown. 1985. *Case studies and theories of organizational decision making*. Vol. 2 of *Advances in information processing in organizations*. Santa Barbara, CA: JAI.

Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2004. The illusion of learning from observational research. In *Problems and methods in the study of politics*, ed. Ian Shapiro, Rogers M. Smith, and Tarek E. Masoud, 251-73. Cambridge: Cambridge University Press.

Gerber, Alan S., Donald P. Green, and David Nickerson. 2001. Testing for publication bias in political science. *Political Analysis* 9 (4): 385-92.

Gerring, John. 2007. *Case study research: Principles and practices*. Cambridge: Cambridge University Press.

Gerring, John. 2011. *Social science methodology: Tasks, strategies, and criteria*. Cambridge: Cambridge University Press.

Glynn, Adam N., and Kevin M. Quinn. 2009. Why process matters for causal inference. Unpublished manuscript, Department of Government, Harvard University, Cambridge, MA.

Goldstone, Jack A. 1997. Methodological issues in comparative macrosociology. *Comparative Social Research* 16:107-20.

Goldthorpe, John H. 1997. Current issues in comparative macrosociology: A debate on methodological issues [with response to commentaries]. *Comparative Social Research* 16:121-32.

Goodin, Robert, and Charles Tilly, eds. 2006. *The Oxford handbook of contextual analysis*. Oxford: Oxford University Press.

Harman, Gilbert. 1965. The inference to the best explanation. *Philosophical Review* 74:88-95.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81 (396): 945-60.

Hume, David. 1960. The idea of necessary connexion [from *An enquiry concerning human understanding*, sec. 7]. In *The structure of scientific thought: An introduction to philosophy of science*, ed. Edward H. Madden. London: Routledge & Kegan Paul.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47 (1): 5-86.

Kelman, Herbert C. 1982. Ethical issues in different social science methods. In *Ethical issues in social science research*, ed. Tom L. Beauchamp, Ruth R. Faden, R. Jay Wallace Jr., and LeRoy Walters, 40-98. Baltimore: Johns Hopkins University Press.

Kenny, Charles, and David Williams. 2000. What do we know about economic growth? Or, why don't we know very much? *World Development* 29 (1): 1-22.

King, Gary. 1995. Replication, replication. *PS: Political Science and Politics* 28 (3): 443-99.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

Lebow, Richard Ned. 2007. Counterfactual thought experiments: A necessary teaching tool. *History Teacher* 40 (2): 153-76.

Lieberman, Evan S., Marc Howard, and Julia Lynch. 2004. Symposium: Field research. *Qualitative Methods* 2 (1): 2-15.

Lieberson, Stanley. 1985. *Making it count: The improvement of social research and theory*. Berkeley: University of California Press.

Manski, Charles F. 1995. *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.

Marini, Margaret, and Burton Singer. 1988. Causality in the social sciences. *Sociological Methodology* 18:347-409.

Mazur, Dennis J. 2007. *Evaluating the science and ethics of research on humans: A guide for IRB members*. Baltimore: Johns Hopkins University Press.

Miguel, Edward, and Michael Kremer. 2004. Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 72 (1): 193-217.

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

Miller, Richard W. 1987. *Fact and method: Explanation, confirmation and reality in the natural and the social sciences*. Princeton, NJ: Princeton University Press.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge: Cambridge University Press.

Neuman, W. Lawrence. 1997. *Social research methods: Qualitative and quantitative approaches*. 2nd ed. Boston: Allyn & Bacon.

Paluck, Elizabeth Levy. 2008. Methods and ethics with research teams and NGOs: Comparing experiences across the border of Rwanda and Democratic Republic of Congo. Unpublished manuscript, Princeton University, Princeton, NJ.

Pearl, Judea. 2009. Causal inference in statistics: An overview. *Statistics Surveys* 3: 96-146.

Popper, Karl. 1934/1968. *The logic of scientific discovery*. New York: Harper & Row.

Rindskopf, David. 2000. Plausible rival hypotheses in measurement, design, and scientific theory. In *Research design: Donald Campbell's legacy*, vol. II, ed. Leonard Bickman, 1-12. Thousand Oaks, CA: Sage.

Robinson, Gregory, John E. McNulty, and Jonathan S. Krasno. 2009. Observing the counterfactual? The search for political experiments in nature. *Political Analysis* 17 (4): 341-57.

Rosenbaum, Paul R. 1984. From association to causation in observational studies: The role of strongly ignorable treatment assignment. *Journal of the American Statistical Association* 79 (385): 41-48.

Rosenbaum, Paul R. 1999. Choice as an alternative to control in observational studies. *Statistical Science* 14 (3): 259-304.

Rosenbaum, Paul R. 2007. Interference between units in randomized experiments. *Journal of the American Statistical Association* 102 (477): 191-200.

Rubin, Donald B. 2008. For objective causal inference, design trumps analysis. *Annals of Applied Statistics* 2 (3): 808-40.

Sales, Bruce Dennis, and Susan Folkman, eds. 2000. *Ethics in research with human participants*. Washington, DC: American Psychological Association.

Samuelson, Paul A. 1959. What economists know. In *The human meaning of the social sciences*, ed. Daniel Lerner. New York: Meridian Books.

Sekhon, Jasjeet S. 2009. Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science* 12:487-508.

Shadish, William R., and Thomas D. Cook. 1999. Design rules: More steps toward a complete theory of quasi-experimentation. *Statistical Science* 14 (3): 294-300.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Trochim, William M. K. 1989. Outcome pattern matching and program theory. *Evaluation and Program Planning* 12:355-66.

Van Evera, Stephen. 1997. *Guide to methods for students of political science*. Ithaca, NY: Cornell University Press.

Wood, Elisabeth. 2006. The ethical challenges of field research in conflict zones. *Qualitative Sociology* 29 (3): 373-86.